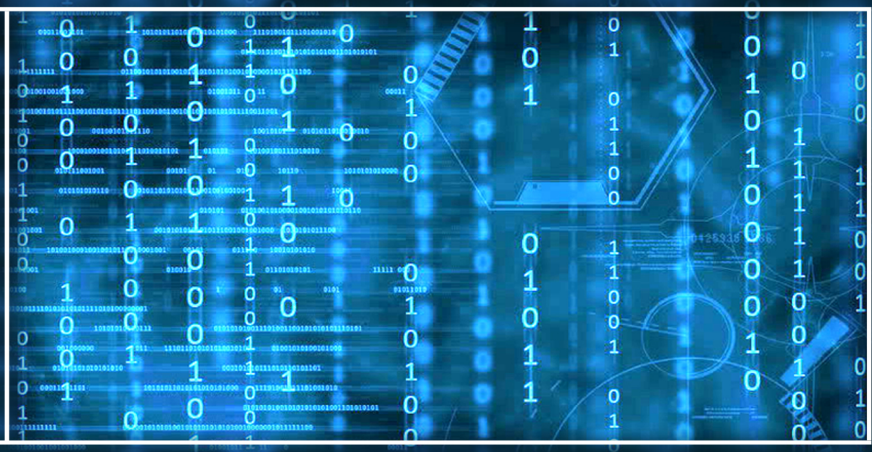
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 11 Issue 11

November 2020



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

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 11 Issue 11 November 2020
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

CONTENTS

Classification of Imbalanced Datasets using One-Class SVM, k-Nearest Neighbors and CART Algorithm

Maruthi Rohit Ayyagari
College of Business, University of Dallas
Irving, Texas, USA

***Abstract*—In this paper a new algorithm, OKC classifier is proposed that is a hybrid of One-Class SVM, k-Nearest Neighbours and CART algorithms. The performance of most of the classification algorithms is significantly influenced by certain characteristics of datasets on which these are modeled such as imbalance in class distribution, class overlapping, lack of density, etc. The proposed algorithm can perform the classification task on imbalanced datasets without re-sampling. This algorithm is compared against a few well known classification algorithms and on datasets having varying degrees of class imbalance and class overlap. The experimental results demonstrate that the proposed algorithm has performed better than a number of standard classification algorithms.**

Keywords*—SVM; k-NN; CART; OKC; classification; machine learning

I. Introduction

Classification is a task of categorizing the instances of a specified class from amongst the given set of classes. This task is done by a classifier that is demonstrated on a dataset of training cases. Most of the classification algorithms expect balanced class, i.e. there will be practically equivalent number of cases from all classes in the preparation dataset. But in many real world domains, like fraud detection, medical diagnosis, etc., the number of examples that belong to one class may severely outnumber the instances that belong to another class/classes. Such datasets, in which significant differences in the proportion of cases having a place with various classes are possible, called imbalanced datasets. The imbalance in class distribution could prompt high misclassification rates of minority class cases. One of the real explanations for this is the majority of the classification algorithms deal with the objective of enhancement of accuracy. As the majority class instances are much higher in number than the minority class ones, the classifier would give high accuracy, even if it classifies all instances as majority class and misclassifies all the minority class instances. This is called class imbalance problem. Besides the imbalanced datasets, other data intrinsic characteristics like overlapping between classes, presence of small disjuncts and lack of density of the minority class in training datasets could also impact the performance of the classifier significantly. The issue of class imbalance becomes more serious in the presence of one or more of such data on intrinsic characteristics. A few arrangements have been proposed in the past to manage these issues independently. In this paper, we have proposed a new algorithm, namely, OKC classifier (hybrid of One-class SVM, K-nearest neighbor and CART) to overcome this problem.

A. Imbalanced Datasets

In many real life applications, the situation of imbalanced datasets every now and again shows up. A dataset in which one class extremely outnumbers other can be considered as an imbalanced dataset. The class with moderately less number of cases in a dataset is called 'minority class' and alternate class is called ‘majority class’. The minority class usually represents the most essential idea to be learned, and it is hard to distinguish it since it may be related to huge and remarkable cases, or because the data acquisition of these cases is costly [1-2]. The imbalance of data distribution between different classes is known as between-class imbalance [3]. Such imbalance could be a consequence of the intrinsic nature of the data. For example, in the fraud detection domain, it is difficult to get the data related to the fraudulent transactions than the data that belong to legitimate transactions. Within a class imbalance is said to happen when a class is comprised of various sub-groups and the quantity of cases having a place with each sub-bunch is altogether not quite the same as those of other sub-bunches inside a similar class [4].

B. Class Overlapping

The class overlaps problem appears when a region in data space contains training data from more than one class. In such case, there is no clear partition between various classes causing difficulty in the classification process. The performance of a classifier is extraordinarily influenced when the issue of class overlapping is present along with an imbalance in the dataset. It has been proved that for the datasets that have clean clusters, i.e. no overlapping and are linearly separable, classifier performance on such datasets is not influenced by any degree of imbalance [1, 5]. In other works, it has been proved that if the data in the overlapping region are imbalanced, then the imbalance ratio affects the performance more than the size of overlap [1].

C. Lack of Density

The issue of lack of density emerges when there is almost no information accessible to represent the minority class concept. In the event that the cases of the minority class are less, then it becomes difficult to distinguish between minority class and noise. The majority of standard classifiers aim to

obtain a good generalization capability. In case of lack of density of a minority class, the classification rules that predict the minority class are highly specialized whereas due to the large number of majority class cases, the classification rules that predict the majority class seem to be more general to the classifier as their coverage is very high as compared to the minority class ones [6]. So, in this case the rules that predict the minority class are discarded by the classifier leading to high misclassification of a minority class.

II. Background

Verma et al. [7] used median filter, Gaussian filter and un-sharp masking J for the image enhancement. Entropy based segmentation is used to find the region of interest and then KNN and SVM classification techniques for the analysis of kidney stone images. The accuracy of KNN was found 89% and that of SVM was 84%. Li and Wang [8] used SIFT (Scale-invariant feature transform) algorithm to extract feature and the extracted features are clustered by K-means clustering algorithm. After clustering BOW (bag of word) of each image is constructed and multi-class classifier is trained using SVM (Support Vector Machine) to classify images. Authors reviled that SVM gave better results in small sample training set. Accuracy of image classification was about 90% with this method. Guo et al. [9] proposed SVM-based sequential classifier training (SCT-SVM) approach for remote sensing image classification. This technique help in reducing required number of training samples for classifier training. Different experiments were conducted with Sentinel-2A multitemporal data and accuracy of 76.18% to 94.02% achieved with the proposed technique.

McDermott et al. [10] in this study investigate Support Vector Machine (SVM) classifiers for detecting brain hemorrhages using Electrical Impedance Tomography (EIT) measurement frames. A 2-layer model of the head with series of hemorrhages is designed by means of numerical models and physical phantoms. Authors reported that phantom models are more challenging with maximum specificity of 75% when used with the linear SVM. The detection are was increased when radial basis function (RBF) SVM classifier and a neural network classifier were applied. Badgujar and Deore [11] proposed a hybrid algorithm using Migrating Bird Optimization and Support Vector Machine (MB-SVM) classifiers. Gaussian filters are used to eradicate the noise from the fundus retinal image. Experimental validation on a publicly available STARE data-set demonstrates the improved performance of the proposed method over existing method. Ma et al. [12]. Presented weight-KNN the KNN-based model acquire the test image's k-nearest neighbors and get the prediction of the image according to the contribution of its neighbors. Hu et al. [13] combine color, texture and shape feature towards multi-type feature. These features were integrated with k-nearest neighbor classifier. Experiment were conducted on 4500 aerial images and recognition rate of 99% was achieved using this multi-type feature. Gul et al. [14] propose an ensemble of subset of k-NN classifiers, (ESkNN) for classification. Experiments were conducted on benchmark data sets and results are compared with usual k-NN, bagged k-NN, random k-NN, multiple feature subset method, random forest and support vector machines. The proposed ensemble gives better classification performance than the usual k-NN and its ensembles, and performs comparable to random forest and support vector machines. Guo et al. [15] proposed a guided filter-based method and used two fusion methods for spectral and spatial features. Hyperspectral images were classified using SVM. The proposed method were fast in execution and easy to implement.

A. Proposed OKC Classifier

The proposed algorithm is a hybrid of one class SVM, k-Nearest Neighbour and CART (Classification and Regression Tree) algorithms. In this algorithm, Hellinger distance and Gini impurity are used as splitting criteria for choosing the best feature and best value to split, respectively. Hellinger distance has been proved to be skewed insensitive [16] i.e. it is not affected by the situation of class imbalance. On each leaf node of this tree where the illustrations have diverse classes, feature selection is done to choose two features that could best discriminate among the classes and then k-Nearest Neighbours is trained on all examples and one class SVM is trained on the minority class samples. When a new prediction is to be done, it is first classified to the leaf node and then it is categorized as inlier or outlier by the one class SVM. If it is predicted as inlier, it is assigned the minority class otherwise after feature selection it is assigned the class predicted by the k-Nearest Neighbor algorithm with k=1 i.e. the class of its nearest neighbour. This algorithm is designed to handle the class imbalance problem even if other data intrinsic characteristics like class overlap and lack of density is also present. As the feature selection is done at each leaf, only those features that play a significant role in classification are selected. It means that overlapping features will be discarded and thus the class overlapping problem can be handled to a great extent. The feature selection is done by using Hellinger distance [17]. The one class SVM algorithm is trained on the minority class tests at each leaf with mixed samples, so it is ensured that all minority class illustrations are learnt by the classifier.

B. One Class SVM

Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper-plane. The conventional 2-class classifier finds a hyper-plane that isolates one class from another. The one-class SVM finds the hyper-plane that separates all of the in-class points from the origin; it is essentially a two-class SVM where the origin is the only member of the second class. So, basically it separates all the data points from the origin and maximizes the distance from this hyper-plane to the origin. This results in a binary function, which captures regions in the input space and returns +1 in the region capturing the training data point & -1 elsewhere [18].

C. K-Nearest Neighbors

In the K-Nearest Neighbour algorithm, an object is classified by a majority vote of its neighbors, with the object being classified to the class most common among its k nearest

neighbors. If k=1, the object is simply classified to the class of that single nearest neighbour. It is typically in light of the Euclidean separation between a test sample and the specified training samples. For n-dimensional space, the Euclidean distance between two points x and y is calculated as following:-

$$d = \sqrt{\sum_{k=1}^{N}(x_k - y_k)^2}$$

It has been observed that the k-NN algorithm suffer from the curse of dimensionality [19] i.e. it cannot perform well when the number of features of the dataset is large. To deal with this issue, we are doing feature selection to select the best features that could best discriminate among the classes before applying k-NN. This feature selection is done at each leaf, with mixed class samples, separately so that the problem of class overlap could be minimized as different features may be prevalent in different places in the data space.

D. CART

Classification and Regression Tree (CART) is a binary recursive partitioning algorithm that is fit for handling nominal and continues attributes both as targets and predictors [20]. The classification tree is built by recursively splitting parent nodes into two child nodes that have maximum homogeneity. This homogeneity is determined by an impurity function. CART searches through all values of the attributes to find the best value to split. There are several impurity functions like Gini index, Towing splitting rule, etc. The process of splitting is stopped when a node becomes pure. Otherwise, it is repeated until a split result into a child node with less number of observations than a predefined number, or when the change in impurity function is less than the predefined minimum change number. Classification of a new observation is made by assigning the dominating class of the leaf node to which the new observation belongs to. In case of imbalanced datasets, when there is the problem of absolute rarity or lack of density of the minority class, the dominating class at the leaf nodes is usually the majority class. This results into misclassification of the observations that belong to the minority class. To sort out this problem, we are using the One-class SVM and k-NN at the leaf nodes with mixed classes instead of voting. One-class SVM is trained on the minority class, to cover all minority class examples so that the problem of lack of density of minority class can be handled to at least some extent. Then after selecting two features using Hellinger distance, k-NN is trained on all samples of the leaf.

E. Splitting Criteria for OKC Classifier

In the proposed algorithm, the splitting criteria used for the choice of the best features, is Hellinger distance and the criteria used for the selection of best value of the chosen feature is Gini impurity. Hellinger distance is a good criterion to be used with imbalanced datasets as it is not affected by the class distribution skew [16]. Assuming a binary class problem (class + and class-), let x+ be class+ and x- be class-, x_{+j} is the number of positives in bin j and x_{-j} is the number of negatives in bin j. For a feature that has p number of bins, the Hellinger distance is given below:

$$d_H(x_+, x_-) = \sqrt{\sum_{j=1}^{p}\left(\sqrt{\frac{|x_{+j}|}{|x_+|}} - \sqrt{\frac{|x_{-j}|}{|x_-|}}\right)^2}$$

The Hellinger distance for all features is calculated before each split and the feature with maximum Hellinger distance is chosen to split. After that, the choice of best value, of the selected feature, to split is made by using Gini impurity. Gini impurity is the expected error rate if one of the results from a set is randomly applied to one of the items in the set [20]. Gini impurity can be computed by summing the probability of each item being chosen times the probability of a mistake in categorizing that item. To compute Gini Impurity for a set of items, suppose i {1, 2,…m}, and let f_i the fraction of the items labeled with the value i in the set, the Gini impurity as given below:-

$$I_G(f) = \sum_{i=1}^{m} f_i(i - f_i) = \sum_{i=1}^{m}(f_i - f_i^2) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} f_i^2 = 1 - \sum_{i=1}^{m} f_i^2 = \sum_{i \neq k} f_i f_k$$

The value with the lowest Gini impurity is selected for split.

F. Stopping Conditions for OKC Classifier

The process of splitting of nodes is done recursively until some stopping condition is met. In the proposed algorithm, there are three stopping conditions:

1) When the node becomes pure i.e. all samples on that node belongs to a single class.

2) If the change in impurity functions, i.e. Gini index, after splitting, would be less than the predefined minimum value.

3) If the split would result into a child node with less number of samples than the predefined minimum number of samples.

G. Algorithm

Input: A set S of labeled instances, threshold values for minimum number of samples at leaves and minimum change in utility function i.e. Gini Impurity.

Output: A binary tree with class labels and/or one class SVM, list of selected features and k-NN classifiers at leaves.

Step 1 If all samples at the current node have the same labels, assign that label to the current node and return.

Step 2 For each attribute, evaluate the hellinger distance and choose the attribute A with a maximum value of the hellinger distance.

Step 3 For each distinct value of A, evaluate Gini impurity and choose the value V, with the lowest value of Gini Impurity.

Step 4 Evaluate the difference between the utility of the current node and the utility that would result after split is performed on value V of attribute A.

Step 5 If the difference in utility is less than the threshold value or if the split would result into nodes with the less number of samples than the threshold value, fit a one-class SVM on the minority class samples and calculate hellinger distance on all attributes to choose two attributes with highest and the second highest value of hellinger distance. On the chosen attributes fit a k-NN classifier and return.

Step 6 Partition S with value V and attribute A.

For each child node, call the algorithm recursively.

III. Evaluation and Discussion

In this work, we have considered public dataset of five categories, namely, Yeast, CTG, Wilt, Fraud, and Semiconductors. Brief information about these databases also depicted in Table I.

A. Experimental Results

In order to evaluate the performance of the proposed algorithm, we have considered five different public datasets as described in Section 3. These five different datasets are normalized and taken from the UCI repository [21]. The results of the proposed algorithm are compared with standard machine learning algorithms decision tree, neural network, SVM, Naïve Bayes, k-Nearest Neighbors, Naive Bayes tree and CART. The proposed algorithm is also compared against random over-sampling, random under sampling, hybrid over-under sampling and meta-cost techniques applied to all the standard algorithms discussed in this section. In meta-cost, the cost of misclassification of minority class is set to double than the cost of misclassification of the majority class. The results obtained after performing various experiments without sampling, after random under-sampling and after random over-sampling are depicted in Tables II to IV, respectively. Experimental results based on hybrid of random under-sampling and random over-sampling are presented in Table V. In Table VI, we have presented experimental results achieved after setting meta-cost double for misclassification of minority class than the misclassification of the majority class. We have seen that proposed classification algorithm, namely, OKC performs better than existing algorithms.

TABLE I. Datasets used for Experiments

Dataset	Number of Attributes	Size of Training Data		Size of Testing Data	
		Class-I	Class-II	Class-I	Class-II
Yeast	9	20	464	10	199
CTG	22	365	1124	106	531
Wilt	6	74	4265	187	313
Fraud	25	100	600	200	100
Semiconductor	3	76	924	28	539

TABLE II. F-Score (%age) without Sampling

Algorithm	CTG	Wilt	Yeast	Fraud	Semiconductor
Decision Tree	41.38	0	0	0	0
Neural Network	51.41	14.85	18.18	28.33	0
SVM	55.06	0	0	0	0
Naïve Bayes	41.03	3.16	8.42	34.68	0
k-NN	11.64	0	0	7.69	0
Naïve-Bayes Tree	22.47	15.76	0	23.38	0
CART	21.95	4.19	0	0	0
Proposed Algorithm	61.27	59.23	22.9	80.7	21.05

TABLE III. F-Score (%age) under Sampling

Algorithm	CTG	Wilt	Yeast	Fraud	Semiconductor
Decision Tree	22.1	41.41	9.04	63.8	9.24
Neural Network	43.24	35.65	9	61.4	7.9
SVM	49.83	20.85	8.99	57.62	6.4
Naïve Bayes	49.04	40.33	8.7	65.59	7.35
k-NN	19.59	55.42	9.57	59.42	7.35
Naïve-Bayes Tree	46.46	37.68	8.29	69.57	0
CART	36.96	44.11	9.04	69.32	8.7
Proposed Algorithm	61.27	59.23	22.9	80.7	21.05

TABLE IV. F-Score (%age) over Sampling

Algorithm	CTG	Wilt	Yeast	Fraud	Semiconductor
Decision Tree	41.84	28.05	10.22	79.41	8.81
Neural Network	49.83	26.05	7.92	36.65	9.9
SVM	52.17	7.14	8.75	65	7.16
Naïve Bayes	40.82	45.85	8.7	72.05	7.23
k-NN	22.5	23.26	10.53	54.42	12.95
Naïve-Bayes Tree	21.43	10.1	8	30.95	8.76
CART	44.02	29.46	8.22	46.59	12.77
Proposed Algorithm	61.27	59.23	22.9	80.7	21.05

TABLE V. F-Score (%age) after Hybrid Under-Sampling and Over-Sampling

Algorithm	CTG	Wilt	Yeast	Fraud	Semiconductor
Decision Tree	32.21	37.66	10.69	61.16	9.51
Neural Network	51.27	19.23	5.37	50.87	9.63
SVM	50.67	11.94	7.82	60.32	6.73
Naïve Bayes	45.42	45.96	7.95	64.42	7.43
k-NN	22.32	57.33	8.39	56.19	10.77
Naïve-Bayes Tree	27.35	33.62	6.50	52.20	6.45
CART	51.33	50.97	14.46	53.64	9.21
Proposed Algorithm	61.27	59.23	22.90	80.70	21.05

TABLE VI. F-SCORE (%AGE) AFTER SETTING META-COST

Algorithm	CTG	Wilt	Yeast	Fraud	Semiconductor
Decision Tree	14.94	0	0	8.92	0
Neural Network	50.21	22.75	0	51.45	0
SVM	52.31	0	9.31	0	0
Naïve Bayes	49.8	15.32	0	36.84	0
k-NN	27.03	7.14	0	12.88	3.57
Naïve-Bayes Tree	25.47	4.17	20	22.16	0
CART	20.65	5.18	0	45.54	0
Proposed Algorithm	61.27	59.23	22.9	80.7	21.05

IV. CONCLUSION

In this paper, a new classification algorithm based on a hybrid combination of one class SVM, k-NN and CART algorithms has been proposed. This algorithm is outlined to such an extent that it could perform well in classification of imbalanced datasets that are non-linearly separable without any need of resampling. Also, it can deal with the circumstances of class overlap and lack of density of the minority class in imbalanced datasets. Our experiments have shown that the proposed algorithm could outperform a number of standard classification algorithms. However, this work is focused only on the binary classification tasks. The task of multiclass classification in the presence of class overlaps, lack of density of the minority class in imbalanced datasets is left for future scope.

REFERENCES

[1] Asuncion A and Newman D (2007) UCI machine learning repository. http://archive.ics.uci.edu/ml/datasets.html.

[2] Cieslak DA, Hoens TR, Chawla NV, and Kegelmeyer WP (2012) Hellinger distance decision trees are robust and skew-insensitive. Data Mining and Knowledge Discovery, 24(1):136-158.

[3] Haibo He and Garcia E (2009) Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9):1263-1284.

[4] Japkowicz N (2001) Concept-Learning in the presence of Between-class and Within-class Imbalances. Advances in Artificial Intelligence, 67-77.

[5] Kouiroukidis N and Evangelidis G (2011) The effects of dimensionality curse in high dimensional k-nn search. In the proceedings of the 15th Panhellenic Conference on Informatics, 41-45.

[6] López V, Fernández A, García S, Palade V, and Herrera F (2013) An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250:113-141.

[7] Verma J, Nath M, Tripathi P and Saini KK (2017) Analysis and identification of kidney stone using Kthnearest neighbour (KNN) and support vector machine (SVM) classification techniques. Pattern Recognition and Image Analysis, 27:574. https://doi.org/10.1134/S1054661817030294.

[8] Li Q and Wang X (2018) Image Classification Based on SIFT and SVM. IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), 762-765.

[9] Guo Y, Yin X, Zhao X, Yang D and Bai Y (2019) Wireless Com Network. EURASIP Journal on Wireless Communications and Networking. https://doi.org/10.1186/s13638-019-1346-z.

[10] McDermott B, O'Halloran M, Porter E, Santorelli A (2018) Brain haemorrhage detection using a SVM classifier with electrical impedance tomography measurement frames. PLoS ONE 13(7):e0200469. https://doi.org/10.1371/journal.pone.0200469.

[11] Badgujar R and Deore P (2018) MBO-SVM-based exudate classification in fundus retinal images of diabetic patients. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 1–12.

[12] Ma Y, Xie Q, Liu Y and Xion S (2019) A weighted KNN-based automatic image annotation method. Neural Computing and Applications. https://doi.org/10.1007/s00521-019-04114-y.

[13] Hu G, Yang Z, Zhu M, Li H, and Xiong N (2018) Wireless Com Network. https://doi.org/10.1186/s13638-018-1195-1.

[14] Gul A, Perperoglou A, and Khan Z (2018) Osama Mahmoud Miftahuddin Miftahuddin Werner Adler Berthold Lausen. Advanced Data Analysis and Classification, 12: 827. https://doi.org/10.1007/s11634-015-0227-5.

[15] Guo Y, Jia X and Paull S (2018) Effective Sequential Classifier Training for SVM-Based Multi-temporal Remote Sensing Image Classification, IEEE Transactions on Image Processing, 27(6):3036-3048.

[16] Luengo J, Fernández A, García S, and Herrera F (2011) Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary under sampling. Soft Computing, 15(10):1909-1936.

[17] Prati RC, Batista GE, Monard MC (2004) Class imbalances versus class overlapping: an analysis of a learning system behavior. Advances in Artificial Intelligence, 312-321.

[18] Schölkopf B, Williamson R, Smola A, and Shawe J (1999) Support Vector Method for Novelty Detection. In the proceedings of the 12th International Conference on Neural Information Processing Systems, 12:582-588.

[19] Seragan T (2007) Programming collective intelligence: building smart web2.0 application.

[20] Weiss GM (2005) Mining with rare cases. The Data Mining and Knowledge Discovery Handbook, Springer, 765–776.

[21] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, and McLachlan GJ (2008) Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1):1-37.

Involving American Schools in Enhancing Children's Digital Literacy and Raising Awareness of Risks Associated with Internet Usage

Mohammed Tawfik Hussein[1]
Department of Accounting, Finance and Management Information Systems
Prairie View A and M University, Prairie View
Texas, United States

Reem M. Hussein[2]
Seventh Grade Teacher and Grade Level Administrator
International Leadership of Texas-CS
College Station
Texas, United States

***Abstract*—The purpose of this study is to shine the light on the importance of educating students on digital literacy and netiquette, for technology has become a common denominator in most of our tasks. This study is mostly concerned with involving schools in educating students on this matter since students spend most of their time in schools. The paper expresses the urgency of increasing the dose of digital literacy taught in schools to help raise students' awareness to potential risks the internet has. It breaks down the risks that young users are prone to face as well as ways to safely avoid them. Further, the paper analyzes the state standards practiced in the US. to serve as a wake-up call for schools to work on improving their standards to protect young users from the versatile harms. Therefore, schools are conjured to take on the role of enhancing students' digital literacy and their understanding of the potential risks present online.**

Keywords*—Digital literacy; e-learning; internet risks; online education and safety

I. Introduction

With every decade and every century of time comes unique trends that capture the interest and curiosity of the generation living within that time frame. The trends peculiar to those ages tend to be relevant to the people and thus it occupies most of their daily tasks by default. With the arrival of the 21st century, the most common phenomenon that spread across the world was technology and its various functionalities that mostly includes the internet. The rapid spread of the internet and technology caused drastic changes in the World that made such an epidemic become the peak of interest for researchers. However, less research has been executed on the ramifications of technology and internet surfing etiquette (netiquette) especially for children of all ages. Novice internet users and technology holders, such as students at schools, can be prone to many of the dangers that technology carries within its folds [1]. Therefore, this paper will firstly shine the light on the literature that covers issues related to technological effects on minors. Also, this paper will discuss the risks that students are bound to face as well as methods to possibly minimize students' risks and enable them to surf the internet safely and more responsibly. This research paper is organized as follows: Section II presents the literature search and review used; Section III gives the definition of what is the digital literacy; Section IV covers the importance of teaching student's digital literacy and online safety measures while Section V covers the potential 21st century risks facing young internet users; and Section VI deals with how schools contribute effectively to minimizing students' online risks. In Section VII, conclusion and contributions are discussed, and finally future work is discussed in Section VIII.

II. Literature Review

The Worldwide proliferation of technology that initially started as a trend, is becoming a norm and a must in almost all organizations and for all its users. The internet, specifically, is found to be very effective in many realms, especially in the realm of education [2]. The technological advances are attracting more school boards across the world to adopt its many resources into its curriculum [3]. Consequently, 21st-century education is now majorly defined by students' digital use throughout their school years, and many of its tools are holistically reliant on the internet [4].

With its rapid spread and development, children are easily gaining access to many of its forms at home, at school, and at other different locations [5], which is raising many concerns among scholars. The product so-called "internet" is granting minors access to its many constituents almost for free but without a manual of instructions or precautions that they should take while surfing the web [6]. The dearth of studies concerned with raising awareness on internet safety measures is causing an imbalance leading to children's exposure to many risks without a shield to protect them [7].

While the internet is massively used as an entertaining outlet or an educational media, others use it to target vulnerable users in ways to bully them, mislead them with wrong data, hurt them, or sexually harass them. Unfortunately, these phenomena are part of the package when becoming an internet user and it's widely spreading to where it's part of the internet's culture [8]. Children's scarce knowledge concerning digital literacy is allowing child predators and other adversaries to use these gaps in ways that harm children without them knowing.

According to a learning.com study, among 110,000 students in America who took the learning.com survey, more than %75 of fifth and eighth graders are not equipped with the

proper tools to use the internet safely [9]. Without the right preparations, students are prone to fall into the many black holes that are woven into the internet system.

Therefore, the urge to teach students of all ages about online safety measures simultaneously while using the internet is imperative [10]. The author in [11] suggested that schools and parents need to cooperate in an effort to minimize students' exposure to online risks as well as to increase their digital knowledge so they can become vigilant users.

III. What is Digital Literacy

The term 'digital' encompasses computers, the internet, cell phones and any form of technology used nowadays. Becoming digitally literate means to harness all necessary skills that allows the user to carefully navigate through the internet, respect authorship rules, understand social responsibility and netiquette, and react professionally to any attempts of bullying or harm [12].

IV. The Importance of Teacginh Students Digital Literacy and Online Safety Measures

Due to the continuous advances in technology, new names to describe this era are being created. "In 1998, Don Tapscott talked about the 'Net Generation', in 2001 Marc Prensky coined the term 'digital natives'." [13]. Students' preferences for how they want to learn and preform their daily tasks are redefined and are majorly dependent on technology [5].

The escalating digital use is becoming a norm that requires a high level of understanding and literacy to avoid potential risks. Adversaries are using the unlimited online potential to target those who are digitally illiterate via different outlets including virus infected websites, social media sites, website scams and false advertising, sexual exposing contents, and self-harm-oriented websites.

Studies have shown that the possibility of children staying out of the internet's risks is by having parent, school, and peer supervision. However, The EU Kids Online survey shows that half of children's access is from private places, including their bedrooms, where supervision is not available which makes them prone to many risks [14]. Therefore, the urge to have students educated on digital literacy is highly needed since they are not always surrounded with those who will supervise and protect them. The more children are equipped with the right knowledge to work out any problems or obstacles they may face online, the less others will need to step in to protect them while online.

Furthermore, the 21st century curriculum is based on teaching student's college preparatory skills that involves technology-based tasks. If children were taught the proper ways to use technology and utilize it for beneficial causes, they will grow professionally and become strong candidates for future jobs that are predicted to be holistically reliant on technology [15]. Hence, the crucial role of schools in promoting digitally literate generations [5].

V. Potential 21ST Century Risks Facing Young Internet Users

While the internet exudes many positive traits and benefits, it is also a medium used for many dangerous acts that are set by many parties including hackers, bullies, pedophiles, self-harm advocates, and thieves. The arising threats against young users, specifically, are raising parents' concerns towards their children's exposure to the internet [8].

Among the problems that children are expected to stumble upon is meeting strangers. Some predators use social media and other forms of online communications to trick children into sending pictures, sharing personal information, or arranging meetings [16]. Children can also be exposed to sexual content accidentally or deliberately. As part of the unlimited internet access and the limited policies and rules controlling the giant web, students are most likely to bump into inappropriate and explicit advertisements when surfing untrusted or unprofessional sites.

Furthermore, children who have at least one social media account is prone to cyberbullying [17], which is becoming an alarming epidemic [18]. Students at a young age are still building their personalities and are trying to discover their interests and talents, which puts them in a sensitive stage to where negative comments become vital to their health and self-esteem [19]. In fact, the term cyber bullying developed as a result of adolescents (ages 12-18) profoundly using the social network to negatively shame, embarrass, and break their peer's confidence, and thus ostracize them from their social groups [19].

Moreover, the ubiquitous use of websites is creating a pathway for adversaries to spread phishing threats and hidden computer viruses through websites that are familiar to the user. Those threats tend to be among the most dangerous methods used to obtain personal information as it looks very similar to legitimate websites, applications, and emails [20], without the right training on this issue, children can easily fall into the nets of fraudsters without knowing [21].

Additionally, according to the EU Kids online survey [22-23], among the top activities that children use the internet for is playing games and watching video clips [13], which is also the same methods self- harm advocates use to encourage suicide and other deviant acts. The recent challenge known as the "Momo challenge", which features a female Japanese sculpture with giant eyes and a sharp grin, has been spreading around among children via social media asking children to hurt and kill themselves [24]. Earlier in the year, rumors of a similar game familiar as the 'Blue Whale Challenge' took place in different countries asking students to also harm themselves and execute 50 harming challenges within a period of 50 days [25]. Such games were built on concepts of competitivity, peer pressure and the lack of adult supervision. The key to not get sucked in these games is by being digitally literate enough to know that these games are threats and another form of cyber bullying, that needs to be reported immediately [26].

VI. How can Schools Contribute Effectively in Minimizing Students' Online Risks

According to the National Center on Education and the Economy (NCEE), the average time spent in schools across countries is around 175 to 220 days with an average of 6.8 hours spent daily [27]. These numbers show that students spend a great amount of time in schools every year, which can be very imperative and fundamental to their growth if they are given the right knowledge. Thus, the numerous hours spent in schools play a role in increasing the trust levels between teachers and students. Based on the EU Kids Online Survey [28], when Australian children are bothered by something online, they are most likely to turn to a teacher or a parent for help. Even though this study mostly specifies Australian kids, it is very common for students to address teachers when needing help, especially those they trust [29]. Therefore, teachers' effects on the issue of preventing online risks are imperative and schools should use its platform to raise awareness on the former online risks.

Firstly, schools are encouraged to organize multiple student seminars that addresses online risks and ways to gain immunity against such harms. These seminars will help serve as reminders for students as well as it will keep them up to date with the skyrocketing development of technology and its many flaws and risks.

The technology standards that are practiced in the states following the common core mostly revolves around students' literacy on formulas, functions, spreadsheet knowledge, presentation creation, importing data and exporting data, problem solving specific to technology tools, keyboarding, typing, toolbar functions, plagiarism, and personal data sharing [30]. On the contrary, the common core technology standards briefly focus on the many forms of cyber bullying and how social media is playing a major role in that. It also didn't mention anything about the phishing phenomenon, computer viruses, online pedophiles and websites that seek the attainment of personal information. While it is crucial for students to know about the computers' many functions and benefits, it is also important for them to know about the risks they are highly expected to encounter while using the computer.

While forty-one states adopted the former technological standards, the other nine states left and developed their own standards including technology related standards. After examining each of the nine states' recently adopted standards for the 2019-2020 school year, that are specific to technology, it was found that Oklahoma and Alaska did not mention the 21st century risks talked about above (cyberbullying, phishing scams, virus-based risks, self-harm invoking games and videos, or online offenders…etc.) Rather, its major focus was mainly on proper computer and internet usage through utilizing its hardware and software tools correctly to problem solve and execute research projects successfully. Indiana, and Virginia; however, talked about the major computer rules as well as slightly mentioning the dangers of cyberbullying. The 3-5 grade technology standards found in the Texas Essential Knowledge Standards (TEKS) did not have strands focusing on the 21st century risks, whereas the 6-8 technology TEKS did.

According to the TEA website, the technology TEKS briefly talked about cyber bullying and virus-related risks as most of its strands were dedicated more towards computer usage, keyboarding, and its related tools. Similarly, to the technology TEKS in Texas, Indiana's standards were also brief on cyberbullying while the bulk of its standards focused on functionality and utilization of the internet and the computer. South Carolina and Florida were the only states that talked about three or more of the potential 21st century risks including cyberbullying, phishing, and virus-related risks. The only state that did not have standards specific to technology on its website was Nebraska [31-40], which could either mean that technology might not be a top priority subject or that it might be part of another subject's standards instead of its own.

From the previous information, it can be concluded that technology as a subject is not among the schools' priorities as it's not part of the standardized testing process compared to math, English and science [41], [42]. Despite its insignificance, most of the schools in America have technology standards that address basic knowledge on the use and functionality of advanced devices, such as computers and cell phones, and the internet. The deficiency found in those standards; nonetheless, was that they either lightly focused on the 21st century online risks (cyberbullying, phishing scams, virus-related risks, offenders…etc.) or they did not mention them at all.

Therefore, schools in America are implored to maximize the attention on digital literacy specifically literacy related to the spreading risks of the 21st century. Even though children are not tested on technology, they are highly in need of guidance that teaches them how to roam the internet safely as part of the default trend of owning a device and having access to the internet.

Since schools in America are truly dedicated to preparing students for the future and for real life situations, then they ought to dedicate more time in teaching students technology literacy skills and netiquette because that is the future [43]. Since students as young as seven years old are owning phones nowadays [44], digital literacy and online risks awareness lessons should be boosted in schools from kindergarten all the way to high school.

VII. Conclusions

The 21st century brought the world a double-edged component that is quickly growing to become among the most fundamental components of life. Technology might have not been that important because it was not so widespread as it is currently [45]. The evolving technology that involves the internet, is bringing its users insurmountable benefits as well as potential harm if the user was not vigilant and digitally literate. This paper mostly focuses on young users who are not literate enough to surf the internet freely, yet they do. [23]. Parents normally would supervise their children to make sure they are using the internet safely; however, parents are not always present everywhere the child goes [23]. Children tend to surf the internet in private places where supervision is not present, which can be risky if they are not familiar with potential 21st century risks. Therefore, schools are conjured to take on the role of enhancing students' digital literacy and their understanding of the potential risks present online. In fact, even

though technology is not part of standardized tests, it is among the few skills that students will remember and use in the real world. It is an undeniable fact that technology is becoming part of every institution, organization, restaurant, entertainment place, and home. The importance of intensifying students' dose of digital literacy in schools is discussed and explained as an urgent issue requiring awareness, which achieves the stated objective for this study. A review of the literature reveals an alarming lack of attention to the prevalent threat of low-technology, or low-complexity phishing attacks. Accordingly, here is a primer on the prominent exploit known as phishing, illustration of several cases, and the necessity for organizational and societal education of data users as to appropriate computer hygiene [46].

The unintended consequences facing humans as they attempt to govern the process of artificial intelligence, machine learning, and the impact of billions of sensory devices connected to the Internet is a challenge to all involved [47].

VIII. Future Work

As future work, the authors will select a school for a pilot study to conduct training for teachers and raise technology awareness for students as well as test their readiness for minimizing the risk associated with internet usage and compare it with the other schools that did not get same training and awareness.

Acknowledgment

The authors would like to thank the College of Business at Prairie View A&M University and International Leadership of Texas for the support and providing the healthy environment to conduct and execute this research.

References

[1] Dias, P., Gomes M.J., & Correia, A. P., Disorientation in hypermedia environments: Mechanisms to support navigation. Journal of Educational Computing Research, 20(2), 93–117, 1999.

[2] Liaw, S. S., Information technology and education: Student perceptions of computer and web-based environments. Doctoral dissertation, Seattle Pacific University, 2000.

[3] Trucano, M., A new research hub on the use of technology in education in developing countries. Edutech. Retrieved from https://blogs.worldbank.org/edutech/new-research-hub-use-technology-education-developing-countries, 2019.

[4] Pavlova, M., Roebuck, D., Learning in technology education challenges for the 21st century. Griffith University. Retrieved from https://pdfs.semanticscholar.org/408f/892cc5a8ee1c4fcd1470bda2e520b197a0ee.pdf, 2002.

[5] Livingstone, S., Haddon, L., & Anke, G. (Eds.), Children, risk and safety on the internet Research and policy challenges in comparative perspective. Retrieved from http://ebookcentral.proquest.com, 2012.

[6] Aggarwal, S., Breeden, B., Henry., P., Mulholland, J., in International Federation for information processing. Advances in Digital Forensics Il, eds. Olivier, M.. Shenoi, S., (Boston: Springer). pp. 317-330, 2006.

[7] Valcke, M., Schellens, T., Van Keerand, H., & Gerarts, M., Primary school children's safe and unsafe use of the Internet at home and at school: An exploratory study. Computers in Human Behavior, 23(6), 2838-2850, 2007.

[8] Finkelhor, D. (2011). The internet, youth and the problem of "juvenoia". Crimes against children research center. Retrieved from http://unh.edu/ccrc/pdf/Juvenoia%20paper.pdf, 2011.

[9] Blackwell, H. (2017). More than 75 percent of fifth and eighth graders are non-proficient in 21st century skills, according to learning. com study. Learning.com. Retrieved from https://www.learning.com/press-releases/75-percent-fifth-eighth-graders-non-proficient-21st-century-skills-according-learning-com-study, 2017.

[10] Anastasiades,P.& Vitalaki, E., Promoting internet safety in Greek primary schools: the Teacher's Role. Educational Technology &Society, 14(2), 71-80, 2011.

[11] Dombrowski, S. & Gischlar, K., Keeping children safe on the internet: Guidelines for parents. National Association of school psychologists. Retrieved from file:///C:/Users/reemh/Downloads/Keeping_Children_Safe_on_the_Internet_Gu.pdf, 2007.

[12] Loveless, B., The importance of digital literacy in K-12. Education Corner. Retrieved from https://www.educationcorner.com/importance-digital-literacy-k-12.html, 2019.

[13] Sonck, N., Kuiper, E., De Haan, J., Digital skills in the context of media literacy: From Livingstone, S., Haddon, L., & Anke, G. (Eds.), Children, risk and safety on the internet, 2012.

[14] Helsper, E., Which children are fully online?: from the Livingstone, S., Haddon, L., & Anke, G. (Eds.), Children, risk and safety on the internet : Research and policy challenges in comparative perspective. Retrieved from http://ebookcentral.proquest.com, 2012.

[15] Roteman, D., How technology is destroying jobs. MIT technology review. Retrieved from https://www.technologyreview.com/s/515926/how-technology-is-destroying-jobs/, 2013.

[16] Aggarwal, S., Breeden, B., Henry., P., Mulholland, J., in International Federation for information processing. Advances in Digital Forensics Il, eds. Olivier, M.. Shenoi, S., (Boston: Springer). pp. 317-330, 2006.

[17] Lampert, C., & Donoso, V., Bullying: from the Livingstone, S., Haddon, L., & Anke, G. (Eds.). (2012). Children, risk and safety on the internet : Research and policy challenges in comparative perspective. Retrieved from http://ebookcentral.proquest.com, 2012.

[18] Valeeva, R., Ribakova, L., Bullying in schools: Case study of prevention and pshycho-pedagogical correction. International journal of environment & science education. 11(7), 1603-1617, 2016.

[19] Kane, G. (2013). Psychosocial stages of symbolic action in social media. Thirty fourth international conference on information systems. Retrieved from https://pdfs.semanticscholar.org/7afd/4e6a10850f1d012e783db73791d1bcbccd06.pd, 2013.

[20] Boulton, J., Boulton, L., Camerone, E., Down, J., Hughes, J., Kirkbirde, C., Kirkham, R., Macaulay, P., & Sanders, J., Enhancing primary school children's knowledge of online and risks with the CATZ cooperative cross-age teaching intervention: Results from a pilot study. Cyberpsychology, behavior, and social networking, 19 (10). Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/27732078, 2016.

[21] Last Dräger, E., Gallardo, I., Hartel, P., & Junger, M., How effective is anti-phishing training for children. Usenix association. Retrieved from https://www.usenix.org/system/files/conference/soups2017/soups2017-lastdrager.pdf, 2017.

[22] Haddon, L., Görzig, A., and Ólafsson, K. (2011). Risks and safety on the internet: The perspective of European children. Full Findings. LSE, London: EU Kids Online, 2011.

[23] Green, L., Brady, D., Olafsson, K., Hartley, J., & Lumby, C. A., Risks and safety for Australian children on the internet: full findings from the AU Kids Online survey of 9-16 year olds and their parents. Retrived from https://researchers.mq.edu.au/en/publications/risks-and-safety-for-australian-children-on-the-internet-full-fin, 2011.

[24] Davidson, T. (2018). Momo 'suicide challenge': Deaths of boy, 16, and girl, 12, linked to sick WhatsApp game. Mirror. Retrieved from https://www.mirror.co.uk/news/world-news/momo-suicide-challenge-deaths-boy-13185367, 2018.

[25] Rossow, A., Cyberbullying taken to a whole new level: Enter the 'blue whale challenge'. Forbes. Retrieved from https://www.forbes.com/sites/andrewrossow/2018/02/28/cyberbullying-taken-to-a-whole-new-level-enter-the-blue-whale-challenge/#c143c6d2673e, 2018.

[26] Polizzi,G., The blue whale game paradox, digital literacy and fake news. Parenting for a digital future. Retrieved from http://eprints.lse.ac.uk/79151/1/Parenting%20for%20a%20Digital%20Future%20%E2%80%93%20The%20Blue%20Whale%20game%20paradox%2C%20digital%20literacy%20and%20fake%20news.pdf, 2017.

[27] Craw, J., Statistics of the month: How much time do students spend in school? National Center on education and the economy. Retrieved from http://ncee.org/2018/02/statistic-of-the-month-how-much-time-do-students-spend-in-school/. 2019.

[28] Research and policy challenges in comparative perspective. Retrieved from http://ebookcentral.proquest.com

[29] Da Luz, Fredson Soares dos Reis, The Relationship between Teachers and Students in the Classroom: Communicative Language Teaching Approach and Cooperative Learning Strategy to Improve Learning. Retrieved from https://vc.bridgew.edu/theses/22, 2015.

[30] Common core state standards K-12 technology skills scope and sequence. (n.d.). Retrieved from https://www.fresnounified.org/dept/curr/tech/PublishingImages/K12_Technology_Scope_and_Sequence.pdf, 2007.

[31] Computer standards of learning for Virginia's public schools. Retrieved from http://www.doe.virginia.gov/testing/sol/standards_docs/computer_technology/2012/stds_comptech_6-8.pdf. 2013.

[32] Florida computer science standards, Retrieved from http://www.cpalms.org/Uploads/docs/Standards/computerscience/Science_ComputerScience_Standards.pdf, 2016.

[33] Indiana department of education: Computer science resources. (2018). Retrieved from https://www.doe.in.gov/standards/computer-science-resources-6-8, 2018.

[34] Indiana department of education. Retrieved from https://www.doe.in.gov/standards/computer-science-resources-3-5, 2018.

[35] Indiana k-12 computer science standards. (n.d.). Retrieved from https://www.doe.in.gov/sites/default/files/wf-stem/ind-k-12-computer-science-standards.pdf.

[36] Oklahoma state department of education. (n.d.). Retrieved from https://sde.ok.gov/sites/ok.gov.sde/files/C3%20PASS%20gr6_10teched.pdf.

[37] South Carolina computer science and digital literacy standards. Retrieved from https://ed.sc.gov/scdoe/assets/File/instruction/standards/Computer%20Science/FINAL_South_Carolina_Computer_Science_and_Digital_Literacy_Standards_(SBEApproved050917)063017.pdf, 2017.

[38] Texas Essential Knowledge and Skills. Retrieved from file:///C:/Users/reemh/Downloads/TA_Grades%203-5_1110%20(3).pdf, 2010.

[39] University of Alaska Fairbanks. (n.d). Retrieved from https://www.uaf.edu/files/soe/Elementary/AK-Content-Standards.pdf

[40] Nebraska department of education. Retrieved from https://www.education.ne.gov/contentareastandards/, 2019.

[41] Bochniak, V., Schools fall flat of preparing. students for the future. The clark chronicle. Retrieved from https://clarkchronicle.com/opinion/2018/05/22/schools-fall-flat-of-preparing-students-for-the-future/, 2018.

[42] David, L., Research says high stakes testing narrows the curriculum. Educational leadership 68(6). Retrieved from http://www.ascd.org/publications/educational_leadership/mar11/vol68/num06/High-Stakes_Testing_Narrows_the_Curriculum.aspx, 2011.

[43] Cantor, D., Struggle for the future: Schools lag in preparing students for the age of automation. The 74. Retrieved from https://www.the74million.org/article/struggle-for-the-future-schools-lag-in-preparing-students-for-the-age-of-automation/, 2018.

[44] Chen, B., What's the right age for a child to get a smartphone? The New York Times. Retrieved from https://www.nytimes.com/2016/07/21/technology/personaltech/whats-the-right-age-to-give-a-child-a-smartphone.html, 2016.

[45] Leask, A., Student then students now: What's changed? Enable education. Retrieved from http://enableeducation.com/student-then-and-student-now-whats-changed/, 2013.

[46] Trautman, Lawrence J. and Hussein, Mohammed and Opara, Emmanuel U. and Molesky, Mason and Rahman, Shahedur, Posted: No Phishing Emory Corporate Governance and Accountability Review (2021), Available at SSRN: https://ssrn.com/abstract=3549992 or http://dx.doi.org/10.2139/ssrn.3549992, (March 6, 2020).

[47] Trautman, L. J., Hussein, M. T., Ngamassi, L., & Molesky, M. J. Governance of the Internet of Things (IoT), Jurimetrics,Vol. 60, Issue No. 3, (Spring 2020).

Detecting Spam in Twitter Microblogging Services: A Novel Machine Learning Approach based on Domain Popularity

Khalid Binsaeed[1], Gianluca Stringhini[2]
[1]College of Computer and Information Sciences
King Saud University, Riyadh, KSA
[2]Dept. of Computer Science
University College London, London, UK

Ahmed E. Youssef[3]
College of Computer and Information Sciences
King Saud University, Riyadh, KSA
Dept. of Computers and Systems Engineering, Faculty of Engineering at Helwan, Helwan University, Cairo, Egypt

***Abstract*—Detecting Internet malicious activities has been and continues to be a critical issue that needs to be addressed effectively. This is essential to protect our personal information, computing resources, and financial capitals from unsolicited actions, such as, credential information theft, downloading and installing malware, extortion, etc. The introduction of the social media such as Twitter has given malicious users a new and a promising platform to perform their activities, ranging from a simple spam message to taking a full control over the victim's machine. Twitter revealed that its algorithms for detecting spam are not very effective; most of the trending hashtags include unrelated spam and advertising tweets which indicates that there is a problem with the currently used spam detection framework. This paper proposes a new approach for detecting spam in Twitter microblogging using Machine Learning (ML) techniques and domain popularity services. The proposed approach comprises two main stages: 1) Tweets are collected periodically and filtered by selecting the ones that appear more frequently than a decided threshold in the specified period (i.e. common tweets). Then, an inspection is conducted on the common tweets by checking the associated URL domain with Alexa's top one million globally viewed websites. If a tweet is common on Twitter but does not appear on the top one million globally viewed websites, it is flagged as a potential spam. 2) The second stage kicks in by running ML algorithms on the flagged tweets to extract features that help detect the cluster of spam and prevent it in real-time. The performance of the proposed approach has been evaluated using three most popular classification models (random forest, J48, and Naïve Bayes). For all classifiers, results showed the effectiveness of the proposed method in terms of different performance metrics (e.g. precision, sensitivity, F1-score, accuracy) and using different test scenarios.**

Keywords*—Spam detection; phishing detection; domain popularity; machine learning; Twitter

I. Introduction

Nowadays, the relationship between people and the Internet has changed dramatically; social media and microblogging services have taken an essential part in the way we live. From a statistical point of view Alexa's website of the global top 500 most visited sites has shown that five websites of the top twenty-five websites are related to social media [1]. This fact supports the claim that social media sites are amongst the most visited around the world. The wide spread of social media has attracted spammers and hackers to perform their activities on these platforms, giving them a huge opportunity and an easy way to reach networks of users who are potentially good targets; and due to the openness of the design of social media, users trust each other on their networks even if they are controlled by hackers. Although social media have given spam and phishing the ideal environment to live in, malicious activities were popular before that; their main target back was electronic mails and web services such as forums. However, the peak point was not reached until social media sites were introduced. In [2], the authors gave a reason for that, they mentioned that the built by design trust relationship between users of these services gave more confidence to the user to read and/or click on hyperlinks sent by a friend on that service. This fact is appealing for the attacker as if he controls one victim, his friend list will be likely trusting his messages.

In [3] the author reported that the spam on social media sites has raised by 355% in the first half of 2013, he justified that as "spammers are turning to the fastest growing communication media to circumvent traditional security infrastructures that were used to detect email spam". He also reported the impact of spam as "the impact of social media spam is already significant, it can damage brand appearance and turns fans and followers into foes". These facts motivate the necessity for developing effective algorithms to prevent spam and phishing on microblogging services, and in order to do that, an effective detection method must be placed first, then the prevention could be done. Almost all techniques in the industry relies on detecting before preventing. Section 3 in this paper describes the current methods used for spam detection in detail.

According to [1], Twitter website is now the most popular microblogging service on the Internet. In contrast to other social media services on the Internet, Twitter has shown, since its introduction in 2006, that it can be an appealing service to almost every user of the Internet; it can be a foundation for blogging, socializing, news, political activities, knowledge and/or job hunting. The feature that makes Twitter distinct from other social media sites that provide the same services is the privacy by design. This feature allows users to get all services without being obligated to reveal any information about themselves or having any user following them. This is

given by the nature of the relation between users (unidirectional) that allows the user to follow any other user without being forced to let them follow you back. This nature is interesting to malicious users since it will allow them to spread their malicious content on the network without having to friend a single user; meaning that the other users on the network will still see their tweets without the need to have the attacker follow the victim; for example, by searching of hashtags.

Another important feature of Twitter is the hashtags; a hashtag implies grouping similar tweets together in a way that allows users to browse them based on a specific subject. This will help attackers to get the highest views possible by targeting popular subjects (e.g. sports, politics, gaming, etc.) and tweeting their messages into them. Hashtags will aid in reaching potentially all users of the service, each according to his/her own interest. This is a crucial evolution in the way spam is spreading; the attacker does not even need to know the target address or name.

Due to the aforementioned reasons, Twitter's algorithms for detecting spam are not very effective since most of the trending hashtags include unrelated spam and advertising tweets. This indicates that there is a problem with the currently used spam detection framework. Hence, many researchers are concerned with investigating and solving the problem of detecting/preventing spam and phishing on Twitter platform [16,24,29-40]. This paper introduces a new approach for detecting spam on microblogging services using domain popularity and Machine Learning (ML) algorithms. The proposed approach comprises two stages: 1) tweets are collected periodically and filtered by selecting the tweets that appear with a frequency more than a decided threshold in a specified period; these tweets are called common tweets. After that, an examination is conducted on the common tweets by checking their associated URL domain with Alexa's top one million globally viewed websites. If a tweet is common on Twitter but does not appear on the top one million globally viewed websites (e.g. google.com), it is flagged as a potential spam. 2) The second stage kicks in by running ML algorithms on the flagged messages to extract features that can help detect the cluster of spam and prevent it in real-time. The performance of the proposed approach has been evaluated through extensive experiments using three different classification models (random forest, J48, and Naïve Bayes). For all classifiers, results showed the effectiveness of the proposed method in terms of different performance metrics (e.g. precision, sensitivity, F1-score, accuracy) and using different test scenarios.

The rest of this paper is organized as follows: Section 2 gives an essential background on spam detection and Section 3 reviews the related work. In Section 4, we present the proposed approach in detail. In Section 5, the performance of the proposed approach is evaluated through extensive experiments. In Section 6, we discuss operation and limitations of the proposed approach. Finally, in Section 7, we conclude this work and give future research perspectives.

II. Background

Spam and phishing are now spreading faster than ever which means that all users on the Internet are potential targets. This is true since spam and phishing messages are designed to exploit the trust concept of the system; meaning that they will use genuine techniques (e.g. sending email) to spread across networks. In this section, we give essential background relevant to spam and phishing.

A. Spam

Oxford dictionary [4] defines spam as "Irrelevant or inappropriate messages sent on the Internet to a large number of recipients". From this description, we can realize that the messages are sent in the network to a group of recipients, this means that a single user receiving a spam message is most likely not interested in the message. The objective of spam varies depending on the intention of the spammer. Some spammers intend to spread malware; others use spams to build a botnet; or for other objectives based on the interest of the spammer. However, the largest use for spam is in the advertisement industry. In [5], the authors reported "...we estimate that spammers and spam-advertised merchants collect gross worldwide revenues on the order of $200 million per year…", they proceeded by showing why e-spam advertisement can be profitable, they argued that unlike post mail spam, the cost associated with using technology to spread spam is negligible. Still, this does not excuse the depraved side of spam. Spam still leads to wasting the victim's time or losing productivity of a service (e.g. Twitter hashtags).

The idea of spam is not exclusive to the Internet; spam was used before the creation of the Internet. The network back then was between universities and large government sites and spam was used on those networks. Nonetheless, it was easy to contain and was not problematic at that time. During the 90s the age of the Internet began, it was commercialized and used by the public within their home. In [6], the authors reported, "By the spring of 1996 spam made up a significant portion of the email received by customers of the major Internet service providers…"; since that date spam was recognized by the industry as a problem that need to be solved. Researchers began to develop new ways to deal with spam; for example, Microsoft began developing research to filter spam via machine learning, they found that the spam messages share some similar characteristic and it is possible to detect a spam message from a legitimate message, they were able to eliminate a large portion of junk mail just by observing the mail stream [7]. Although it was not a solution, at the time, this was an achievement.

The ease of performing spam on online social media helped increase its appearances in this platform. Still, this is only one of the many possible reasons for the popularity of spam in social media. In [8] the authors mentioned that spam on social media is highly effective and this attracts spammers. On the other hand, in [2], the authors believe that the abuse of trust between users of the services is the cause for spam. Furthermore, the authors of [9] found after analyzing a group of spam accounts on Twitter that more than half of them were genuine accounts at some point in time and then they were compromised by the spammer. This last finding can be used to

support the one before it (exploitation of trust) since the compromised user accounts will exploit the trust of his/her friends. From the aforementioned discussion, it is not hard to see how and why social media are perfect platform for spam, they are faster, more scalable, and more effective than traditional spam. All of the previous findings are just some of many possible reasons to why spam is popular on social media as opposed to other traditional ways.

B. Phising

Phishing is a part of the social engineering cluster of attacks where the attacker tries to trick the victim into stealing their sensitive information by sending a message pretending to be a legitimate entity. There is a variety of phishing techniques that can be done through email, SMS, or using fake websites. Phishing can also come in a verity of types, for instance, if the attack is directed to a specific person it is called spear phishing. Nonetheless, for the purpose of this work the term phishing will always denote the general type of phishing.

In [10], the authors reported that 5% of the attackers are successful in convincing their victims. Two years later another group of researchers conducted an in-depth study on phishing [11], they used 20 websites and brought 22 participants, they started asking the participants which of the 20 websites is fake. As expected, 90% of the participants failed to identify the phishing websites from the legitimate ones. The previous finding shows that phishing can be a strong attack if done correctly, thus it can be used to steal sensitive information from ignorant users of any service. This raises the question on how can one know that a website is trustable? This can be answered by answering the opposite question, what makes a fake website trustable.

The authors of [11] answered the later question, they said "Successful phishers must not only present a high credibility web presence to their victims; but they must also create a presence that is so impressive which causes the victim to fail to recognize security measures installed in web browsers". Hence, the presence of the website is the main influence in the success of the phishing attempt. In our opinion, what makes phishing a dangerous attack is the fact that it allows the attacker to penetrate a system without going through the normal defenses.

III. Related Work

As shown above, social media has become an important platform for cyber criminals. Over the years, researchers and scientists have studied spam and phishing attacks to develop ways that will help in detecting and preventing them. None of the current techniques guarantee its results; however, some of them have achieved a tolerable percentage so that it can be cost effective to use. There is an important relationship between detecting spams and preventing them. In [2] the authors inferred that you cannot have prevention without detection by saying "Detecting spam is the first and very critical step in the battle of fighting spam". In [12], the authors mentioned that the length of a false URL differs from the normal one and, thus, it is possible to distinguish fake URLs from the trusted ones. Moreover, in their study on the behavior of the attackers, they found that they usually misuse the webhosting services (mostly free). In addition, they claimed that the domains that become active immediately after registration is most likely associated with phishing purposes. Finally, they mentioned that it is a fact that the machines hosting the phishing domains are distributed across different countries, this proves that botnets are used in phishing attacks.

In [13], an experiment on Twitter hashtags was conducted; the authors created a hashtag on Twitter and monitored the users using it. Their observation showed that after a hashtag becomes popular spammers start using it. Furthermore, they established some features to distinguish spam accounts from genuine accounts. They claimed that the frequency of tweets between the two groups are different, as the spammers tweet with higher frequency than the legitimate users with a mean of 8.66 Tweets Per Day (TPD). On the other hand, the legitimate user tweets with a frequency of 6.7 TPD. Another feature that they found is the friend to follower ratio; they claim that the legitimate user has a higher ratio than a spammer.

In [14], a new way of detecting spam was introduced by the authors, they created 900 user accounts on three different social media websites (Twitter was one of them). They called the newly created group honey-profiles, from that point they started to log all activities in the accounts being either malicious or legitimate for a year. They stated, "Even if friend requests are unsolicited, they are not always the result of spammers who reach out. In particular, many social network users aim to increase their popularity by adding as friend's people they do not know". Later, they started analyzing the spam on the account and came to interesting findings, they found that the level of activity differs between spammers. They, then, categorized them into four groups: displayer, bragger, whisperer, and poster. The poster showed that it is the most effective out of the four and the displayer was the least effective. Furthermore, the authors built a tool to detect the spam activity on Twitter by working more on their insights. They focused on two groups, the bragger and the poster, as they claim that they do not require genuine profiles for detection. The first strategy is called FF-ratio; it works by comparing the number of friends the user has and the number of friend request sent by him/her.

This can be considered as a variation of the technique introduced by the authors of [13] where they compare the friend to follower ratio, but the focus here is on the request sent not to the friends the users already have. The paper also studied the similarity between messages, where they say that it is possible to detect a bot user from a legitimate human user based solely on the URLs on the message. In addition, they addressed another technique for detecting spam or phishing bots by comparing the number of friends and the messages sent. The authors finalized their work by using machine learning techniques to extract features between the spam/phishing accounts that allowed them to detect spam and phishing in real time on Twitter.

The problem with this work is that the speed of the process is not fast enough since Twitter limits the machine to only run 20,000 API calls per hour. To solve the issue, they decided to get assistance from the users of Twitter by providing them with the ability to flag (mark) tweets as spam then execute the classifier on the profiles. The advantage of this technique is

that it saves time, meaning that if the spams get more inelegant, we do not need to find an alternative way; instead we can retrain the data and get even stronger detection.

In [15], the authors proposed a scheme for detecting spam; the paper was solely focused on Twitter. The authors claimed that it is possible to differentiate spam/phishing tweets from legitimate tweets in two different ways: i) account feature-based relations and ii) message feature-based scheme. This means that they rely heavily on the features they learn from existing spam. However, all these schemes are time and resource consuming; spam is a moving target and difficult to measure.

The authors of [2] introduced a new perspective for detecting spam/phishing on Twitter since their approach takes into consideration the performance factor. They elaborated on [14] by commenting that it can barely reach the near real time requirements by Twitter. The authors continued by reporting that with the increased popularity of Twitter the traditional ways that were used before the age of social media is not effective and should not be used anymore. They thought that detecting spam is not an achievement if you do not have acceptable performance rate in the system. This takes into consideration the plea of near real-time delivery where traditional techniques will consume too much computational power and will not be able to meet the time requirement. They continued describing their new approach by saying "Our work shifts the perspective from individual detection to collective detection and focuses on detecting spam campaigns". This will increase performance dramatically since the focus will be shifted to a cluster of tweet as opposed to one tweet at a time. They proceeded on efficiency by claiming that their approach clusters related spam accounts into a campaign and generates a signature for the spammer behind the campaign. Thus, not only their work can detect multiple existing spam accounts at a given time, but it can also capture future ones if the spammer maintains the same spamming strategies. And in regard to robustness, they reported, "Twitter defines the behavior of posting duplicate content over multiple accounts as spamming. By grouping related accounts, our work can detect such a collective spamming behavior". In our opinion, focusing on the group level is a brilliant idea and should in theory increase the performance of any given system and increase the speed of detect/prevent since spam shares some common characteristics and the future detection feature that they introduced. This makes sense because spammers promote their content in large scale campaigns as [17, 18] described.

The effectiveness of a web spam will increase if the domain associated with it is more popular around the web in particular search engines, as they are the root of finding websites on the Internet. In 2007, Microsoft started a research project to investigate web spam. They described web spam in [19] as "...Web spam refers to pages that use techniques to mislead search engines into assigning them higher rank..." from this definition, we can see that spam is giving itself more undeserved popularity to gain as much visits as possible. They found that the construction of the dataset is crucial to improve accuracy of spam classification. This relates heavily to the idea of this paper, as if the spammer on Twitter could perform web spam to increase the popularity of its domain on the web, this might earn him/her a spot on Alexa's most visited websites worldwide. In this case, the detection algorithm will skip him/her since it is not a suspicious website anymore. Microsoft continued by categorizing the methods of increasing popularity "…There are numerous ways to improve a site's ranking, which may be broadly categorized as ethical, or white-hat, and less ethical, or grey-hat (or black-hat), SEO techniques…". The ethical techniques are not harmful; in fact, they might improve the sites content; the most harmful category are unethical ways. The authors of [20] talked in-depth about web spam and described several techniques organizing them into taxonomy, most importantly they concluded their paper with the fact that their taxonomy leads to some techniques that could be used by the search engine providers to fight web spam.

One year after the launch in 2006, Twitter had pushed its first update in the battle of fighting spam. They announced in [21] the start of the new admin tool as they called it; it was designed to help the support staff in dealing with spam accounts after they are detected by suspending them. In addition, they introduced the community powered alerts to help the administrators identify spam account blocked by users and suspend them. Then, Twitter hired a detected staff to deal only with spam problems. Before this update, Twitter had no spam counter measures at all.

After one year from the first spam related update, a new update was pushed. This time Twitter realized that no one could detect spam as humans, so they allowed the users to help in the process of detection by flagging tweets as spam. In [22], they reported "Today we've added another tool to our spam fighting toolbox that will give users the ability to flag bad accounts on Twitter". This update was a huge step forward to how they deal with spam. Now, if the spam filter failed due to the sophistication of the spam, normal users will still act as a defense and will report the account for the admin to take action. After that, the spam can be fed to the detection system to increase its accuracy for future detections.

In 2010, Twitter started to take action against phishing attacks. They noticed that phishing is becoming more popular on the service and that there is exploitation to the trust relation in the Direct Message (DM) feature. Based on that, they were obligated to release an update that will deal with the issue. In [23] they announced that the DM system is being redesigned in a way that allows users to send/receive DMs from users they follow. They believe that this approach should reduce the number of phishing attacks.

Most lately, in continuing their fight with spam/phishing, Twitter announced in [25] the system of bot maker. It was designed to achieve the following objectives: i) preventing spam content from being created; ii) reducing visibility time of spam in Twitter; iii) reducing reaction time to new spams. The system works as follows, the distributed systems feeds events to the bot maker, and the bot maker goes through the content over a set of rules then act accordingly. The rules are grouped into two parts, condition and action. The conditions are placed to help in deciding whether it is a spam or not, while the action is what will follow the condition if it is met. In their study, the service had 40% less spam since the launch of the bot maker. Ideally spam should be detected at real-time or near real-time,

however, in reality this is hard to achieve because of performance issues. The cleverness that went into the design of the bot maker is that it consists of multiple stages. The first stage is the real-time stage that should provide the system with the capability to detect spam on run time; mechanisms like CAPTCHA are placed at this stage. The second stage is the near real-time, when the first stage fails the second stage kicks in, ML is a key concept in this stage to train and classify the objects on the system. The final stage is the periodic jobs stage, this stage consists of a model that extracts features and similarities between user accounts by evaluating the user's activities over a period of time, this stage can be run off line.

To sum up, spam is a real issue that affects the user experience in social media and there are multiple research papers [26-40] aimed to fight the existence of spam. Many of them focus on social media as a broad category and since Twitter is considered a microblogging service with different user requirements, this broad category of research does not always fit to Twitter. To be as precise as possible, we have focused as much as we can on the papers that explicitly mentioned Twitter as a service. Overall, the draw-back of the current literature are usually one of two, either it is not accurate enough, or it is not fast enough. The proposed work aims to provide a solution that is accurate and fast enough to be used in near-real-time application.

IV. Outline of the Proposed Approach

The main objective of this work is to introduce and develop a new model for detecting/preventing spam messages in near real-time. The proposed approach focuses on filtering and flagging tweets based on domain popularity then analyze them using ML algorithms to extract features that can help in future spam detection. Our goal is to detect spam messages that could lead to further damage, not just general spam messages. The focus of the work will be on the URLs associated with the message itself since it is the most common way to spread malicious content on the Internet. As shown in Fig. 1, the proposed approach comprises the following phases:

1) Collection phase: Collecting tweets periodically; (e.g. tweets in one hour).

2) Filtering phase: Selecting the tweets that appears with a frequency more than a predefined threshold.

3) Flagging phase: Examining selected tweets via popular domains on the web and flag the potential spams.

4) Feature extraction phase: Running ML algorithms on flagged tweets to extract the features that could be useful in detecting spam tweets.

5) Detection phase: Detect spam in real time using the features learned by the ML algorithm.

In periods, tweets will be collected and filtered by selecting the tweets that appear more than the decided threshold in the specified period (i.e., common tweets). After that, an examination will be conducted on the common tweets by checking the associated URL domain with Alexa's top one million globally viewed websites. The assumption is, if a tweet is common on Twitter and does not appear on the top one million globally viewed websites (e.g. google.com), it will be flagged as a potential spam. Thus, the common tweets on Twitter, but not on Alexa's will be flagged as potential spam message. Furthermore, the proposed model is reinforced by ML techniques for feature extraction to increase detection accuracy. Therefore, after flagging the potential spam messages, ML algorithms will be run on the flagged messages to extract features that help identify the cluster of spam in the future and prevent it in real-time.

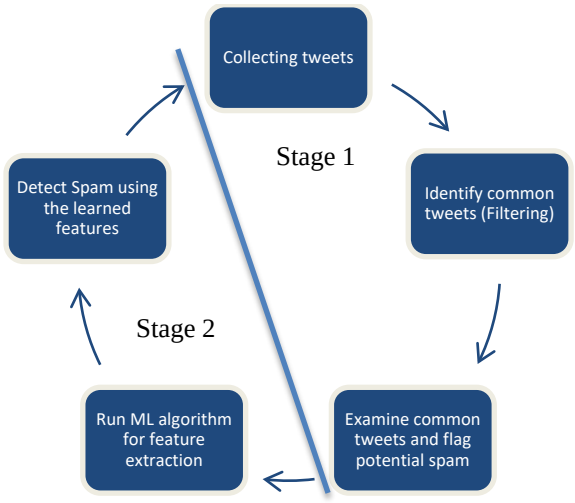


Fig. 1. The Proposed Model Outline.

V. Performance Evaluation

In this section, we describe in detail the proposed detection model, the datasets used to develop it, and the set of experiments conducted to validate it and evaluate its performance.

A. Collection Phase

We have collected a dataset of 27 days' (42TB) worth of tweets with a total of 268,921,568 records each record represents 1 tweet. This can be considered as the initial dataset and will be referred to as dataset1. Moreover, Spamhause is a company that collects and releases a list of confirmed spam domains, these domains will be helpful in detecting some of the false negatives results.

B. Filtering Phase

After collecting tweets and constructing dataset1 (step 1), the proposed model will need to filter the collected data to have a training set for the ML algorithm. The second step is to decide a threshold for the frequency of the tweets in the one-hour period; the tweets in dataset1 that have frequency exceeding this threshold are selected for popular domain test. These tweets are called common tweets. The initial value for the threshold was 120 tweets/hour (2 tweets per min); the test started with this value. Later, after few iterations of the process; manual analysis of the results showed that this value seems a bit low as the percentage of the domains that showed a frequency between 120 and 186 was benign with a 67.8%, thus the threshold was increased to 200 (3.33 tweets per min). The message that have frequency 200 or more are gathered in dataset2 and any spam messages that has a frequency bellow 200 will not be included. At the end, dataset2 had a size of 75,678,885 common tweets; among them are 19,658,349 are actual spam and 56,020,162 are not spam.

C. Flagging Phase

In the third step, the common tweets (i.e. those appear with a frequency 200 or more in a period of 1 hour) are tested via popular domains on the web; a common tweet is flagged as a potential spam if it does not appear on the top one million globally viewed websites. After applying this rule, 23,026,928 tweet messages were extracted from dataset2 in a list of 1131 distinct domain, this dataset will be referred to as dataset3. The distinct domains have been tested manually; we have visited each and every domain using a virtual machine to protect our own systems.

The confusion matrix after applying phase 3 (i.e., at the end of stage 1) is shown in Table I. In order to evaluate the performance of the first stage; we have used the performance metrics expressed in Equations 1-4. The precision was valued at 84.5% and the sensitivity is at 99%. Even though the precision is quite low, it is still incredibly good for the first stage. The performance values of stage 1 are shown in Table II.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Sensitivity\ (recall) = \frac{TP}{TP+FN} \quad (2)$$

$$F1-score = 2 \times \frac{Precision \times Sensitivity}{Precision+Senstivity} \quad (3)$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative.

TABLE I. CONFUSION MATRIX PRODUCED BY STAGE 1

	Actual Spam (+) 19,658,349	Actual Not Spam (-) 56,020,536
Flagged as Spam (+) 23,026,928	TP 19,461,766	FP 3,565,162
Flagged as Not Spam (-) 52,651,957	FN 196,583	TN 52,455,374

TABLE II. PERFORMANCE EVALUATION AFTER STAGE 1

Precision	Sensitivity	F1-Score	Accuracy
84.5%	99%	91%	95%

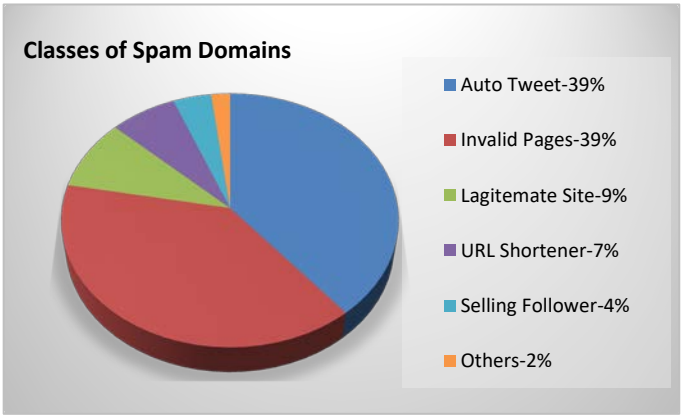


Fig. 2. Spam Domain Classification.

Fig. 2 shows the spam domain classes; we can see that highest percentage is for the invalid pages (i.e. pages that are not working anymore). This shows that spammers create their website for a specific purpose and then dispose them after it completes its objective. Another large percentage was for the tweeting services, many users on Twitter use such services to auto tweet some content they want on their timeline at specific periods.

D. Feature Extraction Phase

To extract spam features, we need to find similarities between users of the social media websites. First, we need to extract features from each user of the service; according to [26] there are two categories of features: user-based features where the focus is on the relations of the user (e.g. followers/friends) and content-based features where the focus is on the content that the users post like tweet messages. Three different classifiers (Random forest, J48 and Naïve Bayes) implemented in Waikato Environment for Knowledge Analysis (Weka), were employed in stage 2 to ensure the accuracy of the results. Weka is free software licensed under the GNU General Public License and developed at the University of Waikato, New Zealand.

1) User-based features

a) User reputation: The first feature studied is user reputation, the authors of [26] commented on user reputation by saying "Spam accounts try to follow large number of users to gain their attention". In [27] the author mentioned that spam accounts have the tendency to follow a large amount of users to gain attention. Thus, he created the following formula to calculate the reputation, R(vi), of a single user (vi),

$$R(vi) = \frac{f(vi)}{f(vi)+o(vi)} \quad (5)$$

where f(vi) the number of friends and o(vi) the number of followers. From this formula, we can see that if the number of friends is small compared to the number of followers then the reputation will be low (i.e. close to zero) and according to the author these accounts have a high probability of being spam. The reputation study showed that auto tweeting services cannot be classified as spam even though, in principal, they could satisfy the definition of spam given in section 1 as they might be an irrelative message sent automatically; the users of these services have a high reputation score with a mean of 0.47rep. The reputation study also confirmed that the users that tweet outdated URLs are defiantly spammers; the top 3 invalid domains have scored a low mean of 0.19rep; it involved 983,273 users. Finally, the reputation also confirmed that shortened URL in general cannot be used to classify any group of users since it is used by all. Hence, the reputation helped in finding some of the domains that are used heavily by spammers due to the low reputation score such as changerion.inf and coconut.chips.jp scoring 0.21 and 0.29 respectively. In general, we think that detecting spam using reputation is an outdated technique because of shortened URL and auto tweeting services that dominate the messages on the social media.

b) User followers and friends count: In the previous feature (reputation) the focus was on the percentage given by the reputation formula, so two users could have the same reputation, but this does not necessarily mean that they have the same followers and friends count. The number of followers and friends is an important feature to distinguish different clusters of users. In order to study this feature, a random training set was chosen from the database with a size of 21K record (tweet). The legitimate users scored a mean of 253,032.7 followers and 4244.993 friends count. This makes sense due to the fact that active users and well-known icons like celebrities will have a high count of followers and a low count of friends. On the other hand, spammer have shown that they have a low mean number of followers compared to the legitimate users scoring; a mean of 4429.76 follower count and 3592.11 friend count. This comes from the fact that users usually follow back the user who followers them, so spammers exploit this habit by sending a flood of friend requests to a group of user account in the hope that some of them follow back.

c) User verification: Twitter provides a service of verifying known users such as celebrities, this feature could be used for detecting spam account since verified users are most likely legitimate users. A random set of 17171 tweets was chosen from the database as a training set, after classifying the dataset using the three different classifiers the result was as shown in Table III.

All the three classifiers (Random forest, Naïve Bayes and J48) gave the same results. The classification result shows that verified account are usually not spammers with a probability of 99.5% (i.e. if an account is verified by Twitter it has a 99.5% chance of not being a spam account). We can see that this feature is useful in detecting if an account is not spam (verified), but not the other way around. It is important to note that if the account is not verified, this does not mean it is a spam account as the results have shown 50.1% chance for this, this means that in order to detect spam we will need to add more features. This can be useful as a first filter after the flagging (Alexa comparison) to eliminate the accounts that are not relevant.

d) User listed count: In Twitter, each user has several public lists that he/she is a member of. By studying spammers on Twitter and visiting their pages, we have noticed that most of them are listed in different kinds of lists; most of them are in advertisement groups. However, for the purposes of this research, we believe that using the count of how many times a spammer has been listed is more accurate than checking the actual list itself due to the fact that the lists don't have a standard naming system which can make each list unique. A random training set of 17540 labelled tweets was selected from the database to test the validity of the feature. The feature did prove as a useful feature for detecting spammers. The J48 classifier was able to distinguish spammers with the statistics shown in Tables IV and V.

Like the verified feature, the count of lists can be used to find spammers, but not the other way around, the classifier has classified 7288 tweets correctly as spam (TP) and 1810 tweet classified wrong (FP). Even though the performance is not high it is still an acceptable feature and can be added to the overall classifier.

e) User statuses count: Every post on Twitter is counted as a status of the user, this means that if a user tweets or retweets or even replies publicly to another user, the counter will count every instance. Obviously, spammers will have high statuses associated with their accounts. Hence, old and active user accounts (aka veterans) will still have a high count as well, so this feature needs to be tested by a classifier to check if it is an acceptable feature. A random training set of 26986 tweets was selected from the database for testing feature by classifiers. The first classifier (Random forest) gave expected results with good statistics; this is shown in Tables VI and VII.

The second classifier, J48, classified 12939 tweets as a spam correctly (TP) and only 1033 was classified as spam wrongly (FP). This shows that the feature (statuses count) could be considered as a strong feature to add to the final classifier. The results given are unexpected as it was stronger in detecting the legitimate tweets rather than the spam tweets. These results are shown in Tables VIII and IX. As for the Naïve Bayes classifier, the results are more in favor of detecting spam tweets not the other way around as shown Tables X and XI. However, it is clearly shown that the feature is unreliable since 88.97% of the results are classified as spam. Finally, this feature has shown verity in the result in all the classifier. Except for Naïve Bayes classifier, the feature is valuable and can be used in the final classifier to distinguish between the two classes.

TABLE III. CONFUSION MATRIX (USER VERIFICATION FEATURE)

	Spam (+)	**Not Spam (-)**
Not verified (+)	TP=5001	FP=4978
Verified (-)	FN=38	TN=7154

TABLE IV. CONFUSION MATRIX BY J48 (LISTED COUNT FEATURE)

	Spam (+)	**Not Spam (-)**
Classified as spam (+)	TP=7288	FP=1810
Classified as not spam (-)	FN=2564	TN=5878

TABLE V. PERFORMANCE EVALUATION BY J48 (LISTED COUNT FEATURE)

Precision	Sensitivity	F-Measure	Accuracy
0.801	0.74	0.769	0.751

TABLE VI. CONFUSION MATRIX BY RANDOM FOREST (STATUSES COUNT FEATURE)

	Spam	***Not spam***
Classified as spam	13321	651
Classified as not spam	2325	10689

TABLE VII. PERFORMANCE EVALUATION BY RANDOM FOREST (STATUSES COUNT FEATURE)

Precision	Sensitivity	F-Measure	Accuracy
0.9534	0.8514	0.8995	0.8897

TABLE VIII. CONFUSION MATRIX BY J48 (STATUSES COUNT FEATURE)

	Spam	***Not spam***
Classified as spam	12939	1033
Classified as not spam	171	12843

TABLE IX. PERFORMANCE EVALUATION BY J48 (STATUSES COUNT FEATURE)

Precision	Sensitivity	F-Measure	Accuracy
0.9261	0.9869	0.9555	0.9553

TABLE X. CONFUSION MATRIX BY NAIVE (STATUSES COUNT FEATURE)

	Spam	***Not spam***
Classified as spam	13397	10612
Classified as not spam	1030	1947

TABLE XI. PERFORMANCE EVALUATION OF NAIVE (STATUSES COUNT FEATURE)

Precision	Sensitivity	F-Measure	Accuracy
0.5579	0.9286	0.6970	0.5685

f) User favorite count: Users of tweets can flag tweets they like as a favorite, this allows users to group the messages they like and view them at any time. This feature does not only benefit the user him/herself, but other users can go into his/her account and check out his favorite list. The user's favorite count can be used as a feature in the classifier to detect spammers. Similar to other features, a proper testing has been conducted on a training set to check if the feature is acceptable in distinguishing spammers from legitimate users. The three classifiers have been used to test the feature. Random forest and J48 classifiers has shown that spammers do not have a favorite list associated with them (i.e. the count of the list is zero), this makes sense since spammers do not care about other tweets and most of them are bots. Our first thought was; this would not be a proper feature since many users do not use the favorite flag at all. However, the classifiers have shown that this is indeed a reliable feature, not for detecting spammers but for detecting legitimate users with a sensitivity of 0.962.

2) Content-based features

a) Number of hashtags: Hashtags are features used by users of Twitter in order to group relevant tweets together. This feature introduces an opportunity for spammers to spread their content to all the users without mentioning them directly. A careful inspection of the dataset has revealed that a high appearance of hashtags is most likely associated with a spam message. A 15K random tweets were plugged in the classifiers, which gave the results listed in Tables XII and XIII.

TABLE XII. CONFUSION MATRIX (NUMBER OF HASHTAGS FEATURE)

	Spam	*Not spam*
Classified as spam	8740	5288
Classified as not spam	9	963

TABLE XIII. PERFORMANCE EVALUATION (NUMBER OF HASHTAGS FEATURE)

Precision	Sensitivity	F-Measure	Accuracy
0.6230	0.9988	0.767	0.4468

The statistical summary above is for the dataset that includes 4 or more hashtags. This shows incredibly good results with a sensitivity of 0.99. Hence, it could be considered as a reliable way for detecting spam accounts, but not the other way around.

b) Number of Mentions: In Twitter the users have the option to mention other users in their tweets by adding the (@) sign before his/her username. Spammers can use this feature to mention as much users as they can to spread their content directly to them. The classifiers have shown similar result to the number of hashtags shown above. The higher the count of mentions the more likely it is a spam; the classification gave sensitivity of 81% for messages that includes four or more mention to be spam.

c) Sensitivity of tweets: The sensitivity field in the tweet record is a Boolean field (true, false) that is only available when a tweet has a link associated with it. Obviously, this makes sensitivity feature seems to be relevant to our work since our main focus is on domain popularity (i.e. all the tested tweets must have domains associated with them). The denotation of this feature does not describe the content of the tweet itself, but in its place, it is used as an indicator that the hyperlink associated with the message may contain content identified as sensitive. This shows that the sensitivity feature may be used for classification because of its relevancy to the main idea. However, after testing a random training set of 26K tweets, the three classifiers have shown that this feature does not help in identifying spam at all. Fig. 3 shows that the sensitivity of the URL has no effect on the message being a spam or a legitimate tweet since spam tweets are not targeting one specific type of content. Thus, this feature has been dropped from the final classifier.

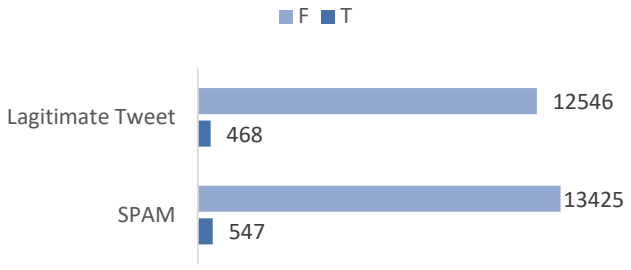


Fig. 3. Tweet Sensitivity.

E. Spam Detction Phase

After selecting the proper features, now it is time to put them together and evaluate the spam detection model. Just like the individual test for features, the same three classifiers (Ransom forest, J84 and Naïve Bayes) were used to evaluate the final spam detection model. To get the highest accuracy possible, two methods for evaluation were considered. The first method (raw record method) involves building a training set of raw tweets; this means that the labelled tweets are taken as they are from the database directly and plugged in the classifier. The second method (grouped record method) groups the tweets that advertise the same domain into one record using mean and standard deviation for each feature. Both methods have the potential to be accurate. In the first method, ML algorithms work better with raw data as they do the calculation and the classification more accurately, while in the second method each set has anomalies that may change the result; by aggregating the anomaly list, anomalies will be removed or have virtually no impact on the dataset.

Each evaluation method involves two test options: 1) in the first option (percentage split test), half of the records (tweets) are used as training set to see how well the classifier can distinguish between them and then the other half are utilized as a test set to examine how well the classifier works on unlabeled data; 2) the second test option (cross-validation test) divides the database into 10 folds, iteratively runs training on 9 folds and leaves one for testing; the test fold is changed in each iteration. This is considered an accurate method because if a fold is used for training it is not used for testing.

1) Evaluation using raw record method: A database of size 26,986 records was constructed to include 50% spam and 50% legitimate tweets, the tweets where selected at random from each type, For the purpose of the first test option (% split) the dataset has been split into two subsets, the first half to be used as a training set and the other half as test set. Next, for the second test option (cross validation) the complete dataset will be plugged into the classifiers with its entirety, the algorithm splits the dataset into ten equal folds then work on them accordingly. The evaluation process is accomplished in 10 iterations; in each iteration nine folds are chosen for training and one is left for testing; the testing fold is changed in each iteration. The records will contain 11 features (user verification, user followers count, user friends count, user reputation, user listed count, user statuses count, user favorites count, user language, tweet language, number of mentions and number of hashtags) plus the label (spam or benign).

a) Percentage split test: The Random Forest classifier was able to perfectly distinguish between the 13,493 instances giving us perfect results for the training dataset as shown in Table XIV. Unlike the random forest classifier, the J48 and Naïve Bayes could not perfectly distinguish between the two classes. However, the result is still on the good side with J48 coming second and Naïve Bayes as the worst out of the three. This shows that the features are in fact good features and they can be used to detect spam messages.

After classifiers learned how to distinguish between spam and legitimate tweets in the training stage, the other half of the dataset (test set) is used to test how well the classifiers can distinguish between unlabeled new data with the model built from the training data in the training stage. The details of the performance are shown in Table XV.

Table XV shows that random forest and the J48 classifiers have built a strong model for detecting spam that could be relied on with 0.981 and 0.946 precision and an accuracy of 92.9% and 92.5%, respectively. On the other hand, the Naive Bayes classifier had the highest and almost perfect precision with 0.988, however, the model also classified falsely nearly half of the classified set which affected its accuracy to be only 76.2%. Thus, the naïve Bayes classifier is not a reliable method of classifying such data and is not recommended to be used.

The confusion matrix in Table XVI explains the above statistics in terms of number of records each classifier has predicted correctly or wrongly. Random forests were able to classify 6877 tweets as spam correctly and only 136 tweets were classified as false positive. On the other hand, the classifier failed to classify 816 tweets that we can call false negatives. Next, the J48 classifier also have some good results, the model classified 6634 tweets correctly as spam and only 379 false positive. In contrast 257 tweets were flagged falsely as legitimate (benign) tweets. Finally, the Naïve Bayes gave a surprising result of 6931 tweets classified correctly as spam and only 82 false positive. Yet, even though Naïve Bayes showed particularly good numbers in detecting spam, it still has a high number of false negatives with 3129 records that should be flagged as spam. Therefore, we can say from the results above, that tree-based classifiers such as J48 and random forest, work very well and are accurate enough to call them valid spam detection techniques.

TABLE XIV. PERFORMANCE EVALUATION FOR SPAM CLASS (TRAINING)

Classifier	Precision	Sensitivity	F-Measure	Accuracy
Random Forest	1	1	1	1
J48	0.972	0.947	0.959	0.969
Naïve Bayes	0.794	0.989	0.880	0.892

TABLE XV. PERFORMANCE EVALUATION FOR SPAM CLASS (TESTING)

Classifier	Precision	Sensitivity	F-Measure	Accuracy
Random Forest	0.981	0.894	0.935	0.929
J48	0.946	0.963	0.954	0.925
Naïve Bayes	0.988	0.689	0.812	0.762

TABLE XVI. CONFUSION MATRIX FOR THE THREE CLASSIFIERS (TESTING)

Classifier		Actual Spam	Actual Legitimate
Random Forest	Classified as spam	***6877***	136
	Classified as legitimate	816	***5664***
J48	Classified as spam	***6634***	379
	Classified as legitimate	257	***6223***
Naïve Bayes	Classified as spam	***6931***	82
	Classified as legitimate	3129	***3351***

TABLE XVII. PERFORMANCE EVALUATION FOR SPAM CLASS (CROSS VALIDATION)

Classifier	Precision	Sensitivity	F-Measure	Accuracy
Random Forest	0.981	0.883	0.929	0.9257
J48	0.954	0.965	0.96	0.9599
Naïve Bayes	0.989	0.689	0.812	0.8407

b) Cross validation test: Before concluding the first method (raw record), another test option will be run to verify and ensure the results. For this option, the same dataset of 26K tweets has been used. The results in Table XVII have shown uniformity with the first option test with the following detailed accuracy. The numbers are almost identical with a slight increase in the cross validation option. This makes sense because in cross validation the 10-fold option increases the variety of the training and test records.

2) Evaluation using grouped record method: In this method, the idea is to group all the tweets that advertise the same domain into one record using the mean and the standard deviation for the numeric fields and count of distinct values for the other types. The dataset contains 630 labelled grouped record (50% spam) each record represents one domain. To create the grouped record, 1000 tweets have been chosen randomly from the tweets that advertise the same domain and the record was built according to the aggregate values of the tweets. Each record will contain 21 features named: count of verified, count of not verified, mean user followers count, STD user followers, mean user friends count, STD user friend, mean reputation, STD reputation, mean user listed count, STD listed, mean user statuses count, STD status, mean user favorites count, STD user favorites, count tweet possibly sensitive, count tweet possibly not sensitive, domain of URL, mean number of mentions, STD mentions, mean number of hashtags and STD hashtags.

a) Percentage split test: Similar to the first method (single record) the random forest was able again to distinguish between the classes perfectly in the training dataset. The J48 has a decrease in sensitivity; only 0.873 and a slight decrease in the precision to 0.975. Finally, the Naïve Bayes has more sensitivity 0.958 than J48 but with the least accuracy 88.91%; this is shown in Table XVIII. With the test dataset, random forest and J48 gave very good accuracy while Naïve Bayes recorded relatively low accuracy as shown in Table XIX. This again shows that tree-based classifiers are accurate enough to call them valid spam detection techniques.

b) Cross validation test: The cross-validation check will be conducted on the entire list which have been plugged into each classifier. The test gave the results shown in Table XX. As expected, and similar to the first method, the results of the cross-validation test are quite similar to the percentage split test.

3) Comparison: Firstly, the two test options (percentage split and cross validation) have shown similar results, which are expected due to the similarity in which how each one of them work as explained previously. However, even though the difference can be considered negligible, we believe that cross validation is the more reliable option to choose in our work.

Secondly, the three classifiers (Random forest, J48 and Naïve Bayes) have been used in a variety of tests and with the same data plugged into them. The tree-based classifiers (Random Forest and J48) have shown better results in this kind of setup with random forest being better in building a solid classification model to distinguish between the classes. While on the other hand, the Bayes based classifier (Naïve) has shown the tendency to group most of the record in one class (usually spam class), this made the classifier unreliable and not recommended to be used with a similar environment.

Lastly, similarly to the two test options, the two method of presenting the data (single and grouped) gave similar results. In the comparison between the two methods, only the random forest classifier will be considered since it has shown that it is the most suitable in this setup. To begin with, both methods scored a perfect score with the training stage meaning that the learning algorithm was able to build a classification model that distinguishes spam and legitimate tweets perfectly using the training data in both methods. However, in the testing stage the single (raw) record approach performed better in terms of all evaluation metrics (precision, sensitivity, f-measure, and accuracy).

To sum up, the proposed approach has been validated through three different classifiers with the random forest classifier being the most reliable one in detecting malicious spam using domain popularity in a micro blogging environment like Twitter. The two evaluation approaches showed very similar results with the single record approach being the favored and more accurate. Cross validation with 10 folds is the most suitable test option for this work. This comparison is shown in Fig. 4.

TABLE XVIII. PERFORMANCE EVALUATION FOR SPAM CLASS (TRAINING - GROUPED)

Classifier	Precision	Sensitivity	F-Measure	Accuracy
Random Forest	1	1	1	1
J48	0.975	0.873	0.921	0.9158
Naïve Bayes	0.801	0.958	0.873	0.8891

TABLE XIX. PERFORMANCE EVALUATION FOR SPAM CLASS (TESTING - GROUPED)

Classifier	Precision	Sensitivity	F-Measure	Accuracy
Random Forest	0.993	0.885	0.936	0.933
J48	0.98	0.903	0.94	0.938
Naïve Bayes	0.798	0.933	0.86	0.869

TABLE XX. DETAILED ACCURACY FOR SPAM CLASS (CROSS VALIDATION - GROUPED)

Classifier	Precision	Sensitivity	F-Measure	Accuracy
Random Forest	0.958	0.876	0.915	0.911
J48	0.962	0.873	0.915	0.911
Naïve Bayes	0.797	0.958	0.87	0.880

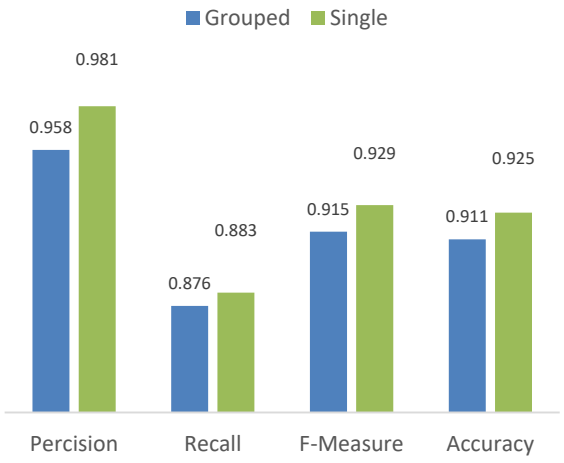


Fig. 4. Comparison between the Single and the Grouped Approach for (Random Forest - Cross Validation).

VI. Discussion

A. Operational Systems Accuracy

In this section, we discuss how Twitter may operate such a system and how accurate will it be in a real operational environment. The operation part will be the same as discussed in this paper by running the system in periodic time and performing the two stages of filtering and evaluation. However, what is not mentioned above is that the accuracy of the system will be much higher than the numbers shown in the evaluation part because of the changes between the testing environment and the real operational environment. Even though the results were good and promising in testing environment, the accuracy will be higher in the actual operation environment since the data considered in this system only referees to tweets with URLs associated with them. Thus, if the whole database was considered in the evaluation part, the numbers will be much more accurate and most likely will jump dramatically since only 8.5% of dataset 1 contains URLs associated with them. This exclusion was a necessary step to increase the accuracy as much as possible making all the records on the proposed system a possible spam message.

B. Limitations

The system may be evaded by some techniques that we can consider as limitations. The first limitation is using URL shortened, this limitation is a very powerful way to make this system useless since the spammer can mask his/her URL into another short URL using URL shorten services. However, it can be easily dealt with if the service that uses this system checks the URLs and gets the complete URL before posting the tweet and saving it as meta data in the record itself. Twitter is not vulnerable for this kind of limitation because they do in fact check the complete path of the URLs. The other limitation is auto tweeting services, in theory, these services are spreading spam since the message is going automatically from the user profile in specific times. However, one could argue that since the user is registered with them and the tweets are not random it is not a spam. The problem is that those tweets will have the domain of the tweeting service which will obviously be a popular domain in Twitter since all the users registered to those services are tweeting their domain, on the other hand, the tweeting service will most likely not be in the Alexa top visited domains. This will make the system flag those service as potential spam even though most of its users are legitimate users.

VII. Conclusions

History has shown that the fight against spam is a cat and mouse game, it is a never-ending battle. Whenever a countermeasure is introduced spammers find a way to evade it. Though, history have also shown that instead of trying to defeat spam entirely we should focus on reducing it to an acceptable rate. In this paper, we have introduced a new way of detecting malicious spam that has never been used before in popular online micro blogging services, focusing mainly on Twitter service. The problem in hand is not to detect spam in general, but to detect malicious spam that could escalate to other type of attacks. Thus, the idea focuses on URLs associated with the messages since they are the most common way used to spread malicious content. Based on Twitter spam policy, the content-based and the user-based features are used in ML algorithm to detect spam messages. This work has added a filtering stage before the ML stage to increase the efficiency and accuracy of detecting such spam type. Furthermore, three different classification algorithms have been studied and used to analyses the data. The results show that filtering the popular domains that appear in Alexa's top one million most visited websites to separate the potential spam before using the ML algorithms is a valid and accurate approach. In Addition, the classifiers were able to identify the similarities between spam messages which allows for future real time detection.

References

[1] "Top Sites," [Online]. Available: http://www.alexa.com/topsites. [Accessed 28/02/2015].

[2] Z. Chu, I. Widjaja and H. Wang, "Detecting Social Spam Campaigns on Twitter," Applied Cryptography and Network Security, Springer Berlin Heidelberg, pp. 455-472, 2012.

[3] H. Nguyen, "State of social media spam," Nexgate, San Francisco, California, 2013.

[4] "Oxford Dictionary," [Online]. Available: http://www.oxforddictionaries.com/us/definition/american_english/spam [Accessed 18/02/2015].

[5] J. M. Rao and D. H. Reiley, "The Economics of Spam," Journal of Economic Perspectives, vol. 26, no. 3, pp. 87-110, 2012.

[6] L. F. Cranor and B. A. LaMacchia, "Spam!," Communications. ACM, vol. 41, no. 8, pp. 74-83, Aug. 1998.

[7] M. Sahami, S. Dumais, D. Heckermany and E. Horvitzy, "A Bayesian Approach to Filtering Junk E-Mail," AAAI-98 Workshop on Learning for Text Categorization, Madison, WI, 1998.

[8] D. Wang, S. B. Navathe, l. liu, D. Irani, A. Tamersoy and C. Pu, "Click Traffic Analysis of Short URL Spam on Twitter," The 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing, Austin, TX, USA, 2013.

[9] C. Griery, K. Thomas, V. Paxson and M. Zhang, "@spam: The Underground on 140 Characters or Less," 17th ACM conference on Computer and communications security, New York, NY, USA, 2010.

[10] R. Basnet, S. Mukkamala and A. H. Sung, "Detection of Phishing Attacks: A Machine Learning approch," in Soft Computing Applications in Industry, Berlin, Germany, Springer Berlin Heidelberg, pp. 373-383, 2008.

[11] R. Dhamija, J. D. Tygar and M. Hearst, "Why Phishing Works," Human Factors in Computing Systems (CHI), Montr´eal, Canada, 2006.

[12] M. D. Kevin and G. Minaxi , “Behind Phishing: An Examination of Phisher Modi Operandi,” San Francisco, CA, 2008.

[13] S. Yardi, D. M. Romero, G. Schoenebeck and d. boyd, “Detecting spam in twitter network,” Firstmonday, vol. 15, no. 1, 2009.

[14] G. Stringhini, C. Kruegel and G. Vigna, “Detecting Spammers on Social Networks,” The 26th Annual Computer Security Applications Conference(ACSAC), New York, NY, USA, 2010.

[15] C. Zi , G. Steven, H. Wang and S. Jajodia, “Who is Tweeting on Twitter: Human, Bot, or Cyborg?,” The 26th Annual Computer Security Applications Conference, New York, NY, USA, 2010.

[16] H. . D. Nannaware., T. . P. Dhangar., A. S. Dhanrao. and P. V. D. Badgujar., “A Run-time Detection System for Malicious URLs in Twitter,” International Journal on Recent and Innovation Trends in Computing and Communication, vol. 3, no. 1, pp. 144 - 147, 2015.

[17] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen and B. Z, “Detecting and Characterizing Social Spam Campaigns,” IMC '10 Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, New York, 2010.

[18] M. Egele , G. Stringhini, C. Kruegel and G. Vigna, “COMPA: Detecting Compromised Accounts on Social Networks,” NDSS symposium, San Diego, CA, USA, 2013.

[19] K. M. Svore, Q. Wu and C.. J. Burges, “Improving web spam classification using rank-time features,” Proceedings of the 3rd international workshop, New york, 2007.

[20] Z. Gyongyi and H. Garcia-Molina, “Web Spam Taxonomy,” First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005), Japan, 2005.

[21] “Turning Up The Heat On Spam,” Twitter, 21 Aug 2008. [Online]. Available: https://blog.twitter.com/2008/turning-heat-spam. [Accessed 2/ 4/2015].

[22] “Help us nail pammers,” Twitter, 13 Oct 2009. [Online]. Available: https://blog.twitter.com/2009/help-us-nail-spammers. [Accessed 2/4/ 2015].

[23] “Avoid phishing scams,” Twitter, 26 Feb 2010. [Online]. Available: https://blog.twitter.com/2010/avoid-phishing-scams. [Accessed 2/4/ 2015].

[24] “State twitter spam,” Twitter, 23 Mar 2010. [Online]. Available: https://blog.twitter.com/2010/state-twitter-spam. [Accessed 4/4/2015].

[25] “Fighting spam with botmaker,” Twitter, 20 Aug 2014. [Online]. Available: https://blog.twitter.com/2014/fighting-spam-with-botmaker. [Accessed 4/4/2015].

[26] M. McCord and M. Chuah, “Spam Detection on Twitter Using Traditional Classifiers,” Proceedings of the 8th international conference on Autonomic and Trusted Computing (ATC'11), Berlin, 2011.

[27] A. H. Wang, “Don’t follow me: Spam Detection in Twitter,” Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on, Athens, Greece, 2010.

[28] M. P. S. Michael Gilleland, “Levenshtein Distance, in Three Flavors,” university of pittburgh, [Online]. Available: http://people.cs.pitt.edu/~kirk/cs1501/Pruhs/Fall2006/Assignments/editdistance/Levenshtein%20Distance.htm. [Accessed 13/7/2015].

[29] R. J. R. Raj, S. Srinivasulu, and A. Ashutosh, “A Multi-classifier Framework for Detecting Spam and Fake Spam Messages in Twitter,” 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT), Apr. 2020.

[30] N. Imam and V. Vassilakis, “Detecting Spam Images with Embedded Arabic Text in Twitter,” 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), Sep. 2019.

[31] Rajasekaran Rajkumar and Jolly Masih, “Opinion Analysis on Twitter Data and Detecting Spam Tweets,” International Journal of Innovative Tech. and Exploring Eng., vol. 9, no. 2, pp. 711–714, Dec. 2019.

[32] O. Çıtlak, M. Dörterler, and İ. A. Doğru, “A survey on detecting spam accounts on Twitter network,” Social Network Analysis and Mining, vol. 9, no. 1, Jul. 2019.

[33] M. Mostafa, A. Abdelwahab, and H. M. Sayed, “Detecting spam campaign in twitter with semantic similarity,” Journal of Physics: Conference Series, vol. 1447, p. 012044, Jan. 2020.

[34] K Subba Reddy and E. Srinivasa Reddy, “Detecting Spam Messages in Twitter Data by Machine Learning Algorithms using Cross Validation,” International Journal of Innovative Tech. and Exploring Eng., vol. 8, no. 12, pp. 2941–2946, Oct. 2019.

[35] S. Linganur, “Detecting Spam in Twitter and Email using Machine Learning Approach,” International Journal for Research in Applied Science and Engineering Technology, vol. 7, no. 5, pp. 1652–1655, May 2019.

[36] N. Senthil Murugan and G. Usha Devi, “Detecting Streaming of Twitter Spam Using Hybrid Method,” Wireless Personal Communications, vol. 103, no. 2, pp. 1353–1374, Feb. 2018.

[37] Z. Alom, B. Carminati, and E. Ferrari, “Detecting Spam Accounts on Twitter,” 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Aug. 2018.

[38] X. Zhang, Z. Li, S. Zhu, and W. Liang, “Detecting Spam and Promoting Campaigns in Twitter,” ACM Transactions on the Web, vol. 10, no. 1, pp. 1–28, Feb. 2016.

[39] N. Eshraqi, M. Jalali, and M. H. Moattar, “Detecting spam tweets in Twitter using a data stream clustering algorithm,” 2015 International Congress on Technology, Communication and Knowledge (ICTCK), Nov. 2015.

[40] D. Sipahi, G. Dalkılıç, and M. H. Özcanhan, “Detecting spam through their sender policy framework records,” Security and Communication Networks, vol. 8, no. 18, pp. 3555–3563, May 2015.

Critical Success Factors on the Implementation of ERP Systems: Building a Theoretical Framework

Asimina Kouriati[1]
PhD Candidate, Department of Agricultural Economics
Aristotle University of Thessaloniki (AUTH)
Greece, Thessaloniki

Basil Manos[3]
Em. Professor, Department of Agricultural Economics
Aristotle University of Thessaloniki (AUTH)
Greece, Thessaloniki

Thomas Bournaris[2]
Assistant Professor, Department of Agricultural Economics
Aristotle University of Thessaloniki (AUTH)
Greece, Thessaloniki

Stefanos A. Nastis[4]
Associate Professor, Department of Agricultural Economics
Aristotle University of Thessaloniki (AUTH)
Greece, Thessaloniki

***Abstract*—Existing pressure for the confrontation of a radically changing external environment has led many companies to invest in various Information Systems, such as Enterprise Resource Planning (ERP), in order to optimize their production processes and strategies. Despite the fact that ERP system is an important strategic tool, many companies fail to take advantage of its benefits due to their default in many aspects of management and implementation. This study aims to investigate the critical success factors of enterprise resource planning system implementation and build a categorization framework so as to create a theoretical base that enhances any further research approaches in various sectors of the economy. Therefore, 37 ERP critical success factors were identified by using Content Analysis method and classified into relative categories to the ERP orientations of implementation and the ERP life cycle phases. Finally, these two types of categorization were merged in order to examine the critical success factors' behavior during the ERP implementation. This paper and the multilateral theoretical framework it creates, sets out how critical success factors must be taken into account by companies and marks a beginning point that promises a sequence of further research approaches in particular economic sectors or in a set of them. By fulfilling the purpose of this study, a significant contribution to computer science literature and especially to the ERP field is offered.**

***Keywords*—Enterprise resource planning (ERP); ERP implementation; critical success factors (CSFs); content analysis (CA); categorization; theoretical framework**

I. Introduction

Competitiveness and globalization have caused radical changes in the business environment [1] something that led many companies to rely on various Information Systems, such as Enterprise Resource Planning (ERP) [2, 3] in order to optimize their production processes and strategies. Enterprise resource planning (ERP) software is a solution that help various companies to integrate all the business functions [4] using shared information and a common database [5]. In particular, [6] argue that ERP uses a central database that collects and organizes data in real time in order to achieve this integration between business operating domains. Additionally, ERP system leads to the acquisition of many business advantages, such as operational efficiency and effectiveness, easy access to reliable information, maintenance of the competitive advantage, reduce the complexity of processes, profit increase and reduce cost [4, 5, 7-12]. In addition to its wide range of practical applications, ERP system is an equally active area of research interest [13]. Researches about ERP concern: the general frame of literature, the software and its optimization, the identification of factors that contribute to the adoption and the selection of those systems, the advantages and disadvantages of ERP implementation and finally, the efficiency estimation of ERP implementation [11, 13-19]. The identification of factors that are adjudged critical in ERP implementation success [9, 20 etc.] is also considered as an interesting research project.

Although many companies enjoy the benefits that ERP provides, other companies face various challenges regarding the implementation of these systems [4, 21] such as financial failures [8]. This happens because companies fail to take advantage of ERP benefits due to their default in some views of management and implementation [11]. The author in [22], in an annual report, states that 42% of the responders deem ERP implementation as successful. This is an important decline compared with the 82% positive responders of the previous year [23]. It is also mentioned in literature that ERP implementation failure starts from the wrong choice of ERP software, which is its initial state [17]. What is more, the unsuccessful implementation may be caused by organizational change, wrong organizational politics, inefficient project management, incorrect understanding of the system, incompatibility with business processes, poor management support and user education as well [13, 24]. These facts led many researchers to investigate factors that contribute to the successful implementation of ERP system in order to inform managers about the risk of adopting an ERP system and supply them with the necessary managing tools. Also, a corresponding paper that identifies the critical success factors in detail and quotes them as a background for further research approaches has not been done in the past.

The above reasons led this paper to investigate the critical success factors and classify them into categories relative to the ERP (orientations of) implementation and the ERP life cycle phases. CSFs' investigation and categorization, based on a literature survey, build a theoretical framework for further research approach in various economic fields which is the principal aim of this study. By fulfilling this aim, this paper offers a significant contribution to computer science literature and especially to the ERP field. At this point, it should be mentioned that this paper has the following structure: 1) a literature review of previous studies about the identification of critical success factors and their categorization; 2) the examination of the theory of content analysis; 3) the identification of critical success factors that was achieved by implementing the content analysis method; 4) the factors which were classified into two types of categorization and were merged with each other; 5) the conclusions that are drawn about the present study's content, 6) present study's practical implication and originality, 7) proposals for further research approaches and 8) research limitations.

II. Literature Review

Successful implementation of ERP systems often requires the identification of critical factors [7]. Apart from the CSFs' identification, researchers use also the categorization method because makes CSFs' searches easier by assigning concepts to the categories and defining the relationships between them [25].

A. Critical Success Factors of ERP Implementation

The elements that have a significant impact on the success of an ERP system are known as Critical Success Factors (CSFs) [26]. CSFs terminology appeared during 60s' to help many companies to achieve their goals and embrace competitiveness [5, 27, 28]. In an ERP environment, CSFs are defined as a sum of activities [29] that contribute to the creation of a mechanism for providing the information needed to the companies' managers [30].

B. Critical Success Factors' Identification

Many researches focused on the ERP systems' failure [4, 31] which can significantly be reduced by adopting a CSFs' strategic analysis [27, 32]. The identification and understanding of CSFs help a company to take effective measures in order to eliminate the reasons that affect the ERP systems' implementation in a negative way [33, 34]. The author in [7] claimed that the successful implementation of ERP systems often requires the identification of critical factors. This is proven by the fact that CSFs given in the literature are many. The current literature contains numerous researches about the critical success factors' identification [1, 5, 7, 35-37] which is claimed that it started from the moment that ERP system appeared in companies [20].

C. Critical Success Factors' Categorization

Apart from the CSFs' identification, researchers use also the categorization method - or else taxonomy into categories- in order to find out associated concepts for each CSF [25]. A taxonomy analysis matters because helps researchers to organize their knowledge issues surrounding the ERP implementation problems [25]. Categorization of critical factors also highlights the key features of ERP systems based on business process management principles [29] and is widely used in the literature [28, 32, 38, 39].

D. Identification and Categorization of Critical Success Factors: Previous Studies Relevant Content

Some researches with a relevant content is this of [40], who reviewed literature and identified twelve CSFs which were categorized them into strategic and tactical. The author in [41], identified ten factors creating, in this way, a base for a further research in Chinese companies and classified them into ERP implementation relative categories. The author in [29] identified twelve factors and provided a comprehensive taxonomy by dividing them into three parts corresponding to the ERP implementation's characteristics, such as setting-up, deployment and evaluation. Based on the relevant literature and an information system research model, critical factors that influence the ERP implementation success were identified in the research of [38]. These factors were classified also into four environmental categories in order to develop a theoretical framework which examines the relationship between CSFs and ERP implementation success in companies which belong to the Chinese economy [38]. The author in [39] examined 45 articles, identified 26 CSFs by using the content analysis method and divided them into strategic and tactical. After analyzing 95 studies by the methods of content and frequency analysis, [25] suggested 17 CSFs and, subsequently, classified them into ERP implementation categories.

The author in [32], after a literature survey, identified 33 critical success factors related to the implementation of ERP systems and, then, classified them from an operational and organizational point of view. The author in [5] identified nine critical success factors and studied their significance within Greek small and medium-sized enterprises. A literature review, in which 20 CSF were identified, was conducted by [28]. Then, these factors were grouped into organizational, project, users, expertise and software categories. This initial analysis was carried out in order to conduct a further investigation about ERP CSFs' importance in educational institutions [28]. The author in [37] identified 20 CSFs of ERP implementation in higher education by using the content analysis method of 38 studies in total. The author in [42] investigated the ERP critical success factors using the literature and investigated their relationship with the ERP success in the Indian automotive industry. Study of [43] examined the critical success factors on the implementation of ERP systems in some companies in the United Arab Emirates. A review of the relevant literature was held in order to identify some CSFs that have the power to exert a great impact on the success -or not- of the ERP implementation in general. Thereafter, the study conducted a further investigation into the effects of those CSFs on the implementation processes of ERP systems in a number of organizations within the United Arab Emirates.

As it can be perceivable, conducting only a literature review is not sufficient for the analysis of existing literature studied. Therefore, researchers impose further investigation by using specially designed questionnaires in order to specify the critical factors that strongly affect the successful ERP system

implementation in particular economic sectors or in a set of them [5, 9, 28, etc.]. As mentioned in the introductory section, this study restricts only on the first part aiming to create a multilateral theoretical framework that will be used by researchers for further investigation in various economic sectors. The ways in which this theoretical framework can be exploited are presented through the presentation of proposals for further studies at the end of this paper.

III. Content Analysis (CA)

Content Analysis (CA) is also widely used by researchers in the fields of critical success factors and ERP implementation in order to identify these factors by grouping them according to their common meaning and purpose [25, 37, 39, 44] and by accomplishing this way their general mapping [12]. This method, is the most common text analysis technique that constitutes a research tool for concept identification within a set of texts [45].

A. Content Analysis as a Research Methodology

Content analysis is a method that was initially developed for the human communication analysis in the field of social sciences but various empirical information systems' studies, adopted this method as part of research methodology [46]. This method is used to describe systematically, objectively and quantitatively a 'communication' material for the identification of the required characteristics of content. Also, qualitative research method, concerns the quantification of qualitative data and is widely used. In addition to that, it is applied to various media of "communication" through the study of documents and media (books, articles, journals, web-pages, letters and interviews) [46-48]. In this case, the media of "communication" are the scientific publications of the relevant literature [46]. CA method concerns the shrinkage of textual data by following a set of specific codes [49] or else it can be seen as a technique that 'squeezes' the words of a text into fewer categories based on certain coding rules. Coding is a part of the analysis that concerns the designation and categorization of phenomena and, during this procedure, data are separated into meaning parts, are carefully examined and compared to similarities and differences [39, 50]. These facts help researchers to study many data via a systematic methodology. CA is a method of summarizing any form by measuring its various aspects [51] and focuses on how often words or meanings appear in texts. This approach has been used occasionally to address a variety of issues and its basic idea is to obtain a list of concepts so as to measure the occurrence times of each concept in each text [39, 51].

B. Content Analysis Principles

There is a large number of choices that researchers have to make when implementing the content analysis method. These choices define the application of this research method and, also, influence the obtained results, the interpretation and the automation of the coding process [51]. The author in [51] proposed eight stages of choices that have to be made when a researcher wants to apply the content analysis method. These stages are presented below [39, 51]:

1) Selecting analysis level - What do the concepts mean: This stage refers to the selection of the research method. That is, if a single word or a set of words or phrases is searched, the research method is selected. This reason why this happens is that different results are going to be obtained when individual words are used during the research, in contrast to those of whole sentences. The use of individual words is helpful if the researcher wants to compare and contrast the results in specific terms. On the other hand, the use of whole phrases is helpful if the researcher wants to capture general concepts in a particular socio-linguistic community [51]. This paper uses different word combinations and whole phrases in order to find specific articles that are relevant to the critical factors of successful ERP implementation.

2) Select the number of concepts to be coded: This stage involves the choice of encoding a predefined set (number) of concepts or a more generalized and inductive coding approach [39, 51]. An interactive and inductive approach is selected to be used as part of this analysis because it allows the integration of the most recognized critical success factors of ERP implementation [39].

3) Choosing to code the frequency of a concept or its existence: This stage concerns the choice of measuring the relevant concepts or the situation that is implied around a phenomenon. The author in [51], more specifically, argued that the study of a whole phenomenon simplifies the process and eliminates the frequency correlations due to editorial selection. In the present study, it was initially chosen to emphasize on the frequency of concepts in order to construct an initial model of critical factors by taking into account the situation indicated around them according to their meaning [39].

4) The distinction of concepts – Level of generalization: A researcher must choose whether particular concepts should be coded exactly as recorded the same even they appear in various forms [51]. In other words, any word in this research that implies the same meaning is categorized with the same structure. As mentioned above, emphasis is initially laid on the frequency of the concepts. Only in this way will the first identification of the factors be achieved. In the second stage, a review of the factors meaning is done in order to redefine the factors and reduce them. These choices were made to facilitate the further research process.

5) Creating rules for coding texts: To ensure cohesion and coherence in coding, it is necessary to establish a set of "translation" rules that can be applied throughout the coding process [39]. In order to generalize the concepts of the texts systematically, the creation of a set of rules that translate general concepts into more general ones is required [51]. These rules can be adapted under the guidance of the literature as it can be noticed below:

- The articles involved in the content analysis highlight the critical success factors. Essentially, the collection of factors is limited only to those cited by the researchers as key and critical.

- Each one of the factors is recorded in a bibliographic program where their meaning is listed based on the selected research material of the analysis.
- Then, according to the factors nominal similarity, the reference frequency is measured.
- After that, the meaning of these factors is studied, in order to redefine their conceptual similarity, and the reference frequency is recounted.
- When the factors are fully identified, conceptually relevant categories and categories related to the life stages of the system will be created.
- Once the categories are completed, factors will be classified according to the categories established.

6) Management of ''useless'' information: This stage determines how the researcher should deal with the unnecessary information. The author in [51] was wondering whether "useless" information should be deleted, omitted or be used to re-evaluate and modify the coding system [51]. Deletion simplifies the process by minimizing the cost of editing and creates a simplified text which is manageable easily [51]. The author in [39] supported that projects, which refer to the investigation of the critical factors of successful ERP implementation, focus on the aggregation of all concepts regarding success factors. This is the reason why this research method includes the entire content of the literature articles. However, the merge of critical factors that have similar meaning is selected.

7) Text coding – What happens in the case of implied content: The author in [51], was wondering if the media of communication are coded as well as what concepts are present or/and what concepts are implied. Only if explicit concepts are used when coding, then texts can be compared in their writing style and a fully automated procedure can be followed [51]. On the contrary, the use of implicit concepts allows the researcher to compare texts from the point of view of social knowledge. This process is of vital significance to the meaning level where each term is coded [51]. In this case, the meaning of all factors is always examined in the light of the articles of analysis.

8) Analysis of results: The actual analysis stage is the examination of the frequency results [39]. In the present study, the researcher tries to seek a conclusion by examining the results and trends with regard to the critical success factors on ERP implementation. Using content analysis method, the critical factors of ERP successful implementation were extracted first, and, then, they were categorized. In this way, the main objective of this paper is covered by creating a theoretical framework that could form the basis for a further research approach in particular economic sectors or in a set of them.

IV. Critical Success Factors' Identification through Content Analysis (CA) Method Application

As it was mentioned earlier, content analysis method implementation concerns the quantification of qualitative data and is applied to various media of "communication" through the study of documents and media [46-48]. In this study, the media of "communication" are the scientific publications of the relevant literature [46].

A. Used Keawords in the Analysis

The selection of the literature material used came after searching in Google Scholar, Scopus and Research Gate and it was instrumental in shaping the main objective of this study. During this research, various keyword combinations were used in order to make the literature material solely relevant to the CSFs as defined in the first rule of analysis. The keywords that were used during this research are listed below (Table I).

Reviewing the literature [5, 24, 32, 37, 39, 52], the repeated use of some studies in similar research applications was observed [25, 29, 35,40]. This fact had been taken into account when the research material of the analysis was chosen because the literature itself indicates the trends in the identification of ERP critical success factors.

B. Distribution of the Analysis' Selected Research Material

The research material is considered important for the development of a factor list that influences the successful implementation [46] and belongs to the current literature as well as the older one, as in the case of other researchers [46, 37, 39]. Specifically, fifty research papers, fifteen conference papers and two dissertations were selected for the application of this analysis given that they met specific requirements, such as their relationship with the present research topic, the origin from reliable resources (scientific journals, conference proceedings, university and academic portals) and the year of publication (over 1999) [26, 39, 46] which is significant because the analysis refers to last two decades. The research material below is presented according to the way journals, conference proceedings and dissertations are distributed annually [27, 44] (Table II).

The annual distribution of journals, proceedings, and theses is presented in the above table in terms of sums. These sums are also converted to percentages per year (%) via the use of the rule of three. Twelve studies (17.91%) belong to 2018, eight studies (11.94%) to 2017, seven studies to 2014 (10.45%), five studies to 2015 (7.46%), four studies (6.97%) to 2003, 2005, 2010 and 2013, three studies to 2007 and 2011 (4.48%), two studies (2.99%) to 2006 and 2009 and one study (1.49%) to each of the remaining years. As it can be seen below, the material of 2018 outweighs the rest of the other years (Fig. 1).

C. Research Material Sectors

After the research on material collection, an initial review was carried out in order to examine the sectors in which critical factors of ERP successful implementation were investigated. The selected research material, which the corresponding research approach is applied to, focuses on various business sectors. These sectors were divided into those of the general business field, SMEs, industry, education, logistics, public, retail, processing of agricultural products, trade of consumer goods, and banking (Table III).

TABLE I. USED KEYWORDS

Id	Keywords	Id	Keywords
1	Critical Factors + Enterprise Resource Planning	**21**	ERP + Successful Implementation
2	Critical Factors + Enterprise Resource Planning + Implementation Success	**22**	Factors + ERP + Implementation Success
3	Critical Factors + Enterprise Resource Planning + Success	**23**	Key Factors + Enterprise Resource Planning + Implementation Success
4	Critical Factors + Enterprise Resource Planning + Successful Implementation	**24**	Key Factors + Enterprise Resource Planning + Success
5	Critical Factors + Enterprise Systems + Implementation Success	**25**	Key Factors + Enterprise Resource Planning + Successful Implementation
6	Critical Factors + Enterprise Systems + Success	**26**	Key Factors + Enterprise Systems + Implementation Success
7	Critical Factors + ERP + Successful Implementation	**27**	Key Factors + Enterprise Systems + Success
8	Critical Factors + Successful + Enterprise Resource Planning	**28**	Key Factors + ERP + Implementation
9	Critical Factors + Successful + Enterprise Resource Planning + Implementation	**29**	Key Factors + ERP + Implementation Success
10	Critical Success Factors + Enterprise Resource Planning + Implementation	**30**	Key Factors + Successful Implementation
11	Critical Success Factors + Enterprise Resource Planning	**31**	Key Success Factors + Enterprise Resource Planning
12	Critical Success Factors + Enterprise Systems + Implementation	**32**	Key Success Factors + Enterprise Resource Planning + Implementation
13	Critical Success Factors + ERP + Implementation	**33**	Key Success Factors + Enterprise Systems + Implementation
14	CSFs + Enterprise Resource Planning + Implementation	**34**	Key Success Factors + ERP
15	CSFs + Enterprise Systems + Implementation	**35**	Key Success Factors + ERP + Implementation
16	CSFs + ERP	**36**	Successful + Enterprise Resource Planning + Implementation
17	CSFs + ERP + Implementation	**37**	Successful + ERP
18	Enterprise Resource Planning + Implementation Success	**38**	Successful + ERP + Implementation
19	Enterprise Resource Planning + Successful Implementation	**39**	Enterprise Systems + Implementation Success
20	ERP + Implementation Success	**40**	Enterprise Systems + Successful Implementation

TABLE II. ANNUAL DISTRIBUTION OF THE ANALYSIS' SELECTED RESEARCH MATERIAL

Type of Study	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**
Journal Papers	0	0	1	0	2	1	3	2
Proceedings Papers	1	1	0	1	2	0	1	0
Theses	0	0	0	0	0	0	0	0
Sum	**1**	**1**	**1**	**1**	**4**	**1**	**4**	**2**
%	**1.5**	**1.5**	**1.5**	**1.5**	**6**	**1.5**	**6**	**3**
Type of Study	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
Journal Papers	2	0	2	4	3	0	4	4
Proceedings Papers	1	1	0	0	0	1	0	3
Theses	0	0	0	0	0	0	0	0
Sum	**3**	**1**	**2**	**4**	**3**	**1**	**4**	**7**
%	**4.5**	**1.5**	**3**	**4.5**	**5**	**1.5**	**6**	**10.5**
Type of Study	**2015**	**2016**	**2017**	**2018**	**2019**	**Sum**	**%**	
Journal Papers	4	0	7	10	1	50	**74,62**	
Proceedings Papers	1	1	0	1	0	15	**22,38**	
Theses	0	0	1	1	0	2	**2,98**	
Sum	**5**	**1**	**8**	**12**	**1**	**67**	**-**	
%	**7.5**	**1.5**	**12**	**18**	**1.5**	**-**	**100**	

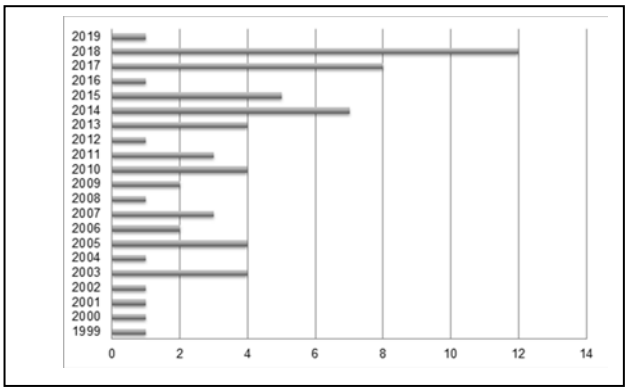


Fig. 1. Annual Distribution of the Selected Research Material.

TABLE III. ECONOMIC SECTORS IN WHICH THE SELECTED RESEARCH MATERIAL FOCUSES

Id	Economic Sector	Studies	Sum	(%)
1	*General Field of economic sectors*	[1, 25, 27, 29, 35, 38-41, 53-72]	**29**	**43.3**
2	*SMEs*	[5, 8-10, 32, 73-80]	**13**	**19.4**
3	*Industry*	[7, 20, 42, 52, 81-85]	**9**	**13.4**
4	*Education*	[24, 28, 36, 37, 86, 87]	**6**	**8.96**
5	*Logistics*	[88, 89]	**2**	**2.99**
6	*Public Sector*	[12, 90, 91]	**3**	**4.48**
7	*Processing of agricultural products*	[21, 34]	**2**	**2.99**
8	*Retail*	[92]	**1**	**1.49**
9	*Consumer goods*	[93]	**1**	**1.49**
10	*Banking*	[94]	**1**	**1.49**
Sum			**67**	**100**

Studies focused on the general investigation of CSFs in various companies – or without precise business activity - belong to the general field. In this case, there are 29 (43.3%) corresponding studies. In the case of small and medium-sized enterprises, there are 13 (19.4%) studies. These studies are equally relevant to the general field's studies. However, it is considered that there should be in a separate category because there are several research applications in the literature that emphasize the size of companies. The research applications of the industry sector are 9 (13.43%) and refer to the automotive, fire protection, manufacturing, electronics and construction companies. Education field studies, that are included in the analysis, are 6 (8.96%). Also, there are 3 studies (4.48%) that are specialized in the public sector. Two studies (2.99%) correspond to the field of processing agricultural products and logistics. Finally, one study (1%) corresponds in each field of the following sectors: retail, trade of consumer goods and banking. The deviations per economic sector are illustrated in the next chart (Fig. 2).

This chart shows that the general business field studies outperform the other sectors' studies. Assuming that the corresponding research applications in these sectors exist also in the actual range of the literature, it could be concluded that the majority of the whole literature surveys do not focus on a specific business activity, but they address the issue of ERP critical success factors in general.

D. Collection of Critical Success Factors

Moreover, a further review was carried out in order to collect the CSFs. Each of the factors and their meanings was recorded in a bibliographic program (Excel program) according to the selected research material. At first, 1069 factors were collected by measuring their appearance one to one. With regard to their nominal identity, critical success factors were decreased to 48. Regarding the rules of content analysis, factors' meaning was reviewed in order to redefine the conceptual similarity and in this way factors were reduced to 37. In the next table, the meaning of the identified critical success factors on ERP implementation is presented. The meanings resulted from reviewing the studies of the selected research material [5, 7, 20, 25, 28, 60 etc.] (Table IV).

In listing the meaning of critical success factors the characteristics that should be taken into account by entrepreneurs in order to successfully implement ERP systems are also essentially identified. The critical success factors' frequency order is presented in the following table in quantitative and percentage terms which are calculated via the rule of three (Table V).

Examining the frequency values, as they resulted from the content analysis's application (by measuring their appearance one to one), it could be considered that the reference level of the CSFs in the selected research material maybe reflects a real situation of the relevant literature in which emphasis is initially placed on top management (provision of the necessary resources for the operation of the system) and, subsequently, on various organizational and financial characteristics. Also, it could be concluded that the literature emphasis lies on the factors that have not been mentioned in detail in it except from some researchers who have considered them worthy of study, such as the number of implemented modules, company-wide support and commitment and knowledge management.

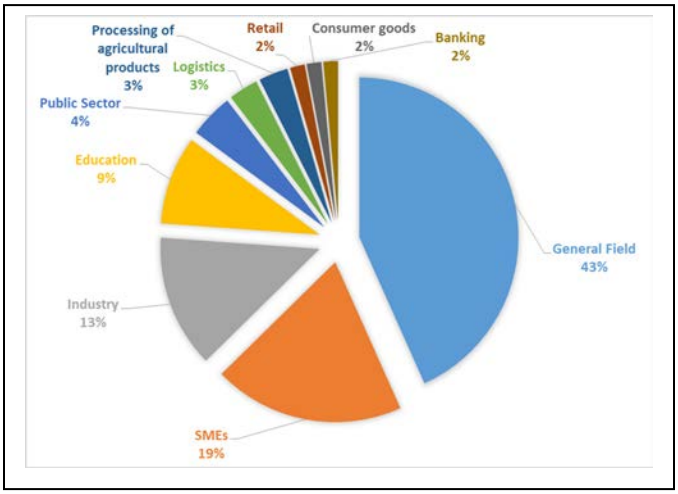


Fig. 2. Deviations' Structure Diagram of the Examined Economic Sectors.

TABLE IV. MEANING OF IDENTIFIED CRITICAL SUCCESS FACTORS

CSFs	**Meaning of CSFs**
Top management support and commitment	Level of commitment and support in obtaining ERP and providing valuable resources (system payments, selection and use of consultants, appointing executives to guide implementation).
Communication, collaboration and trust	Information exchange between the company's departments and partners. Common goals and trust relationships accomplish smooth functioning.
Composition of a capable and balanced project team	When the experience in ERP implementation field doesn't exist, top management comprise a group of people consisting of external consultants, departments' managers and project leaders in order to monitor the implementation. This team ensures the efficient implementation and its success, the goals' achievement and the staff training on ERP system implementation.
Project management	Application of knowledge and use of skills, tools, and techniques to project activities to meet the project's requirements.
Business plan, goals, scope, mission & vision	Provide a clear overview of the desired future situation.
Change Management	It ensures a smooth dealing with the changes which are resulting from BPR. It is supported by top management, partners and the project team. It depends on the desire and readiness of the company for possible changes.
Training	Knowledge delivery to project team members and users. Explanation of the system's logic and practical training. It aims at the proper understanding of the new business processes and the acquisition of information skills resulting in the development of various capabilities.
Business Process Re-engineering (BPR)	It starts by analyzing the existing work activities. New activities are recording by creating new business processes according to the system functions. It reduces the system modifications by matching the business processes with the software.
Service Quality	Users' support is obtained by external partners (vendors, consultants) and skilled business staff.
IT Infrastructure/Business and IT legacy systems	IT Infrastructure: Hardware and network existence before the insertion of a new system. Legacy systems: Used systems before ERP. A source of information about the ERP system's operational activities that helps to address any problems that may arise during the implementation.
Users and stakeholders' involvement	It leads to the system's adaptation in users' requirements, the optimum use and users' acceptance.
Software testing, customization and troubleshooting	Customization of the system is performed in order to maintain specific business processes that create a competitive advantage in a business. During customization, errors may occur that can be tackled only with the prosecution of rigorous testing which will contribute to the proper operation of the system.
Accuracy, Quality & Data Integrity	Accurate and reliable data that enters into the system either by legacy information systems or during business processes and produces correct information.
System Quality	It describes how reliable and functional a system is. It is measured by the way it is used, its functionality, flexibility and accuracy.
Organizational culture	The way the business has learned to understand, think and feel. This is communicated by how these features are shared and channeled among its members.
ERP package selection	An ERP system, by its very nature, imposes its own logic on the company's strategy, organization and mentality and it is imperative that the software package is carefully selected. It depends on business activity and its needs. It is carried out with the external consultants and vendors.
Presence of project champion & adequate role	A personality which understands the functions of the system, defines the objectives, assumes the burden of the project and ensures that its success depends on the involved users. It is often referred to as the project manager/director.
Implementation strategy & goals achievement timeframe	Decisions on how to implement an ERP software by using a specific methodology and timeframe. The selection and the implementation of an ERP system should be the result of a well-thought-out development and application of this ERP strategy.
Performance Monitoring, Evaluation & Feedback	The exchange of information between project team members and users. It concerns the measurement of performance with indicators of users' performance and satisfaction. It is a tactic of measuring progress and controlling efficiency in terms of use.
Use of consultants	Due to the complexity of the system, companies choose to collaborate with consultant companies in order for them to be helped with the functions of the system. The purpose of consultation is to select the right system, align the software with the business processes and configure it according to the company's desires. It plays an essential role in starting the project and is reduced in the last stages of the system's implementation.
Users' characteristics, skills and capabilities	Characteristics of users according to their behavior and, especially, their skills
Recognition of qualifications, reward and motivation	The ability of a project champion to maintain high morale on system users' by recognizing their achievements.

Minimum customization	It exists when the company adopts the procedures and functions that are integrated in the ERP software, instead of seeking to modify it precisely in the company's business processes. It is recommended to avoid the cost of potential modifications that are unavoidable if they are necessary for the system functionality. It is proposed that the customization level be set before applying the system.
Well defined Budget of Project	Creating a detailed and realistic budget before the system's insertion.
Post-implementation audit	It is carried out when the system reaches a growing stage in order to understand whether BPR has a real impact on the organization or not. It assesses if the goals are being met and provides lessons for the future.
Realistic expectations	Presentation of the real needs of a company by eliminating the unrealistic ones during the business plan preparation and the desired goals' setting.
Communication plan	A communication policy that channels information between top management, users and external partners.
Existence of empowered decision-makers	Empowering the team to make decisions on time and providing them with implements in order to bring the desired results during the system implementation.
ERP, business & business processes alignment	The adaptation of an ERP system with the existing business processes or new ones. Based on this adaptation, the company managers judg if the company's processes and culture undergo further changes and to what extent.
National culture	Bureaucracy, political and legal requirements determine how a business operates and therefore the operation of the ERP system.
Competitive & External Pressures	Pressure from the external competitive environment affects the company's internal environment. The pressure that the company receives from the external environment and, at the same time, the creation of innovative and competitive activities.
System support/Maintenance and further training	It is performed when the system is in the final stage of implementation. The system is supported by maintenance procedures during which continuous upgrades are made for optimal system operation and additional training that may be proved necessary.
ERP vendor selection	ERP suppliers, creators of system software. They handle the system's installation in the company, provide initial training and modify the system in collaboration with the consultants. The ERP vendors' selection is based on their history, support and capabilities.
Controlled ROI on ERP implementation	An indicator that evaluates the return of investment on ERP.
Knowledge Management	Acquiring the maximum amount of knowledge from external partners and staff in aiming at an automated system usage.
Company-Wide Support and Commitment	Once ERP enters a company, each user is required to adopt the use of the system by participating in the new responsibilities without invoking the previous working procedures.
Implemented modules	The functionality of the system is affected by the number of implemented modules. The company chooses either to use full range of ERP modules offered or to restrict itself to a specific number of modules according to specific needs and requirements.

TABLE V. FREQUENCY ORDER OF THE IDENTIFIED CSFS AFTER THE APPLICATION OF CONTENT ANALYSIS METHOD

Id	Factors	Frequency	(%)
1	Top management support and commitment	67	6.27
2	Communication, collaboration and trust	66	6.17
3	Composition of a capable and balanced project team	65	6.08
4	Project management	61	5.71
5	Business plan, goals, scope, mission and vision	59	5.52
6	Training	55	5.14
7	Change Management	54	5.05
8	Business Process Re-engineering (BPR)	52	4.86
9	Users and other stakeholders' involvement in evaluation, modification and implementation of the system	51	4.77
10	IT Infrastructure / Appropriate business and IT legacy systems	51	4.77
11	Service Quality	48	4.49
12	Software testing, customization and troubleshooting	42	3.93
13	System Quality	41	3.84
14	Accuracy, Quality & Data Integrity	37	3.46
15	Organizational culture	35	3.27
16	ERP package selection	33	3.09
17	Presence of project champion and adequate role	32	2.99

18	Implementation strategy and goals achievement timeframe	32	2.99
19	Performance Monitoring, Evaluation and Feedback	28	2.62
20	Use of consultants	25	2.34
21	Users' characteristics, skills and capabilities	16	1.5
22	Minimum customization	16	1.5
23	Recognition of qualifications, reward and motivation	15	1.4
24	Well Defined Project Budget	13	1.22
25	Post-implementation audit	11	1.03
26	Realistic expectations	10	0.94
27	Communication plan	8	0.75
28	ERP, business and business processes alignment	8	0.75
29	System support / Maintenance and further training	8	075
30	Existence of empowered decision-makers	7	0.65
31	National culture	6	0.56
32	Competitive and External Pressures	6	0.56
33	ERP vendor selection	4	0.37
34	Controlled ROI on ERP implementation (Return on Investment)	2	0.19
35	Knowledge Management	2	0.19
36	Company-Wide Support and Commitment	2	0.19
37	Implemented modules	1	0.09
Total		1069	100

Specifically, the factor of top management commitment and support has the highest number of frequency noting that the right leadership and management are important tools for the success of ERP implementation. The factor of communication, collaboration and trust has the second highest number of frequency which proves that the existence of the above characteristics between business, employees and external partners is crucial for successful ERP implementation. Project team, project management, business plan, change management and users' education factors follow in the order of frequency. These factors are sufficiently referred to as literature, which is something that highlights their importance to the successful implementation. Business Process Re-engineering to system quality factors follow high values in the order of frequency (Id: 41-52), which makes them particularly successful components of ERP implementation. Factors from data accuracy to recognition and reward (Id: 15-37) could very well be considered as the average values of the frequency order. This proves that the literature gives a relative importance to the existence of these factors without giving them the appropriate priority. Factors that range between the last numbers of frequency (Id: 1-13) were not widely reported in the selected material. However, these factors are considered worthy of study and investigation; this could also be a proposal for any further research approach. At this point, the conclusions after applying the content analysis are drawn. The identified critical success factors can be considered individually by the companies in order for the implementation of the system to be successful. Next step is to create categories and classify the identified factors in them in order to draw further conclusions about ERP success.

V. Critical Success Factors' Categorization

As it was mentioned in the section of literature review, critical success factors can be classified either into relative categories to the ERP system implementation (or else orientations), into relative categories to the ERP life cycle or in categories related to strategy and tactics [9, 25, 27, 34, 38-41, 56, 59, 61, 65, 81]. In this analysis, it was decided that critical factors will be classified according to the first two categories and in a combination of them. These types of categorization were chosen due to the fact that they are widely used in the literature. In addition to that, they were selected in order to strengthen the current theoretical framework with perspectives in which CSFs can be taken into account by companies collectively as dimensions of the ERP implementation and its life cycle. At the end of this analysis, the merge of these two categorization forms is held in order to provide further conclusions regarding the successful ERP implementation [53, 76, 95] and, also, to show another way in which CSFs can be taken into account by combining the implementation aspects with those of the system's life stages. Having all the above into consideration, these categories are related to the ERP orientations of implementation and the life stages of the system.

A. First Form of Categorization: ERP Orientations of Implementation

The first form of categorization concerns the classification of factors into organizational, project, human, technological/ERP and external partners (vendors-consultants) categories [9, 25, 28] (Table VI).

Once the categories were created, factors were classified according to the selected research material (mentioned in Table III) and the relative literature as indicated in the next table. Thirteen (13) organizational factors, six (6) project factors, five (5) human factors, ten (10) technological/ERP factors and three (3) external partners' related factors arose from the CSFs' classification (Table VII).

TABLE VI. FIRST TYPE OF FACTORS' CATEGORIZATION

Id	Categories	Conceptual Content of Categories
1	Organizational	It is related to the company's structure, general administration, processes, goals, culture, and business environment
2	Project	It is associated with a group of people who supervise the ERP project and implementation.
3	Human	It is related to users' relationships with ERP implementation. Skills and characteristics, participation and support in ERP implementation procedures.
4	Technological/ ERP factors	They are related to the system's functionality and the technological characteristics.
5	External partners	They highlight the relationship between company, ERP system and external partners.

TABLE VII. CLASSIFICATION OF FACTORS ACCORDING TO THEIR ORIENTATIONS IN ERP IMPLEMENTATION

Categories	Id	Critical Success Factors
Organizational factors	1	Business Process Re-engineering (BPR)
	2	Well Defined Budget of Project
	3	Business plan, goals, scope, mission and vision
	4	Change Management
	5	Communication, collaboration and trust
	6	Communication plan
	7	Competitive and External Pressures
	8	Knowledge Management
	9	National culture
	10	Organizational culture
	11	Controlled ROI on ERP implementation (Return on Investment)
	12	Realistic expectations
	13	Implementation strategy and goals achievement timeframe
Project factors	1	Existence of empowered decision-makers
	2	Performance, Monitoring, Evaluation and Feedback
	3	Presence of project champion and adequate role
	4	Project management
	5	Composition of a capable and balanced project team
	6	Recognition of qualifications, reward and motivation
Human factors	1	Top management support and commitment
	2	Company-Wide Support and Commitment
	3	Training
	4	Users and other stakeholders' involvement in evaluation, modification and implementation of the system
	5	Users' characteristics, skills and capabilities
Technological/ ERP factors	1	Implemented modules
	2	IT Infrastructure / Business and IT legacy systems
	3	Software testing, customization and troubleshooting
	4	System Quality
	5	Accuracy, Quality & Data Integrity
	6	ERP package selection
	7	Minimum customization
	8	ERP, business and business processes alignment
	9	Post-implementation audit
	10	System support / Maintenance and further training
Extertnal Partners' factors	1	Service Quality
	2	Use of consultants
	3	ERP vendor selection

The differences between the above categories are shown in percentage form below (Fig. 3). It can be noticed that organizational (35%) and technological/ERP related factors (27%) outweigh the rest of the other categories. This fact points out that the selected research material, or possibly most of the literature, gives more weight to these dimensions of the implementation. In particular, it is concluded that the successful implementation of ERP systems in various sectors of the economy depends mainly on the components of business management and technological aspects. Under no circumstances do the above comments negate the importance of the other categories - dimensions in the successful implementation of the system.

B. Second Form of Categorization: ERP System's Life Stages

The second form of categorization involves the classification of factors in the life stages of ERP system [81, 61, 59]. This type of categorization concerns the critical success factors that belong to the pre-implementation phase, implementation phase and the post-implementation phase (Table VIII).

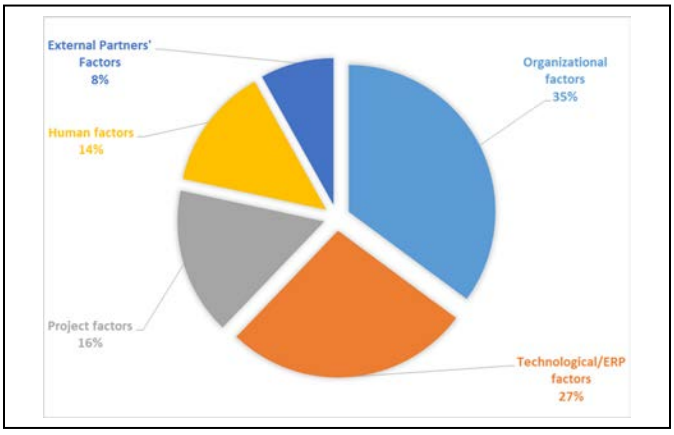


Fig. 3. Deviations' Structure of the First Type of Categorization.

TABLE VIII. SECOND TYPE OF FACTORS' CATEGORIZATION

Id	Categories	Conceptual Content of ERP Stages
1	Pre-implementation phase	Factors related to a company's preparation procedures for acquiring an ERP system. These include the system's investment examination, the software package and external partners' selection and the strategic planning.
2	Implementation phase	Factors related to project-related activities, users' organization, software testing, configuration, conversion, stabilization and finally the ERP implementation.
3	Post-implementation phase	Factors related to activities that take place as long as the use of the system is in the hands of mature users and last as long as the processes of upgrading, maintenance and additional training. At the end of these processes, the system is replaced with a new one.

The use of this type of categorization was made because the ERP success requires the identification and the management of the critical elements in each phase of system life cycle implementation [79]. Once the categories were created, factors were classified, as indicated below according to the selected research material and the relative literature (Table IX).

TABLE IX. FACTORS' CLASSIFICATION ACCORDING TO ERP SYSTEM'S LIFE PHASES

Categories	Id	Critical Success Factors
Pre-implementation phase factors	1	Top management support and commitment
	2	Communication, collaboration and trust
	3	Composition of a capable and balanced project team
	4	Business plan, goals, scope, mission and vision
	5	Change Management
	6	Business Process Re-engineering (BPR)
	7	IT Infrastructure / Business and IT legacy systems
	8	ERP vendor selection
	9	Use of consultant
	10	Implemented modules
	11	ERP package selection
	12	Minimum customization
	13	Implementation strategy and goals achievement timeframe
	14	Well Defined Budget of Project
	15	Controlled ROI on ERP implementation (Return on Investment)
	16	Realistic expectations
	17	Knowledge Management
	18	National culture
	19	Competitive and External Pressures
Implementation phase factors	1	Project management
	2	Training
	3	Communication plan
	4	Users and other stakeholders' involvement in evaluation, modification and implementation of the system
	5	Service Quality
	6	System Quality
	7	Software testing, customization and troubleshooting
	8	Accuracy, Quality & Data Integrity
	9	Organizational culture
	10	Presence of project champion and adequate role
	11	Recognition of qualifications, reward and motivation
	12	Users' characteristics, skills and capabilities
	13	Company-Wide Support and Commitment
	14	ERP, business and business processes alignment
	15	Existence of empowered decision-makers
	16	Performance Monitoring, Evaluation and Feedback
Post-implementation phase factors	1	Post-implementation audit
	2	System support / Maintenance and further training

After the CSFs' classification into the above-mentioned categories, nineteen (19) ERP pre-implementation, sixteen (16) ERP implementation and two (2) post-implementation ERP phase factors arise. The differences between the categories are shown in percentage form (Fig. 4) and, as can be seen, ERP pre- implementation and implementation phase factors occupy the largest percentage.

These results point out that the selected research material pays close attention to the first two phases of the system's life. The literature states that in the event that the last phase of system implementation is examined, further factors and parameters should be taken into account, such as the frequency of upgrades, financial support and the provision of specialized assistance by external partners [96, 97].

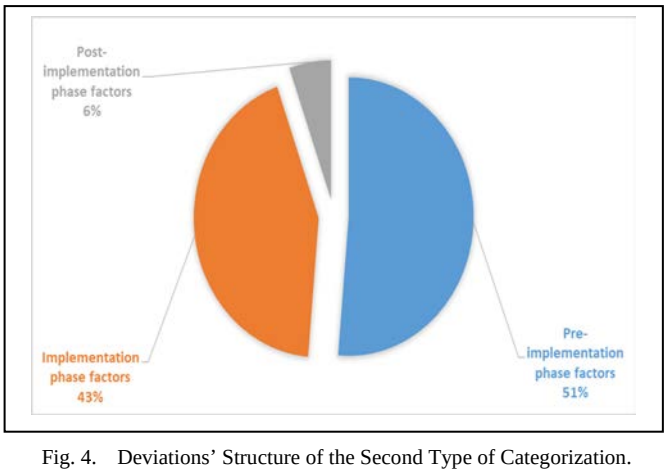


Fig. 4. Deviations' Structure of the Second Type of Categorization.

C. Merging of the Two Categorization Forms

Subsequently, the above types of categorization were merged in order to draw further conclusions about the critical factors, their categorization and the ERP system success [53, 76, 95] (Table X).

The application of the merger shows that eleven organizational, one project, one human, four ERP and two external partners' factors are included in ERP pre-implementation phase. Additionally, two organizational, five project, four human, one external partners' and four ERP system factors are taken into account as implementation phase factors also. Finally, two factors relating to the ERP system are entirely post-implementation phase factors. The above results are presented in percentage, cumulative and diagrammatic form below (Table XI, Fig. 5 to 7).

TABLE X. MERGED FACTORS' CLASSIFICATION

Category	***Pre-implementation phase factors***	***Implementation phase factors***	***Post-implementation phase factors***
Organizational factors	Communication, collaboration and trust	Communication plan	
	Business plan, goals, scope, mission and vision		
	Change Management		
	Business Process Re-engineering (BPR)		
	Implementation strategy and goals achievement timeframe		
	Well Defined Budget of Project	Organizational culture	
	Controlled ROI on ERP implementation (Return on Investment)		
	Realistic expectations		
	Knowledge Management		
	National culture		
	Competitive and External Pressures		
Category	**Pre-implementation phase factors**	**Implementation phase factors**	**Post-implementation phase factors**
Project factors	Composition of a capable and balanced project team	Project management	
		Presence of project champion and adequate role	
		Recognition of qualifications, reward and motivation	
		Existence of empowered decision-makers	
		Performance, Monitoring, Evaluation and Feedback	
Category	**Pre-implementation phase factors**	**Implementation phase factors**	**Post-implementation phase factors**
Human factors	Top management support and commitment	Training	
		Users and other stakeholders' involvement in evaluation, modification and implementation of the system	
		Users' characteristics, skills and capabilities	
		Company-Wide Support and Commitment	
Category	**Pre-implementation phase factors**	**Implementation phase factors**	**Post-implementation phase factors**
Technological/ERP factors	IT Infrastructure / Business and IT legacy systems	System Quality	Post-implementation audit
	Implemented modules	Software testing, customization and troubleshooting	
	ERP package selection	Accuracy, Quality and Data Integrity	System support / Maintenance and further training
	Minimum customization	ERP, business and business processes alignment	
Category	**Pre-implementation phase factors**	**Implementation phase factors**	**Post-implementation phase factors**
Extertnal Partners' factors	ERP vendor selection	Service Quality	
	Use of consultants		

TABLE XI. MERGED FACTORS' CLASSIFICATION IN QUANTITATIVE AND PERCENTAGE FORM

Categ.	**Pre-implementation phase factors**
Organizational factors	11
Project factors	1
Human factors	1
Technological/ERP factors	4
External partners' factors	2
Total	19

Fig. 5. Pre-Implementation Phase.

Categ.	**Implementation phase factors**
Organizational factors	2
Project factors	5
Human factors	4
Technological/ ERP factors	4
External partners' factors	1
Total	16

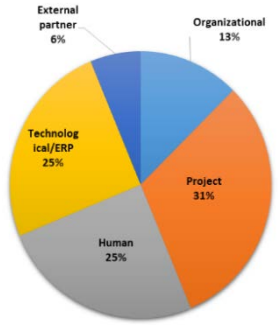


Fig. 6. Implementation Phase.

Categ.	**Post-implementation phase factors**
Organizational factors	0
Project factors	0
Human factors	0
Technological/ ERP factors	2
External partners' factors	0
Total	2

Fig. 7. Post-Implementation Phase

Some of the factors, which were categorized into dimensions of the ERP system implementation, have a prominent place both in the pre-implementation and implementation stage while in the last stage the external partners' critical factors gain a dominant role. In particular, the majority of organizational factors seem to play a dominant role in the pre-implementation phase which makes the organizational aspects and further parameters of them necessary elements for the acquisition of the system. Some of these factors, such as communication plan and organizational culture, are included in the second phase emphasizing the importance of their existence during the implementation of the system. On the contrary, the majority of project factors play a greater role in the implementation phase of the system. One of them -the composition of a capable and balanced project team- appears in the pre-implementation phase, noting that its existence prior to system acquisition is essential for the future implementation success. Human factors appear mainly in the implementation phase of the system. Thus, this is something that indicates that the "human" element is particularly presented in activities related to the system implementation flow and the users' administration.

Some of the technological/ERP factors are identified in the pre-implementation phase, which indicates that important decisions must be made by the outset regarding the pre-existing situation. These factors are also identified in the

implementation phase. Therefore, they define guidelines for the system's configuration and quality. It was also concluded that only ERP system factors play a vital role in the post-implementation phase. This is explained by the fact that specific procedures are carried out as soon as the system used matures in the users' hands until it is replaced with a new one. These procedures concern the software maintenance, upgrades, additional user training and implementation audit in order to be made perfectly clear whether the company's goals were accomplished or not. Finally, factors that are relevant to ERP vendors and consultants are superior to the pre-implementation phase of the system. This is justified by the fact that in the first stage of ERP implementation, processes that take place in it prepare the company for the acquisition of an ERP system. One of these processes is the selection of external partners, whose quality is evident in the implementation stage, and is an important aspect of the system implementation success.

VI. Conclusion

This study aimed to investigate the critical success factors of enterprise resource planning system implementation and build a categorization framework in order to create a theoretical basis that will enhance any further research approaches in various sectors of the economy. By fulfilling the purpose of this study, a significant contribution to computer science literature and especially to the ERP field is offered. In order for the study's legitimate objective to be achieved, the critical factors of successful ERP implementation were initially determined by reviewing the relevant literature through the use of content analysis method. Then, the identified CSFs were classified into relative categories to the orientations of ERP implementation and ERP system's life cycle phases. Finally, these two types of categorization were merged in order to examine the critical success factors' behavior during the ERP implementation so as for useful conclusions to be drawn.

Content Analysis was chosen because various empirical information systems' studies adopted this method as a part of a research methodology. This method was applied according to the literature's standards in relevant scientific studies which had been found by searching in specific databases and by using keywords about ERP implementation and critical success factors. The studies that were selected to take place in this investigation include the research material that meets the present study's topic requirements. Most of this research material helped in shaping the objective goal of this paper and was considered important for the development of a factors' list that influences the successful ERP implementation in various economic sectors. These sectors were divided into many fields which led to the conclusion that corresponding research applications in these sectors may occur also in the actual range of the literature. CA method was applied by reporting the frequency of CSFs' concept in the selected research material content, something that helped thirty-seven factors to be identified. The conclusions drawn with regard to the frequency results from the application of the content analysis were multiple and presented in detail in the corresponding section above. Specifically, it was generally concluded that the CSFs' reference to the selected research material maybe reflects a real situation of the relevant literature in which emphasis is mainly placed on top management, subsequently on various organizational and financial characteristics highlighting the importance of them in the successful implementation and, finally, on factors that have not been mentioned in detail in the literature, except from some researchers who are considered them worthy of study. The meaning of these factors were also recorded and presented emphasizing, in this way, the success characteristics that should be taken into consideration by entrepreneurs during the ERP system implementation.

After the factors' identification, categories, relative to ERP implementation and its life phases, were created and merged with each other in order to strengthen the theoretical framework towards a more integrated approach for decision making policy regarding the critical factors which influence the ERP implementation success. The first type of categorization referred to the creation of organizational, project, human, technological/ERP and external partners (vendors-consultants) categories and the second one the creation of pre-implementation, implementation and the post-implementation phase categories. The categorization framework was set up by the use of the relevant literature which also suggested the classification of thirty-seven critical factors into the above-mentioned categories. The analysis results showed that the organizational and technological/ERP factors outweigh the rest of the other categories. This proves that the selected research material, or possibly most of the relevant literature, gives more weight to these dimensions of the implementation. In particular, it was concluded that the successful implementation of ERP systems in various sectors of the economy depends mainly on the components of business management and technological aspects, which is something that do not negate the importance of the other dimensions (project, human, external partners) in the successful implementation of the system. It was also pointed out that the selected research material pays close attention to the first two phases of the system's life. Therefore, it was suggested that in the event that the last phase of system implementation is examined, further factors and parameters have to be defined and taken into account directly by the enterprises (frequency of upgrades, financial support and specialized assistance by external partners). Finally, through the application of the merger and the discussion of its results, it is evident in which aspects of ERP implementation emphasis is placed on per system's life cycle phase.

VII. Practical Implication and Originality

By fulfilling the purpose of this study, a significant contribution to computer science literature and especially to the ERP field is offered. Firstly, it is worth mentioning that the discussion of the frequency and categorization results probably helps in understanding the literature's prevailing trends about ERP implementation and CSFs. The discussion on the results that occurred from the merging of the two categorization forms, could very well be referred to as the means with which the aspects that characterize and influence the implementation of ERP system can be taken into account by various companies during the three life stages of the system. Last but not least, it must be pointed out that the

creation of a theoretical framework, which involves various (four in total) aspects of the treatment of key topics during the ERP implementation, is innovative and constitutes a promising tool that -as it was mentioned earlier in detail- can be used by many researchers as a basis for many types of researches. This study is purely theoretical but, as an available scientific article, may provide background knowledge for conducting a series of further research approaches in specific economic sectors or in a set of them. Present study's authors promise to explicit this knowledge in order to strengthen the computer science literature with new findings.

VIII. Proposals for Further Research Approaches

The identification of critical factors through the content analysis application created a theoretical framework which allows critical factors to be considered by companies individually as features of the implementation and the system. According to this perspective, the identification of these factors' importance and influence in the degree of the ERP implementation success is suggested. The importance's ranking of critical factors, specifically, might be a proposal for companies, indicating the order in which the elements of each factor should be taken into account during the ERP implementation. However, further investigation should be carried out in order to include the degree of successful implementation and to study its relationship with critical factors. This type of identification will enable further deepening of the existing literature and disseminate new findings about the ERP system and its implementation in specific economic sectors or in a set of them, such as the above-mentioned general business field. The categorization analysis application respectively led to the creation of three more theoretical frameworks (four in total) that allow critical success factors to be considered by companies not only collectively as dimensions of the ERP implementation and its life phases, but also in combination. According to these perspectives, the identification of these approaches' relationships with the degree of the ERP implementation success is suggested. Through this examination, useful findings can be drawn if the implementation success is shaken positively, negatively or not at all, in the case that critical factors are taken collectively into account by companies as aspects of implementation, life cycle and their possible combination. More specifically, carrying out a corresponding analysis, many researchers will be greatly helped to be aware of the area in which the ERP system implementation may present problems and make various proposals for avoiding the undesirable result.

IX. Research Limitations

One of the most significant limitations of this research was the selection of a literature part to be used for the analysis purposes, given that it was not possible to use the whole literature's studies. Analyzing the relevant literature, many studies conduct a literature review and impose a further investigation by using specially designed questionnaires in order to specify the critical factors that strongly affect the successful ERP system implementation. As mentioned in the introductory section, this study restricts only on the first part, which is something that can be taken into account as a limitation. The existence of this limitation is explained by the fact that this paper aims to create a multilateral theoretical framework that will be used by researchers directly in the part of the further investigation establishment and will further contribute to the ERP literature.

References

[1] E.J. Umble, R.R. Haft and M.M. Umble, "Enterprise resource planning: Implementation procedures and critical success factors". European Journal of Operational Research, Vol. 146, pp. 241–257, 2003.

[2] I.C. Ehie and M. Madsen, "Identifying critical issues in enterprise resource planning (ERP) implementation". Computers in Industry Vol. 56, pp. 545–557, 2005.

[3] Madapusi and D. Ortiz, "The Influence of Technical Competence Factors in ERP System Implementations". Journal of Applied Business and Economics, Vol. 16, pp. 27-39, 2014.

[4] A. Baykasoglu and I. Gölcük, "Comprehensive fuzzy FMEA model: a case study of ERP implementation risks". Oper Res Int J,Vol. 20, pp. 795–826, 2020.

[5] P. Chatzoglou, D. Chatzoudes, L. Fragidis and S. Symeonidis, "Critical success factors for ERP implementation in SMEs". Federated Conference on Computer Science and Information Systems, FedCSIS, pp. 1243–1252, 2016.

[6] F. Jalil, A. Zaouia and R. El, "The Impact of the Implementation of the ERP on End-User Satisfaction Case of Moroccan Companies". International Journal of Advanced Computer Science and Applications, Vol. 7, 2016.

[7] H.M. Beheshti, B.K. Blaylock, D.A. Henderson and J.G. Lollar, "Selection and critical success factors in successful ERP implementation". Competitiveness Review, Vol. 24, pp. 357-375, 2014.

[8] E. Abu-Shanab, R. Abu-Shehab, M. Khairallah, "Critical success factors for ERP implementation: The case of Jordan". The International Arab Journal of E-Technology, Vol. 4, pp. 1-7, 2015.

[9] N.A.A. Ahmed and M. Sarim, "Critical Success Factors Plays a Vital Role in ERP Implementation in Developing Countries: An Exploratory Study in Pakistan". International Journal of Advanced Computer Science and Applications, Vol. 8, pp. 21-29, 2017.

[10] S. Chaveesuk and S. Hongsuan, "A Structural Equation Model of ERP Implementation Success in Thailand". Review of Integrative Business and Economics Research, Vol. 6, pp. 194-204, 2017.

[11] M.I. Mahraz, L. Benabbou and A. Berrado, "Implementation and Management of ERP Systems: A Literature Review". International Conference on Industrial Engineering and Operations Management Bandung, Indonesia, IEOM Society International, pp.1684-1694, 2018.

[12] S. Santos, C. Santana and J. Elihimas, "Critical Success Factors For ERP Implementation In Sector Public: An Analysis Based on Literature and a Real Case". Twenty-Sixth European Conference on Information Systems (ECIS) proceedings, Portsmouth, UK, pp.01-15, 2018.

[13] S.E. Hamdi and A. Abouabdella, "Literature review of implementation of an enterprise resources planning: Dimensional approach". 4th International Conference on Logistics Operations Management (GOL), Le Havre, France, IEEE, pp. 01-07, 2018.

[14] S. Alkatheri, and S. Almandeel, S. "An Exploration of Critical Success Factors for Enterprise Resource Planning System Implementation". The International Journal of Humanities and Social Studies, Vol. 7, pp. 10-18, 2019.

[15] M.S. Shakkah, K. Alaqeel, A. Alfageeh and R. Budiarto, "An Investigation Study on Optimizing Enterprise Resource Planning (ERP) Implementation in Emerging Public University: Al Baha University Case Study". International Journal of Electrical and Computer Engineering (IJECE), Vol. 6, pp. 1920-1928, 2016.

[16] N. Karia and M. Soliman, "Factors affecting enterprise resource planning (ERP) systems adoption among higher education institutions in Egypt". International Journal of Advanced and Applied Sciences, Vol. 4, pp. 144-151, 2017.

[17] M. Haddara, "ERP systems selection in multinational enterprises: a practical guide". International Journal of Information Systems and Project Management, Vol. 6, pp. 43-57, 2018.

[18] Z. Yajun, L. Yujie and M. Skibniewski, "Enterprise Resource Planning Systems for Project Based Firms: Benefits, Costs & Implementation Challenges". Journal for the Advancement of Performance Information and Value (PBSRG), Vol. 4, pp. 85-96, 2012.

[19] A. Wibowo and M.W. Sari, "Measuring Enterprise Resource Planning (ERP) Systems Effectiveness in Indonesia". TELKOMNIKA (Telecommunication Computing Electronics and Control), Vol. 16, pp. 343-351, 2018.

[20] E. Reitsma and P. Hilletofth, "Critical success factors for ERP system implementation: a user perspective". European Business Review, Vol. 30, pp. 285-310, 2018.

[21] S. Duangekanong, "Factors Influencing the Success of an ERP System: A Study in the Context of an Agricultural Enterprise in Thailand". Silpakorn U Science & Tech J, Vol. 8, pp. 18-45, 2014.

[22] Panorama Consulting Solutions, "2018 ERP report a Panorama Consulting Solutions research report". Panorama Consulting Solutions, 2018.https://cdn2.hubspot.net/hubfs/2184246/2018%20ERP%20Report.pdf. Assesed 10/6/2018.

[23] Panorama Consulting Solutions. "2017 ERP report a Panorama Consulting Solutions research report". Panorama Consulting Solutions, 2017. content/uploads/2017/07/2017-ERP-Report.pdf. Assesed 10/6/2018.

[24] A. Fadelelmoula, "The effects of the critical success factors for ERP implementation on the comprehensive achievement of the crucial roles of information systems in the higher education sector". Interdisciplinary Journal of Information, Knowledge, and Management, Vol. 13, pp. 21-44, 2018.

[25] S. Dezdar and A. Sulaiman, "Successful enterprise resource planning implementation: taxonomy of critical factors". Industrial Management and Data Systems, Vol.109 No.8, pp.1037-1052, 2009.

[26] A. Tarhini, H. Ammar, T. Tarhini and R. Masa'deh, "Analysis of the Critical Success Factors for Enterprise Resource Planning Implementation from Stakeholders' Perspective: A Systematic Review". International Business Research, Vol. 8, pp. 25-40, 2016.

[27] J. Ram and D. Corkindale, "How "critical" are the critical success factors (CSFs)? Examining the role of CSFs for ERP". Business Process Management Journal, Vol. 20, pp. 151-174, 2014.

[28] F.C.F. Leandro, M.P. Méxas and G.M. Drumond, "Identifying critical success factors for the implementation of enterprise resource planning systems in public educational institutions". Brazilian Journal of Operations and Production Management, Vol. 14, pp. 529-541, 2017.

[29] M. Al-Mashari, A. Al-Mudimigh and M. Zairi, "Enterprise resource planning: a taxonomy of critical factors". European Journal of Operational Research, Vol. 146, pp. 352-364, 2003.

[30] S.Y. Huang, A. Chiu, P.C. Chao and A. Arniati, "Critical Success Factors in Implementing Enterprise Resource Planning Systems for Sustainable Corporations". Sustainability, Vol. 11, pp. 1-53, 2019.

[31] E. Hustad and D. Olsen, "ERP Implementation in an SME: A Failure Case", in: J. Devos, H. van Landeghem and D. Deschoolmeester (ed) Information Systems for Small and Medium-sized Enterprises. Progress in IS. Berlin, Heidelberg: Springer, 2014.

[32] M.M. Ahmad, and R.P. Cuenca, "Critical success factors for ERP implementation in SMEs". Robotics and Computer-Integrated Manufacturing, Vol.29, pp. 104-111, 2013.

[33] E.W.T. Ngai, C.C.H. Law and F.K.T Wat, "Examining the critical success factors in the adoption of enterprise resource planning". Computers in Industry, Vol. 59, pp. 548-564, 2008.

[34] R. Farrokhian, F. Soleimani, Y. Gholipour-Kanani and S. Ziabari, "A Structural Equation Model for Identifying Critical Success Factors of Implementing ERP in Iranian, Kalleh Food Product Company". In the 2014 International Conference on Industrial Engineering and Operations Management Bali, Indonesia, 2014.

[35] F.F.H. Nah, J.L.S Lau and J. Kuang, "Critical factors for successful implementation of enterprise systems". Business Process Management Journal, Vol. 7, pp. 285-296, 2001.

[36] F.F.H. Nah and S. Delgado, "Critical success factors for enterprise resource planning implementation and upgrade". Journal of Computer Information Systems 47:99-113, 2006.

[37] R. Thompson, O. Olugbara and A. Singh, "Deriving critical success factors for implementation of enterprise resource planning systems in higher education institution". The African Journal of Information Systems, Vol. 10, pp. 21-47, 2018.

[38] Z. Zhang, M.K.O Lee, P. Huang, L. Zhang and X. Huang, "A framework of ERP systems implementation success in China: An empirical study". International Journal of Production Economics, Vol. 98, pp. 56–80, 2005.

[39] S. Finney and M. Corbett M, "ERP implementation: a compilation and analysis of critical success factors". Business Process Management Journal, Vol. 13, pp. 329-347, 2007.

[40] P. Holland and B. Light. "A Critical Success Factors Model For Enterprise Resource Planning Implementation|. Seventh European Conference on Information Systems (ECIS), Copenhagen, IEEE Software, pp.30-36, 1999.

[41] L. Zhang, M.K.O. Lee, Z. Zhang and P. Probir Banerjee, "Critical Success Factors of Enterprise Resource Planning Systems Implementation Success in China". 36th Hawaii International Conference on System Sciences (HICSS), Big Island, HI, USA, IEEE, pp. 01-10, 2002.

[42] S. Lakshmanan, S. Edmund and D. Kinslin, "An Empirical Analysis on Critical Success Factors for Enterprise Resource Planning (ERP) Implementation in Automobile Auxiliary Industries:. International Journal of Engineering & Technology, Vol. 7, pp. 447-452, 2018.

[43] M.M. Nkasu, "Investigation of the Effects of Critical Success Factors on Enterprise Resource Planning (ERP) Systems Implementation in the United Arab Emirates", in S. Satapathy, V. Bhateja, J. Mohanty and S. Udgata (ed) Smart Intelligent Computing and Applications. Smart Innovation, Systems and Technologies, vol 159. Signapore: Springer, 2020.

[44] N. Ahmad, A. Haleem and A.A. Syed, "Compilation of Critical Success Factors in Implementation of Enterprise Systems: A Study on Indian Organizations". Global Journal of Flexible Systems Management, Vol. 13, pp. 217-232, 2012.

[45] D. Silverman, "Doing Qualitative Research: A Practical Handbook", Thousand Oaks, CA: Sage, 2000.

[46] M. Nasir and S. Sahibuddin, "Critical success factors for software projects: A comparative study". Scientific Research and Essays, Vol. 6, pp. 2174-2186, 2011.

[47] H.F. Hsieh and S.E. Shannon, "Three Approaches to Qualitative Content Analysis". Qualitative Health Research, Vol. 15, pp. 1277-1288, 2005.

[48] E.R. Babbie, The Practice of Social Research, 12th ed., Belmont: Wadsworth, Cengage Learning, 2010.

[49] A.B. Marvasti, "Qualitative content analysis: A novice's perspective". Forum Qualitative Sozialforschung, 2019.

[50] Strauss and J.M. Corbin, "Basics of qualitative research: Grounded theory procedures and techniques". Sage Publications, Inc, 1990.

[51] K. Carley, "Coding Choices for Textual Analysis: A Comparison of Content Analysis and Map Analysis". Sociological Methodology, Vol. 23, pp. 75-126, 1993.

[52] M. Hassan, M.A. Jabar, F. Sidi,Y.Y. Jusoh and S. Abdullah, "Critical Success Factors and their Influence in ERP implementation success of Organizational Performance". Acta Informatica Malaysia (AIM), Vol. 2, pp. 12-16, 2018.

[53] J. Esteves and J. Pastor, "Towards the unification of critical success factors for ERP implementations". 10th Annual BIT Conference, Manchester, UK, 2000.

[54] B. Wong and D. Tein, "Critical Success Factors for ERP Projects". Australian Institute of Project Management Conference Program, AIPM, Sydney, pp. 01 –08, 2003.

[55] A. Shatat, "Critical Success Factors in Enterprise Resource Planning (ERP) System Implementation: An Exploratory Study in Oman". The Electronic Journal of Information Systems Evaluation, Vol. 18, pp. 36-45, 2015.

[56] A.U. Shedu and T. Masunda, "The review of critical success factors of enterprise resource planning system implementation". Discovery Publications, Vol. 54, pp. 484-495, 2018.

[57] B. Gartner and C. Duller, "Critical Factors of ERP System Implementation – Quantitative Empirical Results from Austria". International Journal of Strategic Management, Vol. 17, pp. 69-88, 2017.

[58] B.C. Villari and S. Jharkharia, "Critical Success Factors for ERP Implementation: A Classification". Twelfth AIMS International Conference on Management, Kozhikode, Kerala, India, pp. 1013-1022, 2014.

[59] J. Desalegn and A. Petterson, "Investigation of Critical success factors for ERP implementation: A user perspective". Dissertation, Jonkoping University, 2018.

[60] E. Reitsma, P. Hilletofth and U. Mukhtar, "Enterprise Resource Planning System Implementation: a User Perspective". Operations and Supply Chain Management, Vol. 11, pp. 110–117, 2018.

[61] J. Motwani, R. Subramanian and P. Gopalakrishna, "Critical factors for successful ERP implementation: Exploratory findings from four case studies". Computers in Industry, Vol. 56, pp. 529–544, 2005.

[62] J. Ram, D. Corkindale and M. Wu, "Implementation critical success factors (CSFs) for ERP: Do they contribute to implementation success and post-implementation performance". Int. J. Production Economics, Vol. 144, pp. 157–174, 2013.

[63] J. Wu, "Critical Success Factors for ERP System Implementation", in L.D. Xu, A.M. Tjoa and S.S. Chaudhry (ed) Research and Practical Issues of Enterprise Information Systems II. The International Federation for Information Processing. Boston: Springer, pp. 739-745, 2007.

[64] K. Al-Fawaz, Z. Al-Salti and T. Eldabi, "Critical success factors in ERP implementation: A review". European and Mediterranean Conference on Information Systems, EMCIS 2008, Brunel University, Uxbridge, United Kingdom, pp.1-9, 2008.

[65] K. Elmeziane, S. Chuanmin and M. Elmeziane, "The Importance of Critical Success Factors of Enterprise Resources Planning Implementation In China". Business Management Dynamics, 2011.

[66] M. Moohebat, A. Asemi and M.D. Jazi, "A Comparative Study of Critical Success Factors (CSFs) in Implementation of ERP in Developed and Developing Countries". International Journal of Advancements in Computing Technology, Vol. 2, pp. 99-110, 2010.

[67] M. Saleh, M. Abbad and M. Al-Shehri, "ERP Implementation Success Factors in Saudi Arabia". International Journal of Computer Science and Security (IJCSS) 7:15-30, 2013.

[68] O. Francoise, M. Bourgault and R. Pellerin, "ERP implementation through critical success factors' management". Business Process Management Journal, Vol. 15, pp. 371-394, 2009.

[69] P. Chatzoglou, D. Chatzoudes and G. Apostolopoulou, "Examining the Antecedents and Outcomes of ERP Implementation Success: An Explanatory Study", in E. Ziemba (Ed.): AITM 2016/ISM 2016, LNBIP 277, 2017, Springer International Publishing AG, pp. 157–178, 2017.

[70] S. Dezdar and A. Sulaiman, "Critical Success Factors for Erp Implementation: Insights from a Middle-Eastern Country". Middle-East Journal of Scientific Research. Vol. 10, pp. 798-808, 2011.

[71] S.M. Jafari, M.R. Osman, R.M. Yusuff and S.H. Tang, "Erp systems implementation in Malaysia: The importance of critical success factors". International Journal of Engineering and Technology, Vol. 3, pp. 125-131, 2006.

[72] T.R. Bhatti, "Critical Success Factors for the Implementation of Enterprise Resource Planning (ERP): Empirical Validation". Second International Conference on Innovation in Information Technology (IIT'05), Dubai, India, UAE, pp.1-10, 2005.

[73] A. Gandhi, "Critical Success Factors in ERP Implementation and their interrelationship using TISM and MICMAC Analysi". Indian Journal of Science and Technology, Vol. 8, pp. 138–150, 2015.

[74] A. Gianopoulos, "Critical Success Factors in ERP Systems Implementation: the case of medium and small sized Enterprises". Journal of Business Management and Applied Economics, Vol. 4, pp. 01-16, 2015.

[75] C. Doom, K. Milis, S. Poelmans and E. Bloemen, "Critical success factors for ERP implementations in Belgian SMEs". Journal of Enterprise Information Management, Vol. 23, pp. 378-406, 2010.

[76] C. Leyh, "Critical Success Factors for ERP Projects in Small and Medium-sized Enterprises – The Perspective of Selected German SMEs". Federated Conference on Computer Science and Information Systems, Warsaw, Poland, IEEE, pp.1181–1190, 2014.

[77] C.L. Gurudatt, B. Ravishankar and R.V. Jayathirtha, "Gross Deviation Analysis (GDA) of delay as a tool for effective ERP implementation in Indian SMEs, Industrial Engineering Journal, Vol. 7, pp. 1-14, 2014.

[78] L. Ganesh and A. Mehta, A "Critical success factors for successful enterprise resource planning implementation at Indian SMEs". International Journal of Business, Management and Social Sciences, Vol. 1, pp. 65-78, 2010.

[79] T.C. Loh and S.C.L Koh, "Critical elements for a successful enterprise resource planning implementation in small- and medium-sized enterprises". International Journal of Production Research, Vol. 42, pp. 3433–3455, 2004.

[80] V. Hasheela-Mufeti and K. Smolander, "What are the requirements of a successful ERP implementation in SMEs? Special focus on Southern Africa". International Journal of Information Systems and Project Management,Vol. 5, pp. 5-20, 2017.

[81] A.Y. Akbulut and J. Motwani, "Critical Factors in the Implementation and Success of Enterprise Resource Planning (ERP)". Seidman Business Review, Vol. 11, pp. 20-23, 2005.

[82] H.S. Woo, "Critical success factors for implementing ERP: the case of a Chinese electronics manufacturer". Journal of Manufacturing Technology Management, Vol. 18, pp. 431–442, 2007.

[83] I. Zouaghi and A. Laghouag, "Aligning Key Success Factors to ERP Implementation Strategy: Learning from a Case Study". 4th International Conference on Information Systems, Logistics and Supply Chain, Creative Logistics for an uncertain world, Quebec (Canada), 2012.

[84] S. Khan and M. Anwar, "Analysis of Critical Success Factors (CSFs) for Implementation of Enterprise Resource Planning (ERP) in Manufacturing Industry". International Journal of Scientific and Engineering Research, Vol. 10, pp: 392-402, 2019.

[85] Y. Lin, M. Lee, H.P. Tserng, "Construction Enterprise Resource Planning Implementation: Critical Success Factors – Lesson Learning in Taiwan". 20th ISARC, Eindhoven, Holland, I.A.A.R.C, pp.623-628, 2003.

[86] A.I. ALdayel, M.S. ALdayel and A.S. Al-Mudimigh, "The Critical Success Factors of ERP implementation in Higher Education in Saudi Arabia: A Case Study". Journal of Information Technology and Economic Development, Vol. 2, pp. 1-16, 2011.

[87] M.A. Al-Hadi and N.A. Al-Shaibany, "Critical Success Factors (CSFs) of ERP in Higher Education Institutions". International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 7, pp. 92-95, 2017.

[88] J. Arvidsson and D. Kojic, "Critical Success Factors in ERP Implementation: The Perspective of the Procurement System User". Dissertation, Jonkoping University, 2017.

[89] V. Yildirim and A. O. Kusakci, "The Critical Success Factors Of Erp Selection And Implementation: A Case Study In Logistics Sector". Journal of International Trade, Logistics and Law, Vol. 4, pp. 138-146, 2018.

[90] A.A. Mengistie, P. Heaton and M. Rainforth, "Analysis of the Critical Success Factors for ERP Systems Implementation in U.S. Federal Offices". Innovation and Future of Enterprise Information Systems, Lecture Notes in Information Systems and Organisation. Berlin, Heidelberg: Springer, 2013.

[91] S. Rahayu and V.J. Dillack, "Key Success Factor for Successful ERP Implementation in State Owned Enterprises". International Journal of Engineering & Technology, Vol. 7, pp. 916-919, 2018.

[92] P. Garg, "Critical Success factors for Enterprise Resource Planning implementation in Indian Retail Industry: An Exploratory study". International Journal of Computer Science and Information Security (IJCSIS), Vol. 8, pp. 01-06, 2010.

[93] M. Agaoglu, E.S. Yurtkoru, A. Ekmekçi, "The effect of ERP implementation CSFs on business performance: an empirical study on users' perception". International Conference on Leadership,

Technology, Innovation and Business Management, Elsevier Ltd, pp. 35 – 42, 2015.

[94] M. Pourjabbar and M. Delgir, "Critical Success Factors (CSF) for Enterprise Resource Planning (ERP) in Financial Institutes (Case Study: Bank Saderat Iran)". Int. J. of Comp. & Info. Tech. (IJOCIT), Vol. 5, pp. 77-88, 2018.

[95] Z. Saadat and A. Afsharnejad, "Critical Success Factors in Implementation of Enterprise Resource Planning Systems: A Case of Golrang Company in Iran". IOSR Journal of Business and Management (IOSR-JBM), Vol. 18, pp. 32-37, 2016.

[96] D.L. Olson, and F. Zhao, "International Federation for Information Processing", in A.M. Tjoa, L. Xu and S. Chaudhry (ed), Research and Practical Issues of Enterprise Information Systems, Boston: Springer, pp. 569-578, 2006.

[97] C. Barth and S. Koch, "Critical success factors in ERP upgrade projects". Industrial Management and Data Systems, Vol. 119, pp. 656-676, 2019.

Autoencoder based Semi-Supervised Anomaly Detection in Turbofan Engines

Ali Al Bataineh[1], Aakif Mairaj[2], Devinder Kaur[3]
EECS Department, College of Engineering
University of Toledo, Toledo
OH, 43606

***Abstract*—This paper proposes a semi-supervised autoencoder based approach for the detection of anomalies in turbofan engines. Data used in this research is generated through simulation of turbofan engines created using a tool known as Commercial Modular Aero-Propulsion System Simulation (CMAPSS). C-MAPSS allows users to simulate various operational settings, environmental conditions, and control settings by varying various input parameters. Optimal architecture of autoencoder is discovered using Bayesian hyperparameter tuning approach. Autoencoder model with optimal architecture is trained on data representing normal behavior of turbofan engines included in training set. Performance of trained model is then tested on data of engines included in test set. To study the effect of redundant features removal on performance, two approaches are implemented and tested: with and without redundant features removal. Performance of proposed models is evaluated using various performance evaluation metrics like F1-score, Precision and Recall. Results have shown that best performance is achieved when autoencoder model is used without redundant feature removal.**

Keywords*—Anomaly detection; autoencoder; bayesian hyperparameter tuning; turbofan engine

I. Introduction

Anomaly detection refers to identification of those situations, which do not conform to pre-defined normal behavior of the system under consideration. Timely detection of such anomalies in machinery has many applications, which include reduced downtime, reduced maintenance cost and less safety hazards. Increasing focus on reliability and maintenance of complex systems like turbofan engines demands intelligent and autonomous ways to manage the health of these safety critical systems [1]. One such way is to deploy autonomous anomaly detection to monitor the health of turbofan engines. Timely detection of anomalies in turbofan engines can enable its operators to take corrective actions timely and prevent catastrophic failures.

In recent years, there is an increasing interest in data driven anomaly detection techniques. Based on nature of available dataset, data driven anomaly detection techniques are classified into three types: supervised, unsupervised, and semi supervised [2].

Supervised anomaly detection techniques require true labels for all training instances: normal as well as anomalous. Bayesian networks in supervised setup are used for intrusion detection in [3].

Researchers have also used multilayer perceptron (MLP) for detection of normal and attack connections [4]. In addition to detection of attack connections, MLP based approach also helps to identify type of attacks. Researchers have also used support vector machines (SVM) and decision trees for detection of anomalies in various applications [5], [6]. One of the biggest challenges in supervised anomaly detection approaches is non availability of representative labels, especially for anomalous class [7]. Another issue is that anomalous class has far fewer instances than normal class instances. This issue of imbalance class distribution is addressed in [8].

Unsupervised anomaly detection algorithms do not require true labels of data instances for training. Basic assumption for these algorithms is that only small percentage of data belongs to anomalous class. Unsupervised anomaly detection approaches classify infrequent data instances as anomalous [2]. K-means clustering based sliding window approach is used to detect anomalies in discrete manufacturing process by [9].

Another clustering algorithm called Fuzzy C-Means (FCM) is also used by researchers for unsupervised anomaly detection [10]. In FCM, one data instance can belong to more than one clusters.

Some researchers have also used expectation maximization (EM) meta algorithm for unsupervised anomaly detection [11]. EM is again a soft clustering technique which maximizes the value of certain parameter in a probabilistic model. One of the biggest issues with these clustering-based anomaly detection techniques is high false positive rate.

In [12], different clustering algorithms were used for intrusion detection and it was observed that false positive rate was more than 20%. Such high false positive rate makes clustering based anomaly detection techniques unviable for some of the real-world problems [12].

Another challenge in unsupervised anomaly detection techniques is the assumption that only small proportion of data represents the anomalous class. In situations where this assumption is not true, these unsupervised anomaly detection approaches may suffer from bad performance.

Semi-supervised anomaly detection approaches require that true labels should be available for only those data instances which represent the normal behavior of system under consideration. As these semi-supervised approaches do not require the labels for anomalous data instances, therefore these approaches are widely applicable in practice as compared to supervised anomaly detection approaches.

In [13], a kernel principal component analysis (PCA) based approach is used to deploy semi-supervised anomaly detection on a spacecraft. This approach first projects original data on to a lower dimensional space and then reconstructs it from that lower dimensional space. This reconstructed data from lower dimensional space is supposed to represent the true nature of data (independent of noise). The reconstruction error between original data and reconstructed data is then used to detect anomalies. However, the performance of this approach is highly dependent upon type and hyperparameters (e.g., degree in case of polynomial kernel) of kernel chosen. In some cases, choice of kernel requires domain knowledge, which is not easily available in all the cases.

In [14], researchers have proposed a semi-supervised support vector machines (S3VMs) for detection of anomalies in complex systems. It has been observed that semi supervised SVMs give high false positive rate when tuned and trained in a semi-supervised setup [15]. In addition to this, curse of dimensionality is also an issue when these SVMs are used with high dimensional data. In recent years, there is an increasing interest in use of artificial neural networks for various applications. Autoencoder is also a type of neural network, which is designed to learn reconstruction of input data [16].

Unlike PCA based approaches, autoencoders perform hierarchical dimensionality reduction by stacking up multiple hidden layers. By reducing number of neurons in subsequent hidden layers, each hidden layer tends to learn the true nature of the data. So, by using multiple hidden layers in auto-encoder framework, more abstract features can be extracted, and better reconstruction of data can be achieved without any dependency on domain knowledge.

In this research, an autoencoder based semi-supervised anomaly detection approach is used to detect anomalies in turbofan engines. As explained earlier, the main advantage of using autoencoders for anomaly detection is that they require only normal data for training and their performance is also not dependent on any user defined parameters (e.g., kernel type).

The rest of the paper is organized as follows. In section II, adopted methodology is explained in detail. Dataset used in this research is explained in section III. Section IV contains the implementation details and results. Conclusion of this research is presented in section V.

II. Methodology

As stated earlier that an autoencoder based anomaly detection approach is used in this research. A detailed explanation of the adopted methodology is given in this section.

A. Autoencoders

Autoencoder is an artificial neural network, which is trained to learn the reconstruction of input signal. Internally, an autoencoder consists of two parts: encoder and decoder. First of all, input signal $x \in [0,1]^d$ is mapped to a hidden representation $y \in [0,1]^{d'}$ through an encoding function $f(x)$. This hidden representation y is also known as the code of autoencoder. Here d and d^0 represents the dimensions of x and y respectively. Hidden representation y or code is then mapped back to reconstruction $z \in [0,1]^d$ through decoding function $g(y)$. The dimension of reconstruction z is same as of x. Here z should be considered as prediction of x by an autoencoder having code y. This structure of autoencoder is presented in Fig. 1.

Mathematical expressions of encoder and decoder functions are presented in in Eq. (1) and Eq. (2), respectively.

$$y = f(x) = \sigma(W_{xy}x + b_{xy}) \quad (1)$$

$$y = g(y) = \sigma(W_{yz}y + b_{yz}) \quad (2)$$

Here W represents the weight, b represents the bias and represents the nonlinear activation function of neural network. Learning process of an autoencoder involves the minimization of loss between x and $g(f(x))$. Loss function used in this research is the mean squared error (*MSE*). This loss function L is presented in Eq. (3).

$$L\left(x, g\left(f(x)\right)\right) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - g\left(f(x_i)\right)\right)^2 \quad (3)$$

where n represents total number of training examples.

An autoencoder which learns to perfectly reconstruct x everywhere (for all values of x) is not generally useful. Therefore, autoencoders are generally designed in such a way that they can perfectly reconstruct only those inputs which resemble to data in training set. One way to restrict perfect reconstruction everywhere is to constraint code y to have lower dimension than x. Type of autoencoder in which dimension of x is greater than the code y (i.e. $d > d^0$) is known as undercomplete autoencoder [17]. In this research we have used an undercomplete autoencoder to build an anomaly detection model.

In this paper, we have trained our autoencoder model on data representing normal behavior of system under consideration. A perfect autoencoder which is trained on normal data should be able to reconstruct only those inputs which are representative of normal behavior of system under consideration. Metric which is used to quantify the quality of reconstruction is reconstruction error. Reconstruction error can be measured in many ways. In this research we have used sum of squared error between x and z to measure the reconstruction error. This is presented in Eq. (4).

$$\text{reconstruction error (r)} = \sum_{i=1}^{k}(x_i - z_i)^2 \quad (4)$$

where k represents the dimension of input signal.

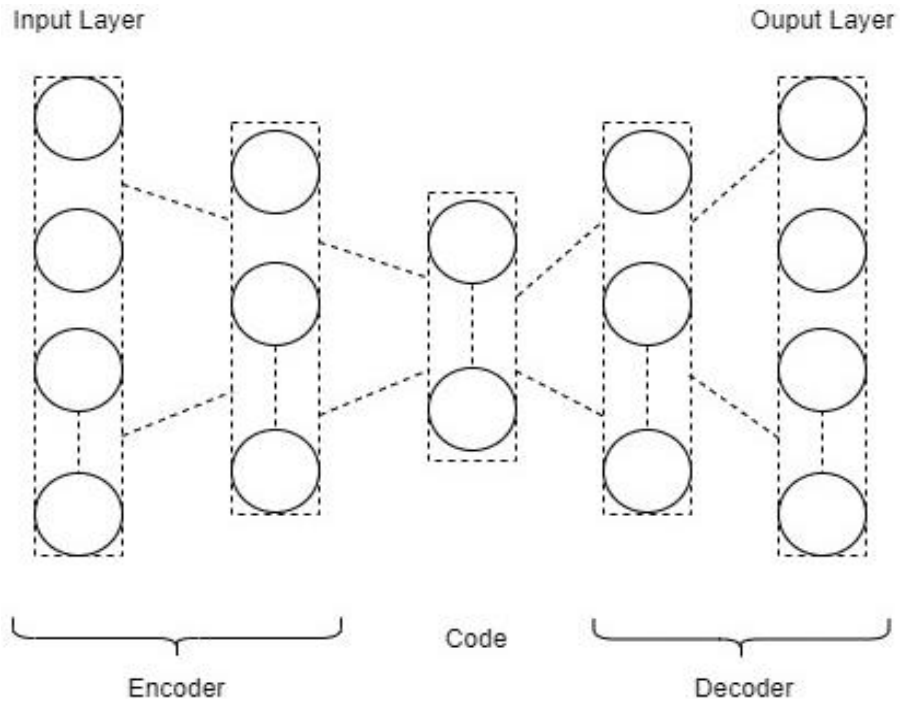


Fig. 1. An Autoencoder Structure.

An autoencoder which is trained on normal data should have a smaller reconstruction error on datapoints which are representative of normal behavior of the system and vice versa. Hence anomalies can be detected by simply using a threshold on reconstruction error. Data points having reconstruction error less than a certain threshold can be classified as normal, whereas data points having reconstruction error greater than a certain threshold can be classified as anomalies. This is presented in Eq. (5).

$$\begin{cases} r > Th\ Anomaly \\ r < Th\ Normal \end{cases} \quad (5)$$

where *Th* represents the threshold value.

The performance of an autoencoder model is highly dependent on the choice of hyperparameters, such as the number of layers, number of neurons and activation function, etc. The approach adopted for hyperparameter tuning in this research is Bayesian optimization, which is explained in the following section.

B. Hyperparameter Tuning using Bayesian Optimization

The aim of the hyperparameter tuning task is to find the set of hyperparameters, which gives the best performance (e.g., F1-score, R2-score, etc.) on the validation dataset for a specific model [18], [19].

For complex models like neural networks, manual tuning of these hyperparameters becomes intractable. There are some approaches like grid search and random search, which perform better than manual search in most of the cases. In the grid and random search, a search grid is being set up and the train-predict-evaluate cycle is executed for a different set of hyperparameters in a loop. However, these approaches are inefficient in the sense that they do not consider the performance of previously chosen hyperparameters while choosing the next set of hyperparameters. Grid and random search will continue to search the whole search space while being uninformed about the past evaluations. As a result, an ample amount of time is usually spent on the evaluation of bad hyperparameters.

In contrast to the grid and random search, the Bayesian approach for hyperparameter tuning considers past evaluations' results while choosing a new set of hyperparameters [20].

There are multiple approaches for Bayesian optimization in literature, differentiated based on the type of regression model and acquisition function they use. A probabilistic regression model is used to model the past evaluations by mapping hyperparameters to score on objective function. This regression model is also known as surrogate model in literature and is represented as $p(s/h)$ [20]. Here s represents the score on objective function and h represents the set of hyperparameters. Whereas next set of hyperparameters (from domain) in each iteration is chosen by optimizing an acquisition function, which uses $p(s/h)$ as a cheap surrogate of actual objective function.

In this work, we have used Tree-Structured Parzen Estimator (TPE) to build the surrogate model of objective function. Tree-Structured Parzen Estimator builds the surrogate model by using Bayes rule. Instead of directly calculating $p(s/h)$, it calculates $p(h/s)$ first and then use Bayes rule as in Eq. (6).

$$p\left(\frac{s}{h}\right) = \frac{p\left(\frac{h}{s}\right) * p(s)}{p(h)} \quad (6)$$

where $p(h/s)$ is probability of hyperparameter given the score and is expressed as in Eq. (7).

$$p\left(\frac{h}{s}\right) = \begin{cases} l(h) \; if \; s < s^* \\ g(h) \; if \; s \geq s^* \end{cases} \quad (7)$$

In Eq. (7), hyperparameters are divided into two distributions: $l(h)$ and $g(h)$. $l(h)$ contains all those set of hyperparameters for which score(s) of objective function is less than a certain threshold s^*, whereas g(h) contains all those set of hyperparameters, for which score(s) of objective function is greater than a certain threshold s^*. Acquisition function used in this research is Expected improvement (EI). The main task of the acquisition function is to find best set of hyperparameters based on surrogate model $p(s/h)$. Mathematical expression of Expected Improvement (EI) is given in Eq. (8).

$$EI_{s^*}(h) = \int_{-\infty}^{s^*} (s^* - s) p(d/h) ds \quad (8)$$

III. Case Study

In this work, we picked up a case study of anomaly detection in a simulated dataset of a turbofan engine [21]. The first principle model required to generate the data is built using a tool known as Commercial Modular Aero-Propulsion System Simulation (C-MAPSS).

C-MAPSS allows users to simulate various operational settings, environmental conditions, and control settings by varying various input parameters. In the chosen dataset, there is run-to-failure data of 249 engines simulated under six different operational settings. Some manufacturing variations and different initial degree of wear are being introduced in all 249 engines in order to make the scenario more real. Initial wear in engines is being introduced by varying efficiencies of various modules. In all the engines, a fault is introduced due to either of two failure modes: High Pressure Compressor (HPC) Degradation and Fan Degradation. At the start of each time series, the engine is running in a normal state and fault is introduced at some point in time, which then leads to engine failure in the future.

The Health state of each engine is measured by a set of 21 sensors installed on different modules of the engine. A list of all sensors is presented in Table I. In addition to these 21 sensor tags, three additional parameters are recorded to represent different operating states of the engine. A list of operational parameters is presented in Table II. The values of all these sensors and operational parameters are recorded at a frequency of one reading per engine cycle.

For semi-supervised anomaly detection, we are required to train our machine learning model on data representing the normal behavior of the system under consideration. As explained earlier, at the start of each time series, all engines are operating in a normal state, therefore in this work, the first 60 percent data of each time-series is considered as representative of the normal behavior of engines. The threshold of 60 percent is decided based on visual analysis of trends. Out of 249 engines, the first 60 percent data of 220 randomly chosen engines is used for training. Data of 20 engines is used as validation data (for hyperparameter tuning), and data of the remaining 19 engines is used for testing the performance of the trained model.

TABLE I. Sensors Names and their Units

Sensor	Description	Unit of Measure
T2	Fan inlet temperature	Rankine (°R)
T24	Low Pressure Compressor (LPC) outlet temperature	Rankine (°R)
T30	HPC outlet temperature	Rankine (°R)
T50	Low Pressure Turbine (LPT) outlet temperature	Rankine (°R)
P2	Fan inlet pressure	Pounds Per Square Inch Absolute (PSIA)
P15	Bypass-duct pressure	Pounds Per Square Inch Absolute (PSIA)
P30	HPC outlet pressure	Pounds Per Square Inch Absolute (PSIA)
Nf	Fan speed	Revolution Per Minute (rpm)
Nc	Core speed	Revolution Per Minute (rpm)
Epr	Engine Pressure Ratio	Nil
Ps30	HPC outlet static pressure	Pounds Per Square Inch Absolute (PSIA)
Phi	Fuel flow ratio to Ps30	pps/psi
NRF	fan corrected speed	Revolution Per Minute (rpm)
NRc	Core corrected speed	Revolution Per Minute (rpm)
BPR	Bypass ratio	Nil
farB	Fuel-air ratio of burner	Nil
htBleed	Bleed enthalpy	Nil

TABLE II. Operational Parameters

Operational Parameter	Description
Tr	Throttle Resolver Angle (TRA)
Al	Altitude
MN	Match Number

IV. Implementation

As detailed earlier, in this work, we have used semi-supervised autoencoders for detecting anomalies in turbofan engines. The overall approach can be divided into two phases: training and testing. In the training phase, training data representing the normal behavior of the engines is used to train the optimal autoencoder model. The effect of the removal of redundant features on model performance is evaluated using Pearson's correlation. If multiple features are found correlated with each other, only one is used in model training/testing. Features selected through this approach are listed in Table III. In this research, the results of both approaches (with and without redundant features removal) are presented.

The optimal architecture of autoencoder for given training data is discovered using the Bayesian optimization-based hyperparameter tuning approach. Search ranges of all the hyperparameters which are tuned using Bayesian optimization are given in Table IV. These search ranges are the same for both approaches (with and without redundant features removal). The final set of hyperparameters discovered by Bayesian hyperparameter tuning for both approaches is

presented in Table V. After figuring out the optimal architecture of the autoencoder; the following task is to train the autoencoder on normal training data. This is achieved using the backpropagation algorithm [22, 23].

TABLE III. FEATURES SELECTED AS RESULT OF REDUNDANT FEATURE REMOVAL

Feature Name	Feature Type
Throttle Resolver Angle	Operational Parameter
Altitude	Operational Parameter
Fan inlet temperature	Sensor
Engine Pressure Ratio	Sensor
HPC outlet static pressure	Sensor
Bypass ratio	Sensor

TABLE IV. HYPERPARAMETERS SEARCH RANGE FOR AUTOENCODER

Hyperparameter	Range
Number of Epochs	1-127
Batch Size	1-256
Number of Layers	[3, 5, 7, 9, 11]
Activation Function	Sigmoid, Softmax, Tanh, ReLU
Optimizer	Adam, Adadelta, RMS, SGD

TABLE V. FINAL HYPERPARAMETER SET

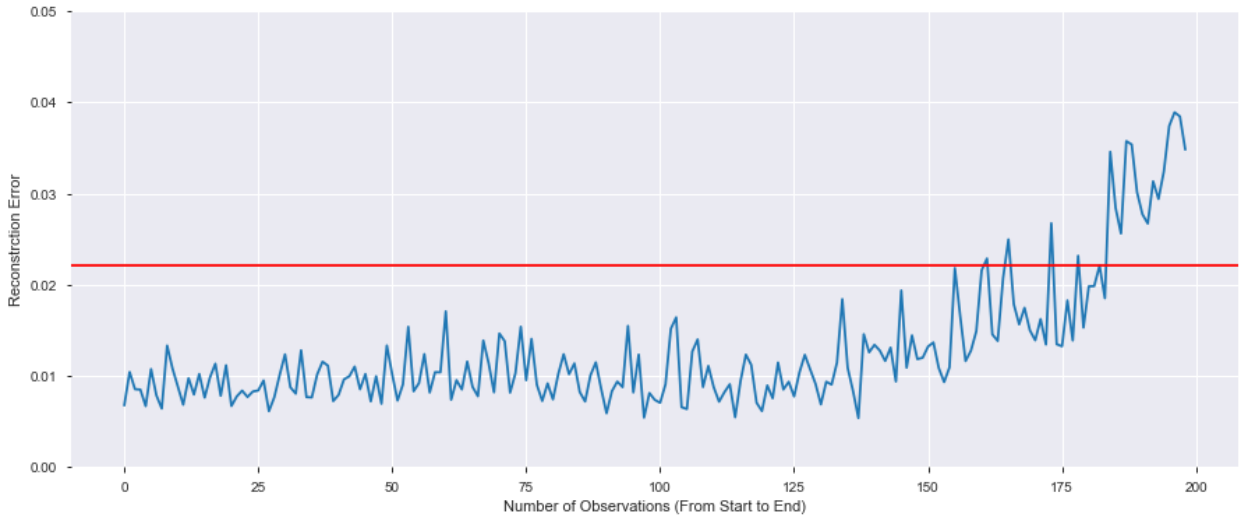
Hyperparameter	Without Feature Removal	With Feature Removal
Number of Epochs	64	22
Batch Size	148	10
Number of Layers	5	3
Number of Neurons per Layer	[24, 12, 8, 12, 24]	[6,3,6]
Optimizer	Adam	RMSprop

Once autoencoder is trained, next task is to compute the threshold value which is required for detection of anomalies during testing phase. For threshold calculation, all training samples (which are representative of normal behavior of the system) are scored through trained autoencoder model. For all scored samples, reconstruction error is computed and 98^{th} percentile of reconstruction error is selected as the threshold value for anomaly detection. Threshold value is a tunable parameter and 98^{th} percentile of reconstruction error is selected based on trial and error on validation dataset. Trained autoencoder model and calculated threshold value is then used to detect anomalies in testing phase.

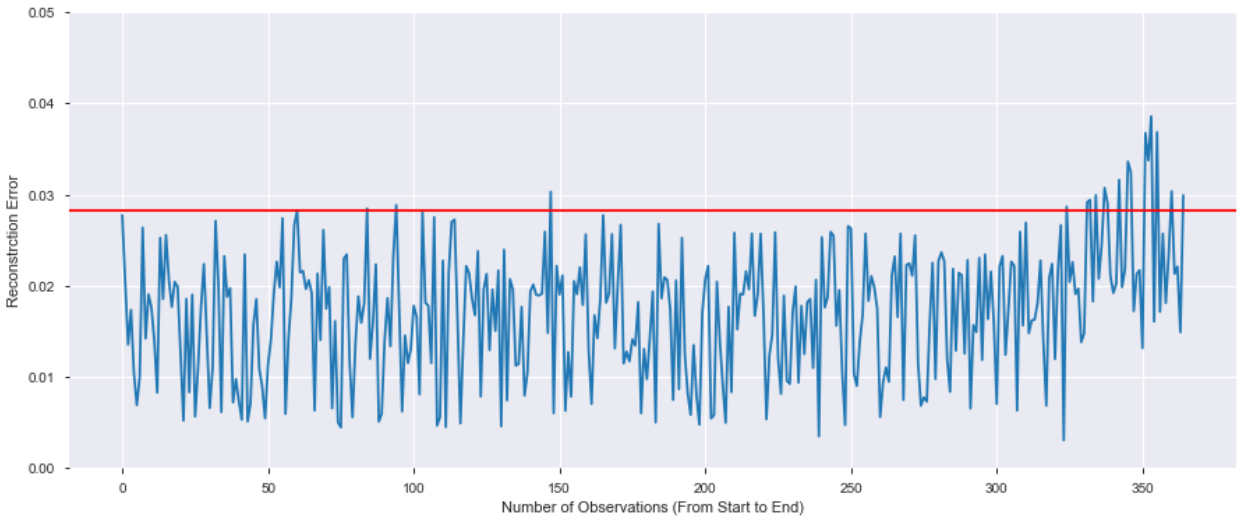
V. RESULTS

Performance of trained model is evaluated on test dataset consisting of run-to-failure data of 19 turbofan engines. As anomaly detection problem can be framed as a classification problem, therefore performance evaluation metrics used in this research are F1-score, Precision and Recall. For computing these metrics, true labels for each testing sample are required. This is achieved by assigning label 'Normal' to first 60% data and label 'Anomalous' to last 5% data of all testing (19) datasets. Reconstruction errors on two randomly chosen test examples for both the approaches are shown in Fig. 2 and Fig. 3. It is evident from both the Fig. 2 and Fig. 3 that reconstruction error increases as engine approaches failure (for both approaches). Threshold value computed by selecting 98^{th} percentile of reconstruction error on training set is also shown in the form of red horizontal line in following figures.

Results in Fig. 2 and Fig. 3 have shown that the best performance is achieved when no feature removal is applied. These results are also verified by F1-score computed on the test dataset for both the approaches. F1-score of approach without feature removal is 0.892 and for approach with redundant feature removal, F1-score is 0.813. These results are also presented in Table VI.

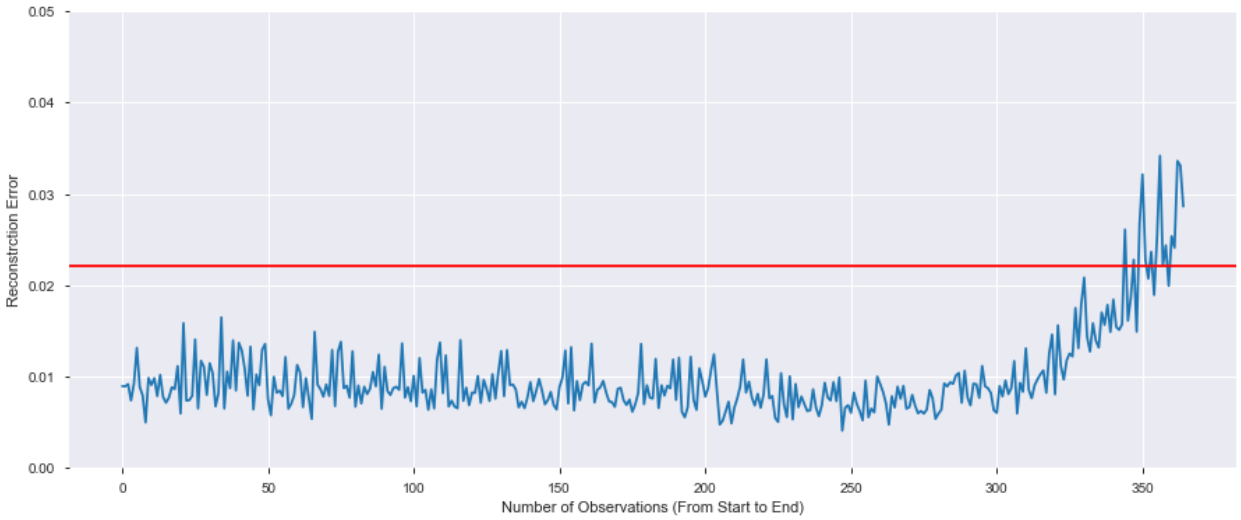


(a) Evolution of Reconstruction Error (without Redundant Features Removal).

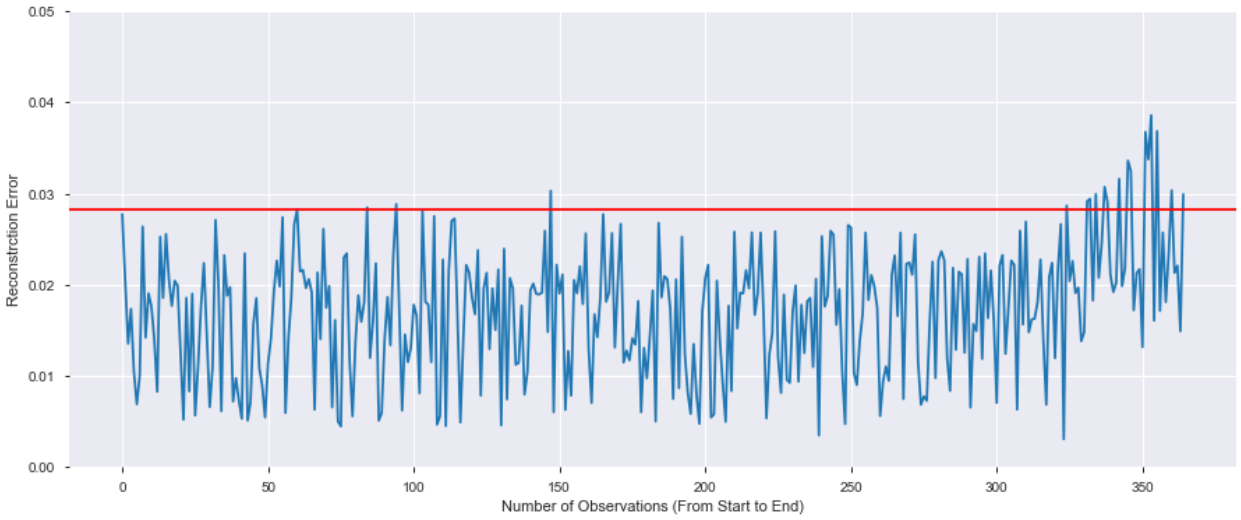


(b) Evolution of Reconstruction Error (with Redundant Features Removal)

Fig. 2. Evolution of Reconstruction Error for Engine # 235.



(a) Evolution of Reconstruction Error (without Redundant Features Removal)



(b) Evolution of Reconstruction Error (with Redundant Features Removal)

Fig. 3. Evolution of Reconstruction Error for Engine # 239.

TABLE VI. MODEL PERFORMANCE

	With Redundant Feature Removal	Without Redundant Feature Removal
F1-Score	0.813	0.892
Precision	0.704	0.896
Recall	0.611	0.724

VI. CONCLUSION

This paper proposed a semi-supervised autoencoder based anomaly detection approach to detect anomalies in turbofan engines. In the training phase, the autoencoder model is trained on data representing the normal behavior of turbofan engines. For tuning the architecture and hyperparameters of the autoencoder model, a Bayesian optimizing based approach was used. To study the effect of redundant features removal, two approaches are implemented and tested: with and without redundant features removal. For the removal of redundant features, Pearson's correlation was used to find a correlated set of features and one feature per set was used in training and testing. Performance evaluation metrics used in this research are F1-score, precision, and recall. Results have shown that the best performance is achieved when no redundant feature removal is applied. Our proposed approach has achieved F1score of 0.892, precision of 0.896 and recall of 0.724. This performance shows that autoencoders with optimal architecture can be a useful algorithm for the detection of anomalies in several real-world systems.

REFERENCES

[1] R. Mohammadi, E. Naderi, K. Khorasani, and S. Hashtrudi-Zad, "Fault diagnosis of gas turbine engines by using dynamic neural networks," in Turbo Expo: Power for Land, Sea, and Air, vol. 43987, 2010, pp. 365–376.

[2] R. Kandhari, V. Chandola, A. Banerjee, V. Kumar, and R. Kandhari, "Anomaly detection," ACM Comput. Surv, vol. 41, no. 3, pp. 1–6, 2009.

[3] P. G. Bringas and I. Santos, "Bayesian networks for network intrusion detection," Bayessian Network, editace A. Rebai, InTech, pp. 229–244, 2010.

[4] D. Rumelhart, G. Hinton, and R. Williams, "Learning internal representation by error propagation, parallel distributed processing," MIT Press, Cambridge, 1986.

[5] S. Peddabachigari, A. Abraham, C. Grosan, and J. Thomas, "Modeling intrusion detection system using hybrid intelligent systems," Journal of network and computer applications, vol. 30, no. 1, pp. 114–132, 2007.

[6] T.-J. Zhou, Y. Li, and J. Li, "Research on intrusion detection of svm based on pso," in 2009 International Conference on Machine Learning and Cybernetics, vol. 2. IEEE, 2009, pp. 1205–1209.

[7] S. Omar and A. Ngadi, "H. jebur, h.(2013)," Machine Learning Techniques for Anomaly Detection: An Overview. International Journal of Computer Applications, vol. 79, no. 2, pp. 33–41.

[8] A. Fernandez, S. Garc´ ´ıa, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, "Foundations on imbalanced classification," in Learning from Imbalanced Data Sets. Springer, 2018, pp. 19–46.

[9] B. Lindemann, F. Fesenmayr, N. Jazdi, and M. Weyrich, "Anomaly detection in discrete manufacturing using self-learning approaches," Procedia CIRP, vol. 79, pp. 313–318, 2019.

[10] J. Dunn, "A graph theoretic analysis of pattern classification via tamura's fuzzy relation," IEEE Transactions on Systems, Man, and Cybernetics, no. 3, pp. 310–313, 1974.

[11] C. Liu, "Maximum likelihood estimation from incomplete data via emtype algorithms," in Advanced Medical Statistics. World Scientific, 2003, pp. 1051–1071.

[12] I. Syarif, A. Prugel-Bennett, and G. Wills, "Unsupervised clustering approach for network anomaly detection," in International conference on networked digital technologies. Springer, 2012, pp. 135–145.

[13] R. Fujimaki, T. Yairi, and K. Machida, "An approach to spacecraft anomaly detection problem using kernel feature space," in Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005, pp. 401–410.

[14] P. Montague and J. Kim, "An efficient semi-supervised svm for anomaly detection," in 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, 2017, pp. 2843–2850.

[15] F. J. Huang and Y. LeCun, "Large-scale learning with svm and convolutional for generic object categorization," in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1. IEEE, 2006, pp. 284–291.

[16] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, 2014, pp. 4–11.

[17] A. R. Triki, R. Aljundi, M. B. Blaschko, and T. Tuytelaars, "Encoder based lifelong learning," in ICCV, 2017.

[18] A. Al Bataineh and D. Kaur, "Optimal convolutional neural network architecture design using clonal selection algorithm," International Journal of Machine Learning and Computing, vol. 9, no. 6, 2019.

[19] A. S. A. Bataineh, "A gradient boosting regression based approach for energy consumption prediction in buildings," Advances in Energy Research, vol. 6, no. 2, pp. 91–101, 2019.

[20] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, and D. D. Cox, "Hyperopt: a python library for model selection and hyperparameter optimization," Computational Science & Discovery, vol. 8, no. 1, p. 014008, 2015.

[21] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in 2008 international conference on prognostics and health management. IEEE, 2008, pp. 1–9.

[22] A. Al Bataineh and D. Kaur, "A comparative study of different curve fitting algorithms in artificial neural network using housing dataset," in NAECON 2018-IEEE National Aerospace and Electronics Conference .IEEE, 2018, pp. 174–178.

[23] A. Al Bataineh, "A comparative analysis of nonlinear machine learning algorithms for breast cancer detection," International Journal of Machine Learning and Computing, vol. 9, no. 3, pp. 248–254, 2019.

Definition of Unique Objects by Convolutional Neural Networks using Transfer Learning

Rusakov K.D[1], Seliverstov D.E[2]
V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65 Profsoyuznayastreet
Moscow 117997, Russia[1]
Plekhanov Russian University of Economics
36 Stremyannylane, Moscow, 117997, Russia[2]

Osipov V.V[3], Reshetnikov V.N[4]
Federal State Institution
"Scientific Research Institute for System Analysis of the Russian Academy of Sciences"
36 Nahimovskiipr-t, Moscow, 117218, Russia

***Abstract*—This article solves the problem of detecting medical masks on a person's face. Medical mask is one of the most effective measures to prevent infection with COVID-19, and its automatic detection is an actual task. The introduction of automatic recognition of medical masks in existing information security systems will allow quickly identify the violator of the mask regime, which in turn will increase security in a pandemic. The article provides a detailed analysis of existing solutions for face detection and automatic recognition of medical masks, method based on the use of convolutional neural networks was proposed. A distinctive feature of the new method is the use of two neural networks at once, using the RetinaFace neural network architecture at the face search stage and using the Resnet neural network architecture at the face mask recognition stage. It is shown that the use of transfer learning on scales, learned to work with faces, significantly accelerates learning and increases the accuracy of recognition. However, with this approach, there are some false positives, for example, when you try to cover your face with your hands, imitating a medical mask. Based on the study, we can conclude that the algorithm is applicable in the security system to determine the presence/absence of a medical mask on a person's face, as well as the need for additional research to solve the problems of false positives of the algorithm.**

***Keywords*—Recognition of medical masks; COVID-19; convolutional neural networks; retina face; Resnet**

I. Introduction

Nowadays the task of recognition the presence of a medical mask on a person's face has become very relevant in the condition of growing incidence of the new COVID-19 coronavirus infection. Medical mask is one of the effective measures for the prevention infection with COVID-19, its use minimizes the risk of spreading this disease.

Under the recognition task have a medical mask on the face of the man in this article means the following: on the input image to find metabologia all persons, for each person to determine the existence of face masks and give a certain confidence to this event.

The solution to the problem of recognizing a medical mask is also solved using convolutional neural networks. So, in [1], the authors solve the problem of face recognition in medical masks. In the real world, when a person tries to hide from systems such as video surveillance, having a face mask is one of the most common ways. If you have a medical mask on your face, the accuracy of facial recognition is reduced. The authors have conducted many studies on face recognition in various conditions, such as changes in posture or lighting, image degradation, and so on. The focus of this work is on medical masks, and especially improving the accuracy of facial recognition in medical masks. The authors solved the problem of detecting masked faces using a multitasking cascading convolutional neural network (MTCNN). Then, facial features are extracted using the Google FaceNet model. Finally, the classification task was performed by the authors using the Support Vector Machine (SVM). In [2], it is discussed that in recent years, face recognition has become a very difficult task due to various types of occlusion or masks, such as sunglasses, scarves, hats, and various types of makeup or disguise. All this affects the accuracy of facial recognition. Despite the fact that many algorithms for face recognition have been developed recently, which are widely used and provide better performance, little has been done in the field of face recognition in masks. Therefore, in this work, the authors chose a statistical procedure that is used in the recognition of unmasked faces, as well as used in the technique of face recognition in masks. RSA is a more efficient and successful statistical method that is widely used. For this reason, the RSA algorithm was chosen in this paper. Finally, there is also a comparative study for better understanding. In [3], the authors propose a new cascade structure based on convolutional neural networks, which consists of three carefully designed convolutional neural networks for detecting masked faces. The authors in their work talk about the applicability of the algorithm for tracking and identifying criminals or terrorists.

As you can see, the task confirms its relevance. Despite the relative simplicity of the task, recognizing medical masks on faces involves solving a number of non-trivial issues. Due to the widespread introduction and development of new information technologies, namely neural network approaches, in many areas of human life, there is a new task of automatic detection of a medical mask on a person's face, which will automatically monitor compliance with the mask mode.

II. Review of Existing Approaches to the Recognition of Medical Masks

From the analysis of various algorithms and methods for recognizing medical masks, it follows that in the structural

scheme of any image recognition method, as a rule, the following two typical types of mask recognition algorithms can be distinguished:

1) Search for a face in the image and then determine the mask on a person's face;

2) Search for a mask in an image without first identifying the face.

In addition, the construction of an algorithm for recognizing medical masks is based on a priori information about the subject area (in this case, on the characteristics of the person's face and the type of mask) and is corrected by empirical information that appears during the development of the algorithm.

The purpose of the face detection process (Fig. 1) is to localize all areas of the image that may contain a face, regardless of external lighting conditions, occlusion, etc. Despite the presence of distinctive features in the facial structure, this is quite a difficult task, since the features vary greatly depending on gender, skin color, and facial expression.

A significant number of factors can affect the detection of faces in a photo [4]:

1) Face position: the face in the image can be rotated at any angle of pitch, yaw, or roll.

2) Some parts of the face may be partially covered, which greatly complicates the ability to detect faces.

3) Different lighting conditions (the type the amount and direction of light sources, the color and brightness, the presence of shadows, color balance of camera, image distortion introduced by the optical system, etc.). For example, when lighting is used, the part of the face is very bright, while the other part is very dark, and it can influence the result.

4) The presence of normal or sunglasses, beards, moustaches, and various accessories make more errors in face detection.

5) The face size can change many times depending on the distance to the image.

6) Faces can be placed on different backgrounds: fixed, low contrast, noisy, etc., which can also make an error in the result of the face detection algorithm.

7) Different expressions and emotions: laughter, anger, surprise, etc., which can also affect the detection of faces.

Since face detection algorithms require a priori information about the face [5], the following categories of methods can be distinguished:

1) Feature-based methods (Fig. 2). For this category of methods, there are three areas, one of which they are based on: high-level feature analysis, low-level analysis, and methods based on form modeling. As a rule, the task of determining the face in an image is associated with the localization of this face on a complex background. Low-level analysis is based on feature segmentation based on image finite elements. Feature-based methods perform operations on image regions, taking into account the geometry and semantics of the face. The representativeness of these features is slightly higher in comparison with low-level analysis.

2) Image classification methods (Fig. 3). These methods are based on the construction of implicit facial patterns based on machine learning and modeling methods. This group of methods uses image recognition tools, presenting the face detection problem as a special case of the recognition problem.

After preliminary selection of the face on the image, it is binary classified for the presence or absence of a medical mask. In order to classify an image as «there is a mask/no mask», the face image is correlated with a feature vector calculated in some way. The most common way to get such a vector is to use the face image itself, with each pixel becoming a component of the vector. The main disadvantage of this method of representation is the extremely high dimension of the feature space, which increases the demand for computing power of the equipment. The advantage is a fairly high accuracy of classification. Thus, the task of automatically classifying images of people's faces by the values «there is a mask/no mask» is a typical task of learning from use cases [6].

The second method of determining the mask on the face is the direct determination of the medical mask on the image immediately, without first determining the faces.

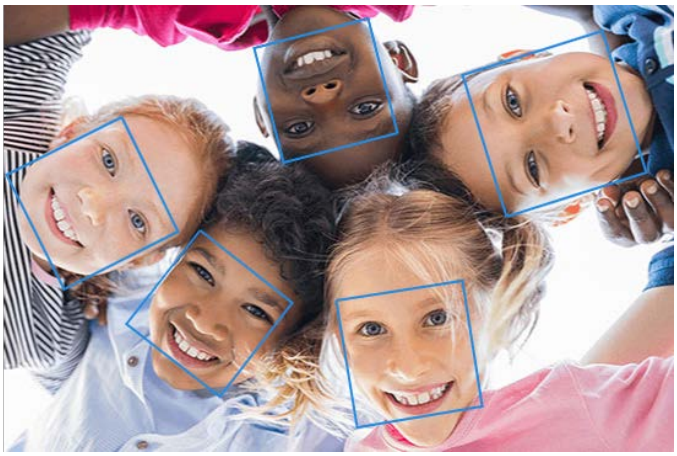


Fig. 1. Face Detection.

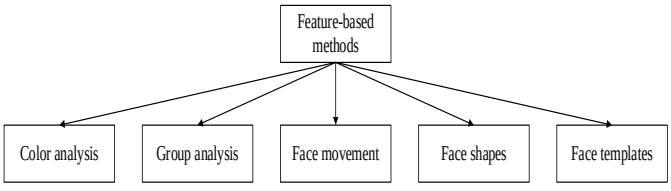


Fig. 2. Structure of Feature-based Methods.

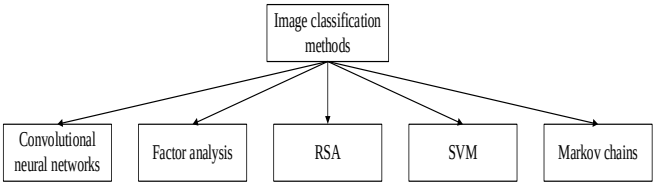


Fig. 3. Structure of Image Classification Methods.

From the analysis, it was revealed that to solve the problem of face detection, it is most preferable to choose the neural network approach as the main method. This choice was made based on the fact that this method provides:

1) The ability to detect a large number of faces in the image.

2) Ability to process images where the face angle is different from the front (with a yaw angle of up to 90°).

3) Ability to predict key points of the face to align it.

III. Architecture of the Medical Mask Recognition Solution

The proposed architecture consists of the following main elements: face detection by a fast one-step detector, face alignment, and face classification by a light convolutional neural network trained using transfer learning on scales for face recognition.

A. Detection and Face Alignment

At the first stage, the RetinaFace neural network architecture was chosen as the basis for the face search algorithm in the image. RetinaFace is a neural network architecture based on the detection and classification of objects at different levels of the "main highway" architecture. The Resnet50 network was used as the main neural network required for feature extraction. Distinctive feature: distinguishing faces in one run using the specified grid of Windows (default box) on the image pyramid (Fig. 4). In this way, both large and small faces are detected in a single network run.

In [7], it is shown that this solution solves the problem of finding and localizing a face quickly and accurately.

Since RetinaFace allows you to predict not only the localization of the face, but also the five key points of the face: eyes, nose and corners of the mouth, we can determine the angle of the face roll as follows:

$$\theta = \arctan(\frac{I}{M})$$

I – the distance between the eyes;

M – the distance between the corners of the mouth.

After getting information about the angle, we can align the image of the face so that the eye level is on a straight line parallel to the abscissa axis. In other words, as a result of this operation, all faces will be aligned.

B. Detection of Mask on Face

Finally, the process of recognizing the presence of a medical mask on the face s determined by learning the classifier in the form of a light convolutional neural network [8, 9, 10]. Convolutional layers transform the input image into a feature vector that is passed through a linear layer [11] of the following type (Fig. 5):

In the case of these classifications, the quality functional is obvious and, therefore, often used [12]:

$$Q(f_{\phi(x)}, I^{train}, I^{test})$$

which determines the percentage of correctly classified objects that are represented in the use cases of the I^{test} test sample constructed using the formalized φ(x) method of the I^{train} learning sample.

Algorithm 1 The Pseudo-code of Mask Detection on PyTorch

Input: image array (w,h,c)

```
dets=predict_face(image)
eye_angle=math.atan(dets[5][1]/dets[7][0])*180/math.pi
crop_image=ndimage.rotate(image,eye_angle)[dets[3]:dets[4],dets[1]:dets[2]]
out = predict_masks(img)
```

Output Class-wise affinity score

It is not common enough for training with random initialization of weights [13] to take place in practice, since a data set of sufficiently large and necessary size is rarely available. To speed up learning, the Transfer Learning technology [14, 15] is used: ready – made weights of the trained model are taken (usually ImageNet [16], containing 1.2 million images with 1000 categories), and then the weights of this trained model are used either as initialization or as a way to extract fixed features for the task of interest.

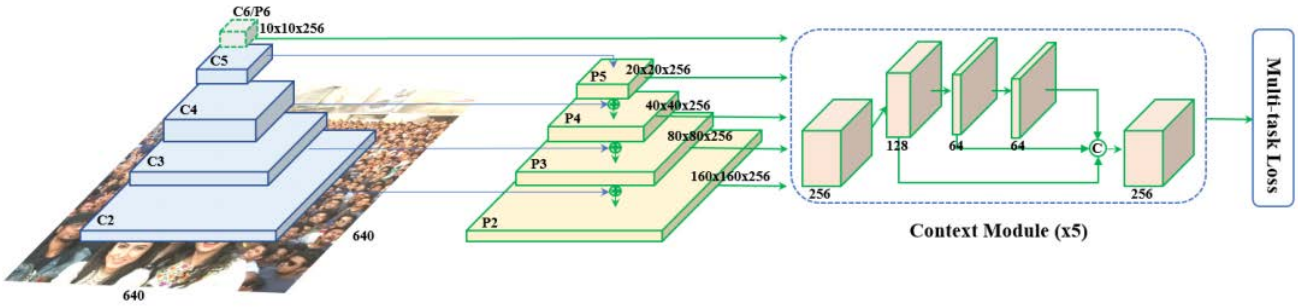


Fig. 4. Fully Connected Layer of the Medical Mask Recognition Neural Network Classifier.

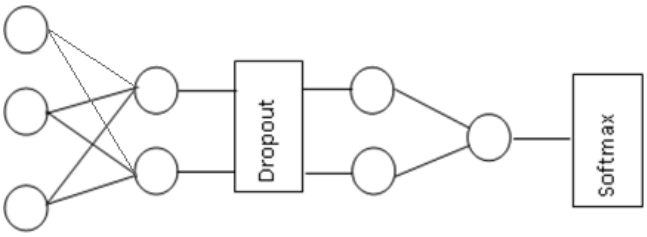


Fig. 5. Fully Connected Layer of the Medical Mask Recognition Neural Network Classifier.

IV. Result

The training was conducted through transfer learning. Two approaches were prepared: in the first approach, the weights of the ResNet neural network were initialized with weights trained on ImageNet, and the second approach: the weights of the ResNet neural network were initialized with weights trained to recognize faces on the ArcFace metric [17]. Comparative transfer learning graphs for different pre-trained weights are shown in Fig. 6.

The graphs show that the model initialized with ArcFace weights [8] converges faster than the model initialized with ImageNet weights.

The validation part of the data sets was used to measure the accuracy of facial recognition in medical masks. Table I shows the recognition accuracy. Accuracy and completeness parameters [18] are used for quality assessment at the selected threshold 0.5.

$$accurancy = \frac{TP}{TP + FP} \quad recall = \frac{TP}{TP + FN}$$

TP – true positive examples;

FP – false positive examples;

FN – false negative examples.

Fig. 7 shows examples of how medical mask recognition works.

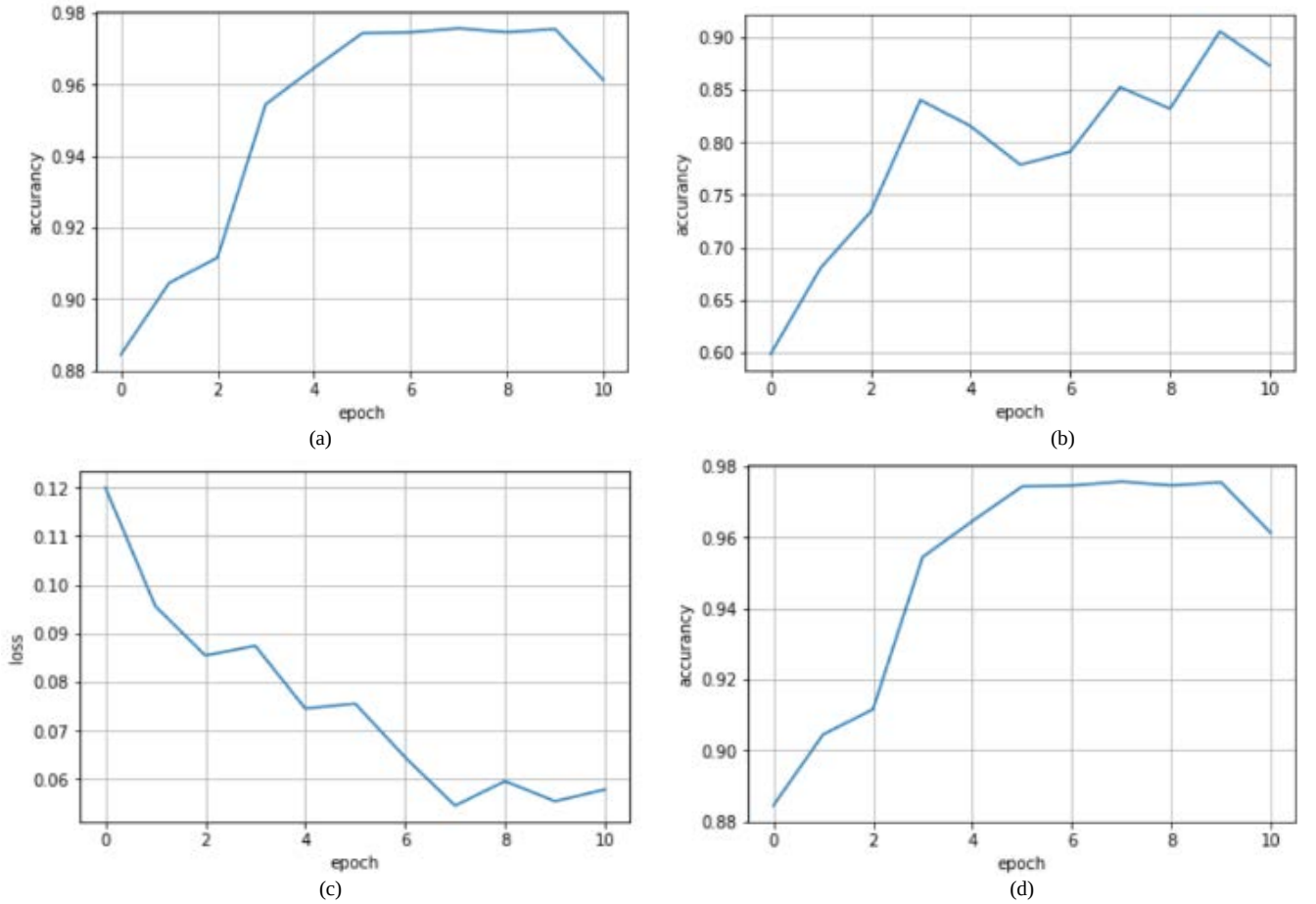


Fig. 6. Graphs of Changes during Validation: a) Errors when Initializing the Model with ImageNet Weights b) Accuracy when Initializing the Model with ImageNet Weights C) Errors when Initializing the Model with ArcFace Weights d) Accuracy when Initializing the Model with ArcFace Weights.

TABLE I. Quality Parameters of Recognition Medical Masks Model

	Accuracy of learning sample	Accuracy of test sample	Completeness of test sample
Recognition medical mask	99.8542	91.5621	92.4562

To demonstrate the performance and accuracy of the classification algorithm, there is an integral characteristic of the ROC curve [19] – another visualization method. The ROC-AUC curve shows what values accuracy and completeness take for different thresholds for making the «there is a mask/no mask» decision. The area under the curve [20] also characterizes the quality of the classification algorithm – the larger it is, the better. High accuracy determines a lower level of false positives.

Fig. 7 shows the ROC curve for the proposed algorithm. The area under the ROC curve is 0.98, which is a pretty good indicator.

Fig. 8 shows the examples of how medical mask recognition works. The left part of the drawing shows the original image with faces selected in the bounding box – the result of the RetinaFace neural network, the right part of the drawing shows the result of recognizing the presence/absence of a medical mask on the found face with some confidence.

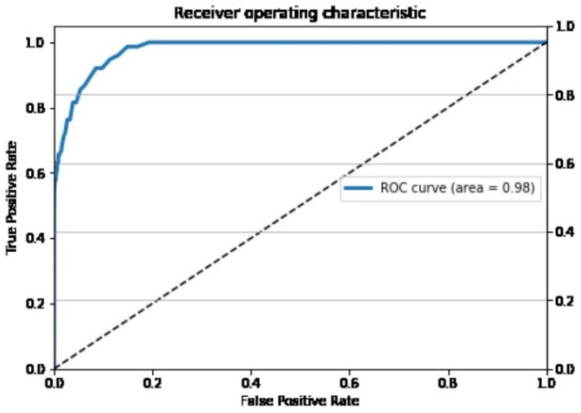


Fig. 7. ROC Curve of the Proposed Classifier.

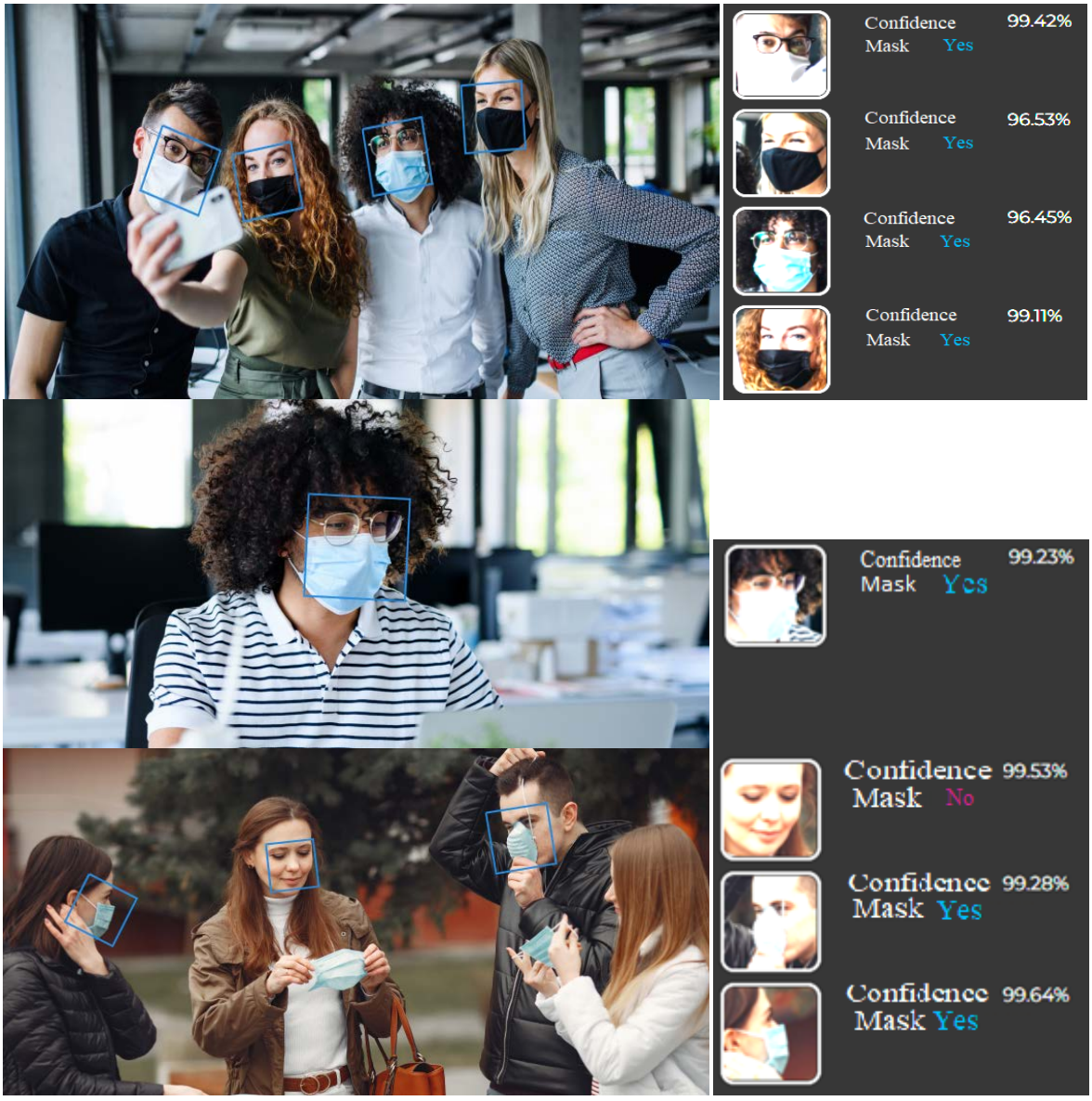


Fig. 8. Examples of Medical Masks Detector Functioning.

V. Comparison with other Methods

As can be seen from the analysis of existing developments in the field of medical mask recognition, most approaches are based on neural network methods [1, 2, 3]. The main difference between the proposed method and the existing ones is the use of pre-trained weights on faces to initialize the neural network before training, which significantly increases both the learning speed (up to 10 epochs) and the quality of the classifier. Table II shows the quality indicators of recognition of medical masks in the test sample.

For works [1,2], the accuracy is taken as the average for all test scenarios.

TABLE II. Verification results (%) of different Mask Detection Algorithms

	Accuracy of learning sample	Accuracy of test sample	Completeness of test sample
MTCNN + FaceNet + SVM	99.7862	82.4862	–
PCA	–	83.11	–
3 Cascade-CNN		86.6	87.8
RetinaFace + Resnet (TL)	99.8542	91.5621	92.4562

VI. Conclusion

This article presents a solution based on the use of two convolutional neural networks to predict the presence of a medical mask on a person's face. As can be seen from the results, the algorithm is able to accurately determine the presence of a medical mask, but there are some false positives, for example, when you try to cover your face with your hands, imitating a medical mask. The speed of the full processing cycle from obtaining a raw image to making a decision about the presence of a medical mask is less than 50 ms. The advantage of using this structure is its flexibility in terms of replacing classification algorithms with more efficient ones in the future. Also, in the course of the work, a unique database of faces in medical masks was collected, which is necessary for training a more effective classifier. The general disadvantage of the proposed algorithm is its computational complexity, and this factor cannot be ignored when building information systems that operate in real time. However, due to the rapid development of computer technology at the present time, these shortcomings can be overcome in the near future. In the future, it is planned to improve the accuracy of recognition of medical masks in more complex scenarios without loss in processing speed. Using graphics accelerators for neural networks is especially promising for solving such problems.

Acknowledgment

Publication is made as part of national assignment forSRISA RAS (fundamental scientific research 47 GP) on the topic No.0065-2019-0001 (AAAA-A19-119011790077-1).

References

[1] Ejaz, M. S., & Islam, M. R. (2019). Masked Face Recognition Using Convolutional Neural Network. 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI). doi:10.1109/sti47673.2019.9068044.

[2] Ejaz, M. S., Islam, M. R., Sifatullah, M., &Sarker, A. (2019). Implementation of Principal Component Analysis on Masked and Non-masked Face Recognition. 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT).doi:10.1109/icasert.2019.8934543.

[3] Bu, W., Xiao, J., Zhou, C., Yang, M., & Peng, C. (2017). A cascade framework for masked face detection. 2017 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM). doi:10.1109/iccis.2017.8274819.

[4] Chandrashekhar, P. Face Detection Techniques-A Review/ P. Chandrashekhar, N.D. Gopal // International Journal of Current Engineering and Technology, 2013. – P. 1809-1813.

[5] Erik, H. Face detection: A Survey/ H. Erik, K. Boon // Computer vision and image understanding, 2001. – Vol. 83, Issue 3. – P. 236-274.

[6] Cha, Z. A Survey of Recent Advances in Face Detection/ Z. Cha, Z. Zhengyou // Technical Report MSR-TR-2010-66, Microsoft research, Microsoft corporation, one Microsoft way Redmond, Multimedia, Interaction, and Communication (MIC) Group, 2010. – P. 1-17.

[7] Deng, Jiankang, J. Guo, Y. Zhou, Jinke Yu, I. Kotsia and S. Zafeiriou. "RetinaFace: Single-stage Dense Face Localisation in the Wild." *ArXiv* abs/1905.00641 (2019): n. pag.

[8] J. Deng, J. Guo, N. Xue and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 4685-4694, doi: 10.1109/CVPR.2019.00482.

[9] K. D. Rusakov, "Automatic Modular License Plate Recognition System Using Fast Convolutional Neural Networks," 2020 13th International Conference "Management of large-scale system development" (MLSD), Moscow, Russia, 2020, pp. 1-4, doi: 10.1109/MLSD49919.2020. 9247817.

[10] Rusakov K.D., Genov A.A., Shil S.Sh. An anti-spoofing methodology for a limited number of photos. Software & Systems. 2020, vol. 33, no. 1, pp. 054–060 (in Russ.). doi: 10.15827/0236- 235X.129.054-060.

[11] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), Antalya, 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

[12] S. Panigrahi, A. Nanda and T. Swarnkar, "Deep Learning Approach for Image Classification," 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA), Changsha, 2018, pp. 511-516, doi: 10.1109/ICDSBA.2018.00101.

[13] L. M. Waghmare, N. N. Bidwai and P. P. Bhogle, "Neural Network Weight Initialization," 2007 International Conference on Mechatronics and Automation, Harbin, 2007, pp. 679-681, doi: 10.1109/ICMA.2007.4303625.

[14] P. Natrajan, S. Rajmohan, S. Sundaram, S. Natarajan and R. Hebbar, "A Transfer Learning based CNN approach for Classification of Horticulture plantations using Hyperspectral Images," 2018 IEEE 8th International Advance Computing Conference (IACC), Greater Noida, India, 2018, pp. 279-283, doi: 10.1109/IADCC.2018.8692142.

[15] M. Shaha and M. Pawar, "Transfer Learning for Image Classification," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2018, pp. 656-660, doi: 10.1109/ICECA.2018.8474802.

[16] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[17] J. Deng, J. Guo, N. Xue and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 4685-4694, doi: 10.1109/CVPR.2019.00482.

[18] N. Seliya, T. M. Khoshgoftaar and J. Van Hulse, "Aggregating performance metrics for classifier evaluation," 2009 IEEE International Conference on Information Reuse & Integration, Las Vegas, NV, 2009, pp. 35-40, doi: 10.1109/IRI.2009.5211611.

[19] Jin Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," in IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 3, pp. 299-310, March 2005, doi: 10.1109/TKDE.2005.50.

[20] M. H. Ferris et al., "Using ROC curves and AUC to evaluate performance of no-reference image fusion metrics," 2015 National Aerospace and Electronics Conference (NAECON), Dayton, OH, 2015, pp. 27-34, doi: 10.1109/NAECON.2015.7443034.

Prelaunch Matching Architecture for Distributed Intelligent Image Recognition

Anton Ivaschenko[1], Arkadiy Krivosheev[2]
Computer Technology Department
Samara State Technical University
Samara, Russia

Pavel Sitnikov[3]
SEC "Open Code"
Samara, Russia

***Abstract*—The paper presents a multi-agent solution for dynamic combination of several artificial neural networks used for image recognition. As opposed to the existing methods there is introduced a dispatcher agent that provides prelaunch matching of possible pro-active identification algorithms through competition. The proposed solution was implemented to solve a problem of stream processing of photo images produced by a number of distributed cameras using an intelligent mobile application. It was probated and utilized in practice to capture the results of electrical meters that are manually monitored by a group of patrol personnel inspectors using hand held devices. Prelaunch matching architecture allowed increasing the quality of digits recognition using various neural networks depending on the operating conditions.**

***Keywords*—Multi-agent technology; artificial neural networks; image recognition; electricity meter data processing**

I. Introduction

Artificial neural networks are widely used for image recognition nowadays. However their application in practice is still concerned with the low versatility caused by filtering property that leads to the lack of multitasking. The image recognizing system needs to process a large variety of data in different operating conditions, and the better processing of one type leads to the failures in other cases. In addition to this the neural network is often being especially trained to reduce the noise affected by surroundings, which can be useful to identify the context sometimes.

Combination of several neural networks within the solid solution [1, 2] is a distinguished approach to improve the quality of recognition making the system adaptive to changing conditions. In this case a multi-agent paradigm becomes suitable and efficient to build an intelligent system with a distributed architecture [3, 4]. Despite the exploitability and potential capacity of this solution the ways of combination remain different, which makes it hard to develop a fully configurable and adaptive system in practice.

For example, electrical or gas meters surveying is so far concerned with visual monitoring of meter readings of various counters, which remains a challenging area for image recognition. This job is usually performed by professional patrol personnel aimed to check the visual status of the system, take the reading and snap a photo for validation. In order to succeed most of them use smartphones or tablet computers that could be upgraded with intelligent software. Variance in meter types and operating conditions makes it problematic to use single neural network.

To cover this gap there was developed a prelaunch matching software solution based on the concept of multi-agent architecture of distributed intelligence. In this case the software agents are introduced not to simulate or be deployed on hand held devices of monitoring inspectors. Each agent represents an autonomous recognizer specializing on processing of a certain type of meters. Additional Matcher agent is supplemented to scope out various objects and assigns them to the most corresponding recognition modules. It can be based by an extra neural network itself or implement preliminary defined rules or reasoning.

II. Materials and Methods

Technologies of intelligent image recognition using neural networks and positive examples of their practical applications are presented in [5, 6]. Image analysis and pattern recognition remain the challenging areas of classification and clustering using the modern technologies of artificial intelligence [7, 8].

Neural networks provide adequate and stable identification of real objects and items with a complex shape. Textual data (words and numbers) are identified with a sufficiently high accuracy even in case of fuzzy or washed-out picture. More specific solutions are presented in [9, 10].

The quality of data processing and analysis can be improved under the context of monitoring procedure and results of preceding identification. Several data sets are combined to analyze multiple layers of a system at once. This approach is widely used for medical data analysis and can be disseminated for a cyber-physical system [11], which can interlink all related data sets (e.g., images, text, measured values, scans) and offer visual analytics to support experts.

Distributed image recognition can be considered as a problem of the Internet of Things. Combination of the Internet of Things as a major data source and Big Data technologies is a powerful tool for information processing and analysis [12, 13] being successfully used at modern enterprises with distributed structure, i.e. electrical networks and oil pipelines.

Modern software architectures improve the performance of data processing in real time using an approach of parallel computing, multi-agent modeling and distributed decision-making support [14, 15]. This approach offers a way of designing adaptive systems with decentralization over

distributed and autonomous entities organized in hierarchical structures formed by intermediate stable forms. Its implementation in practice requires development of new methods and tools for supporting fundamental mechanisms of self-organization and evolution similar to living organisms (colonies of ants, swarms of bees, etc.).

Multi-agent technology can significantly enrich the adaptability of an intelligent system by making it possible to add new components and thus increase the number of options considered. Such an extensive development does not require the changes in an existing logic of already deployed components. The proposed approach is based on the experience of development of distributed image recognition solutions [16 – 18]. This paper describes one further step in this direction.

III. Solution Alternatives Considered

To develop a software solution for distributed intelligent image recognition taking the problem of electrical meters surveying as an example, the following options were analyzed.

Initially, a centralized architecture was created and put into operation, but according to the results of the work, it turned out that mobile devices did not always provide stable communication with the central server. There were also performance problems at times when, in addition to professionals, company management sent other employees to collect data. Such an increase in requests had a negative effect on recognition time. To solve these problems, the architecture of distributed recognition of readings was developed.

Centralized recognition implementation initially seems to become an obvious solution considering the requirements of neural network study and functioning. The logic of the module itself (see Fig. 1) is based on a fairly simple linear architecture. Images of meter readings made by inspectors using smartphones with the Android operating system are transmitted via the Internet to the central computing server, where they are processed in single-threaded mode (in the Main Thread stream) using the PreprocessImagePipeline method.

This method processes the input image in more than 1000 different ways in order to find outlines of readings on the image. Those processed images on which potentially suitable contours were found are transferred to the ProcessImagePipeline method.

The ProcessImagePipeline method takes a closer look at the resulting contours and eliminates the excess ones. Using the remaining contours, it cuts out the numbers from the processed images and passes them for recognition to the RecognizePipeline method. Inside the RecognizePipeline method, digital images are recognized using the neural network described below. For each of the previous stages of image processing options, a recognition result is obtained. Among the whole set of results, the best result is selected.

Based on the results of the operation of the centralized recognition module, it was decided to transfer the recognition process from the central server to specialized autonomous recognizers. These modules can be deployed either on inspectors' smartphones, or on dedicated servers on the cloud. The architecture of the distributed reading recognition module appeared is shown in Fig. 2.

In the main thread, the Controller method receives images from the camera, sends them to the UniversalRecognzier and Tracker methods, and also passes the results to the Aggregator method. The Aggregator method saves and analyzes the results, according to the results of the analysis, returns a report to the Controller method, which indicates whether the final result is ready or if you want to continue collecting results.

In a separate Recognition thread, the UniversalRecognizer method works, which is the centralized recognition module in which the number of processing of the input image is reduced to 10 and the type of processing, is randomly selected.

Another Tracking thread is running the Tracker method. It takes an input image and returns the difference with the previous image, i.e. direction in which direction the camera has moved. This information is subsequently passed through the Controller along with the recognition results to the Aggregator method.

There, it helps to compare the results between themselves, obtained at different intervals, as well as reset the old results with large changes in the position of the camera.

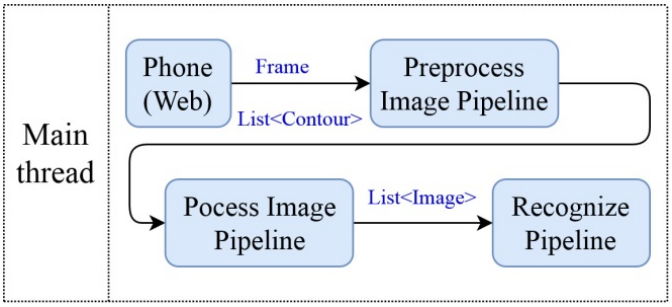


Fig. 1. Architecture of a Centralized Recognition Module.

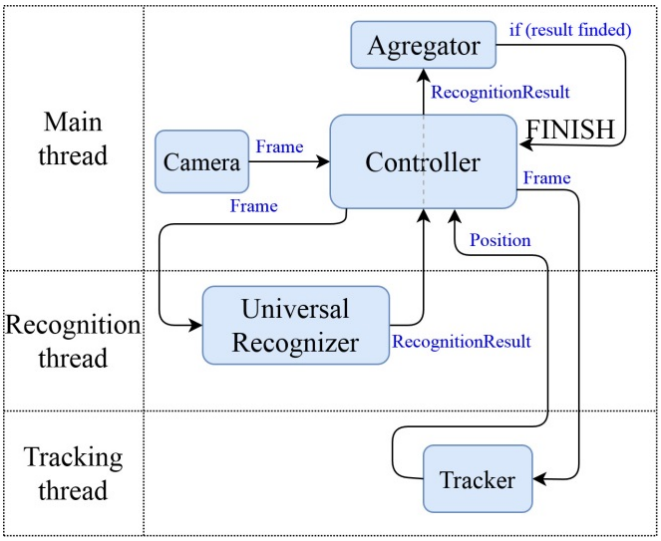


Fig. 2. Distributed Recognition Module Architecture.

IV. Prelaunch Matching Architecture

The achieved results were generalized by a concept of prelaunch matcher. The main idea is to split the system to distributed parts with autonomous behavior. These parts proactively communicate looking for the best combination of services to solve the initial problem. For the perspective of neural networks implementation in practice this approach gives an opportunity to combine several solutions instead of training one neural network, which might require significant costs and time.

Due to the different types of data generated and processed per unit time, classical architectures of the form "one task and one data type – one neural network" can no longer provide sufficient flexibility for new tasks on new data. To solve this problem there was developed two-layer system architecture for processing various types of data, see Fig. 3.

Considering the nature of the proposed approach it can be implemented using multi-agent technology as a software development paradigm. Data from camera is sent to the Matcher agent. Matcher agent operates on the basis of logic, neural networks and knowledge bases; its purpose is to choose the data computing strategy. If Matcher has selected several preprocessors, then each of them is questioned whether it is able to find his patterns on the sample from the received data, according to the results of the answers, the list of the preprocessors selected for processing is specified.

There are three types of Matcher agent strategies: auction, automatic dispatching, and competition:

- The auction is a survey of agents in order to find out which of them is capable of processing data. Based on the survey results, the Matcher agent selects Recognizer agents to process the current data set;
- When automatic dispatching, the Matcher agent independently chooses which agent to prefer for data processing;
- If the competition strategy is applied, then all available agents are involved in data processing, and the best ones are selected based on the processing results.

Agents use basic logic and neural networks in their work. Digit recognizers are used to find numbers in the image. Postprocessor tries to identify the digits that relate to the counter among all the digital symbols in view, the resulting data is transferred to the current storage. When the resulting data are found, the Frame tracker is connected, its purpose is to calculate the movements of the camera between the frames.

This data allows the Matcher agent to compare readings taken at different points in time. Also, as a result of moving the camera away from the counter, the Current storage may be reset. Over time, data from the counter taken from different angles accumulate in the current storage. Matcher selects the best counter results from all successful frames and generates the final recognition result.

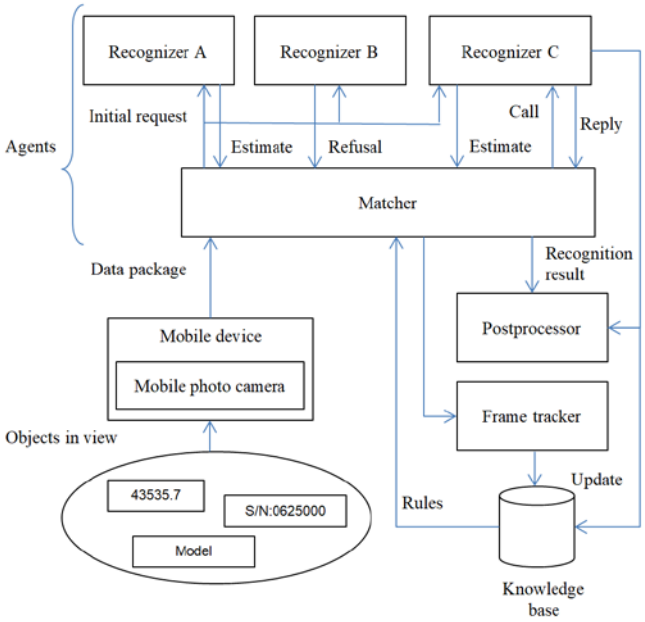


Fig. 3. Prelaunch Matching Architecture.

Auctioning strategy allows distributing the logic of decision-making between the components of the multi-agent architecture. When meeting an unknown object the system can organize a survey sending the requests to all the involved recognizing agents. They can either reply with and accepting of refusing answer based on preliminary understanding of the object type, or try to perform image recognition and reply in case of good quality of identification. After receiving the answers the Matcher can choose the best one and start negotiating with the corresponding recognizer on further identification.

Implementation of the multi-agent approach allows providing high autonomy of the recognizers, introduce new recognition algorithms with minor architecture changes and mix them in case of high level of uncertainty. Rather than other multi-agent implementations of distributed intelligent applications this solution does not require differentiation between the scopes of neural networks. Several different recognizers can be trained using the intersectional sets. Therefore such architecture remains open and provide an opportunity for permanent development by adding new recognizers without replacing the previous ones.

V. Implementation in Distributed Photo Surveying

The problem of electrical meters' photo surveying requires counter reading recognition. This task has to be successfully performed in various conditions, including weak light and darkening, overshadowing, obfuscation, occlusion and other failures. Currently on the market there are quite diverse types of meters both analog and digital. The vast majority of them cannot transmit the values electronically and require photo surveying.

To solve this problem there was developed special software for hand held devices, tablets and smartphones (see Fig. 4) supporting the operator to recognize the readings within the framework of the process collecting and further analyzing the level of energy consumption by the population of a particular region.

The tasks of meter reading analysis include a) identification of display panel and b) digit recognition for its indication evaluation. Targeting primarily the second task improves the quality of neural network application, but limits the prospects of its practical use. Targeting both tasks by one neural network introduces difficult training and low efficiency.

The described above approach was implemented, probated and tested for a convolutional neural network based on the LeNet architecture used to recognize the number symbols. When initializing the model weights, the Xavier Initialization was used.

For convolutional layers, the IDENTITY activation function was set, for a fully connected layer – RELU, for the Output layer – SOFTMAX. To exclude retraining, regularization L2 with parameter 0,0005 was used. The learning speed was set depending on the current iteration: every 200 steps until the 1000th iteration, the speed sequentially decreased from 0.06 to 0.001, after 1000 iterations the speed did not change.

An initial attempt was made to train a neural network based on a set of handwritten digits MNIST. Soon after the start of the work it became clear that it was not possible to get a good result with it. To create a dataset, about 1000 fonts were collected; based on these fonts, images of numbers in the amount of 10,000 copies were generated.

After augmenting this set with rotations and shifts, a dataset was created consisting of 196,000 images of digits, see Fig. 5.

It turned out that 30% of all digits “1” did not differ from each other, they were replaced by additional transformations of the remaining original “1”. The neural network was trained on the received dataset. The results of training the neural network are presented in Fig. 6.

To check the quality of the modules, a test kit was assembled, including 138 images of digital meters (777 digits in total) and 95 images of analog meters (534 digits in total). Because the distributed recognition module receives a video stream as an input, then a sequence of images was simulated for it by means of small offsets of the tested photo. Recognition results on a full set of images will be as follows (see Table I).

Some of the images in the test set were of poor quality, which is even difficult for a person to parse the readings. If you sort the set by image quality and choose the best half, the results can be improved.

One can see that implementation of a multi-agent architecture allows increasing the quality of digits recognition as a part of a distributed intelligent photo surveying solution.

Fig. 4. Application user Interface.

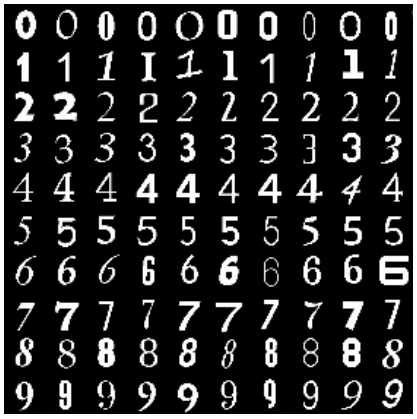


Fig. 5. Updated Training Dataset Images.

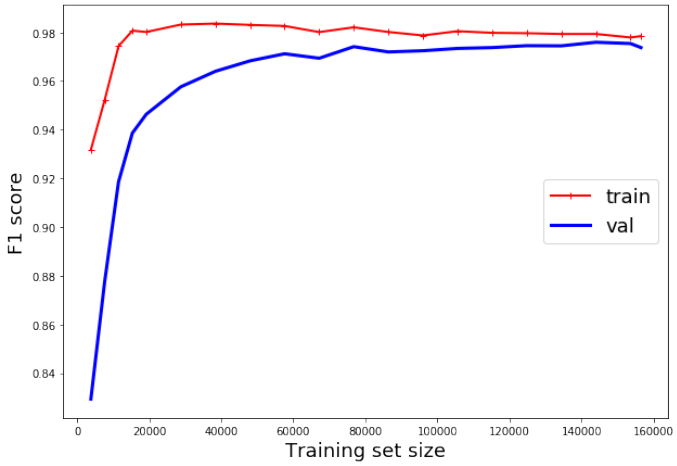


Fig. 6. Dependency of F1 from the Volume of the Dataset.

TABLE I. RECOGNITION QUALITY OF DIFFERENT SOLUTIONS

	Centralized recognition module (FULL dataset)	Distributed recognition module (FULL dataset)	Centralized recognition module (realistic dataset)	Distributed recognition module (realistic dataset)
Analog meters	392/534 = 73.4%	432/534 = 80.9%	197/212 = 92.9%	203/212 = 95.8%
LCD meters	611/777 = 78.6%	640/777 = 82.4%	368/378 = 97.3%	374/378 = 98.9%

VI. APPLICATION TO PRACTICE

The proposed solution was used in the specialized mobile application for photographing the readings of electricity meters, their transmission to the data processing center, recognition and operational analysis by the staff of a regional energy distribution company.

Search for words was carried out by Tesseract. By the relative position of the words, there was a meter mask, by the absolute position of the words in the photo there was an area with readings according to the data obtained from the mask.

As an alternative, a video stream from the phone's camera was implemented in the browser, a frame was displayed on top of the video stream, the user needed to point the phone at the meters that the readings fell into the frame, and then click on the recognition button.

Recognition works according to the following algorithm. The inspector takes a photo; it is processed by a series of filters that increase the quality and sharpness. Then the color image is converted to b / w in about 50 times in different ways, each image is subjected to the following actions:

- contours are searched for in the image, contours are outlined by rectangles;
- among the obtained rectangles, sequences are found that lie on one straight line and having the same size, among all sequences, the best one is taken according to an empirically selected formula with arguments "rectangle size" and "number of rectangles";
- according to the rectangles of the b / w image, sections with potential readings are cut out;
- the cut out images are processed and recognized, the recognition results are saved.

The results of implementation allowed performing the series of experiments and probating the solution in practice.

VII. CONCLUSION

Prelaunch matching architecture provides additional benefits from combining several neural networks into a solid intelligent solution for distributed photo surveying. As opposed to other solution it allows integrating several recognizers trained using the intersectional sets, which makes it open for permanent development by adding new recognizers without replacing the previous ones.

The results of implementation, testing and practical use illustrate the benefits of autonomous pre-processing of photo images using the mobile application with a multi-agent architecture. Next steps are considered with the study of possible agent strategies aiming an increasing the recognition quality in various conditions.

ACKNOWLEDGEMENT

The paper was supported by RFBR, according to the research project № 20-08-00797.

REFERENCES

[1] N. Ueda, "Optimal linear combination of neural networks for improving classification performance", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 2; 2000, pp. 207-215

[2] L. Lenc, P. Král, "Combination of neural networks for multi-label document classification", International Conference on Applications of Natural Language to Information Systems; 2017, pp. 278-282

[3] A. Quteishat, C.P. Lim, J. Tweedale, L.C. Jain, "A neural network-based multi-agent classifier system", Neurocomputing, vol. 72, issues 7–9; 2009, pp. 1639-1647

[4] R. Asadi, N. Mustapha, M. Sulaiman, "A framework for intelligent multi agent system based neural network classification model", International Journal of Computer Science and Information Security, vol. 5, no. 1; 2009, pp. 168–174

[5] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep learning, vol. 1, MIT press Cambridge; 2016

[6] M. Egmont-Petersen, D. de Ridder, H. Handels, "Image processing with neural networks – a review", Pattern Recognition 2002; 35 (10), pp. 2279–2301

[7] B. Pt, P. Subashini, "Optimization of image processing techniques using neural networks – A review", WSEAS Transactions on Information Science and Applications, 8 (8), 2011, pp. 300-328

[8] M. Jena, S. Mishra, "Review of neural network techniques in the verge of image processing", Advances in Intelligent Systems and Computing, vol 628. Springer Singapore, 2018, pp. 345-361

[9] Z. Zhao, P. Zheng, S. Xu, X. Wu, "Object detection with deep learning: a review", IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 11; 2019, pp. 3212-3232

[10] A. Arcos-Garcia, J. Alvarez-Garcia, L. Soria Morillo, "Evaluation of deep neural networks for traffic sign detection systems", Neurocomputing, 316. 10.1016/j.neucom.2018.08.009; 2018

[11] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?", Brain Informatics 2016, Volume 3, Issue 2. pp. 119-131

[12] N. Bessis, C. Dobre, "Big Data and Internet of Things: A roadmap for smart environments", Studies in computational intelligence, Springer 2014. 450 p.

[13] G. Fortuno, P. Trunfio, "Internet of Things based on smart objects: technology, middleware and applications", Springer-Verlag New York Inc., 2014, 250 p.

[14] M. Wooldridge, An introduction to multi-agent systems, John Wiley and Sons, Chichester; 2002, 340 p.

[15] V.I. Gorodetskii, "Self-organization and multiagent systems: I. Models of multiagent self-organization", Journal of Computer and Systems Sciences International 2012, vol. 51, issue 2, pp. 256-281

[16] A. Ivaschenko, A. Krivosheev, "Distributed processing of electrical meters surveying". 2020 Moscow Workshop on Electronic and Networking Technologies proceedings; 2020. pp. 1-4

[17] A. Ivaschenko, A. Krivosheev, P. Sitnikov, "Multi-agent solution for a distributed intelligent photo surveying". Proceedings of the 2019 European Simulation and Modeling Conference 2019. pp. 73-78

[18] Ivaschenko A., Sitnikov P., Surnin O. Accented visualization for Augmented Reality // Emerging Topics and Questions in Infocommunication Technologies, Cambridge Scholars Publishing, 2020. – pp. 74 – 97

STEM-Technology Example of the Computational Problem of a Chain on a Cylinder

Valery Ochkov[1], Konstantin Orlov[2]
National Research University
Moscow Power Engineering Institute
Moscow, Russia

Evgeny Barochkin[3]
Ivanovo State Power Engineering University
Ivanovo, Russia

Inna Vasileva[4]
N.E. Zhukovsky and Y.A. Gagarin Air Force Academy
Voronezh, Russia

Evgeny Nikulchev[5]
MIREA — Russian Technological University
Moscow, Russia

***Abstract*—An application of the STEM technology to the computational problem of the parameters of a closed chain (with and without load) thrown over a horizontal cylinder is considered. The numerical solution is found and its graphical interpretation is made by compiling a system of transcendental equations, as well as carrying out numerical optimization with constraints. The approximating analytical dependence is determined using the fitting functions. In the process of solving a number of concepts from mathematics, physics, computer science are examined. Some possibilities of using specialized mathematical packages (in particular, Mathcad) and of working on online platforms are shown. Additional problems options for using STEM technology are presented.**

Keywords*—STEM technology; math education; closed chain; Mathcad.

I. Introduction

Modernization of higher education is based on the use of new technologies [1]. For engineering and mathematical educations an important area is STEM technology training [2-6]. The problem of the parameters of a closed chain (with and without load) thrown over a horizontal cylinder, showing the application of the STEM technology in education [7–10], is considered.

At first the case of a closed chain without a pendant is analysed. The numerical solution and its graphical interpretation are given using the Mathcad Prime package. By working on online platforms an approximating function is constructed.

Then the more complex case of a closed chain with a pendant is modelled.

In terms of this article a well-known problem of a chain is examined using physics, mathematics, resistance of materials, hydro-gas dynamics, heat transfer, etc.

In solving this problem, a number of concepts from the following disciplines are used [11, 12]:

- Mathematics: function, derivative, integral, antiderivative, curve length, center of gravity of the curve, system of transcendental equations, optimization with constraints, approximation of numerical data;
- Physics: sagging chain with and without load, potential energy of a mechanical system;
- Computer science: work with user functions, numerical solution of transcendental equations, numerical optimization with constraints, smoothing table dependencies, work in Mathcad, solving problems using mathematical online platforms on the Internet.

II. Problem of the Parameters of the Shape of a Closed Chain Without a Pendant

A. Formulation of the Problem

A closed chain of length L is thrown onto a horizontal cylinder of radius R. Determine the shape of the sagging chain, whose length, naturally, is greater than the circumference of the cylinder: $L > 2\pi R$.

This problem is attractive because it can be formulated not only in words, but also tested physically. To do this, simply attach a cylindrical disk to the board in a classroom, put a closed chain on it (see Fig. 1) , take a picture of the sag of the chain with a digital camera, process the image on a computer, and then compare with the corresponding mathematical model (i.e. compare the real object with its digital double). This would make an excellent student laboratory activity.

Fig. 2 shows a scheme of the problem. Choose the origin at a point directly below the center of the circle at a distance h from it. This unknown quantity, along with the other two (xo and a) described below, is the object of the search. Another important search parameter is the angle θ — the angle at which the chain breaks away from the cylinder.

B. Numerical Solution

Nowadays, such problems are increasingly being solved numerically, with the use of mathematical computer programs, rather than analytically [13, 14]. In this article the Mathcad Prime computational package is used [15]. The requirement is to find a formula through which the angle θ is calculated depending on the ratio of L to R. Fig. 3 shows how the six

functions required to solve the problem are specified in Mathcad Prime.

The canonical catenary formula is a·cosh(x/a) [16, 17]. It has a minimum (sagging chain: a > 0) or maximum (arch: a < 0) at the point with abscissa equal to zero and the ordinate equal to a (see Fig. 2). The value of this parameter a needs to be found. The quantity a is also a certain "steepness" of the chain line: if a tends to zero, then the curvature of the chain also tends to zero, it turns into a stretched string. Another unknown is the variable x0: the abscissa of the separation point of the chain from the cylinder (see Fig. 1). There are two such points, but they are symmetric with respect to the ordinate axis.

Fig. 1. A Chain Wrapped around a Wall-Mounted Aneroid Barometer.

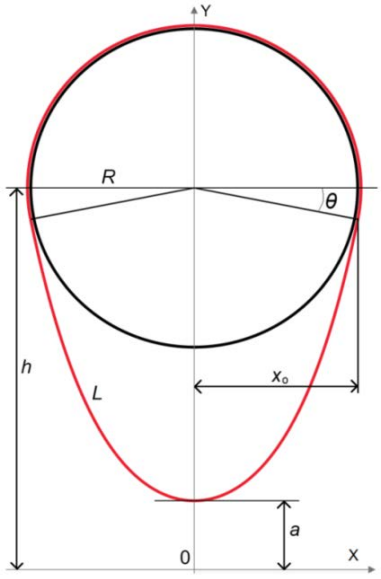


Fig. 2. Scheme of the Problem of the Chain on the Cylinder.

$$F_c(x,a) := a \cdot \cosh\left(\frac{x}{a}\right) \quad \text{Catenary}$$

$$F'_c(x,a) := \frac{d}{dx} F_c(x,a) \rightarrow \sinh\left(\frac{x}{a}\right) \quad \text{Derivative of a catenary}$$

$$S_c(x_o,a) := \int_{-x_o}^{x_o} \sqrt{1+F'_c(x,a)^2}\, dx \quad \text{Catenary length from } -x_o \text{ to } x_o$$

$$F_o(x,R,h) := h - \sqrt{R^2 - x^2} \quad \text{Lower semicircle formula}$$

$$F'_o(x,R) := \frac{d}{dx} F_o(x,R,h) \rightarrow \frac{x}{\sqrt{R^2 - x^2}} \quad \text{Derivative lower semicircles}$$

$$S_o(x_o,R) := R \cdot \left(\pi + 2\ \text{acos}\left(\frac{x_o}{R}\right)\right) \quad \text{Chain length lying on the cylinder}$$

Fig. 3. Auxiliary Functions of the Problem of the Chain on the Cylinder.

The definite integral, which sets the length of the catenary, could be simplified through finding the antiderivative. This work is now being done more and more often using the Internet, e.g. www.wolframalpha.com (see Fig. 4).

As can be seen from Fig. 4, it is not possible to take a specific integral using Wolframalpha (the original expression is returned). But if you work with an indefinite integral, then the problem will be solved (last line in Fig. 4.) After that, it suffices to use the Newton-Leibniz theorem and get the desired chain length formula from -xo to xo: 2a·sinh(xo/a) (see Fig. 5, where the csgn function is the sign of the argument, in this case it is positive).

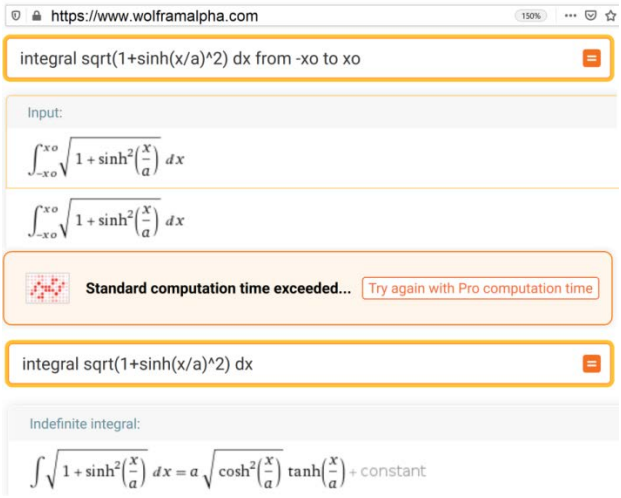


Fig. 4. Internet Search for the Primitive Function.

$$int(x,a) := a \cdot \sqrt{\cosh\left(\frac{x}{a}\right)^2} \cdot \tanh\left(\frac{x}{a}\right)$$

$$int(x_o,a) - int(-x_o,a) \xrightarrow{simplify}$$

$$\rightarrow a \cdot \sinh\left(\frac{x_o}{a}\right) \cdot \text{csgn}\left(\cosh\left(\frac{x_o}{a}\right)\right) + a \cdot \sinh\left(\frac{x_o}{a}\right) \cdot \text{csgn}\left(\cosh\left(\frac{x_o}{a}\right)\right)$$

Fig. 5. Simplification of the Catenary Length Formula.

Fig. 6 shows the continuation of Mathcad calculation: the input data and the calculation of an important parameter of the problem—the ratio of the chain length L to the circumference of the cylinder 2πR, on which the chain is thrown. This parameter can vary from one (the chain fits the cylinder tightly: θ = 90°—see Fig. 2) to infinity (the lashes of the chain sag almost vertically at the circle: θ = 0).

The solution of the problem is reduced to the numerical search for the root of a system of three transcendental equations. This operation in Mathcad (Solve block) is shown in Fig. 7: reasonable initial approximations to the solution are set, constraints are introduced (these are equations in this case, but, in general, inequalities may be included) and the Mathcad built-in function Find is called, which returns the numerical values of its arguments, which turn the equations into identities. Or, rather, almost in identities since the left and right sides of the equations differ by a small amount. Numerical methods for solving problems have another name—approximate methods.

The equations are:

- The length of the closed chain L remains constant and consists of two parts: the part lying on the cylinder (So) and the sagging part (Sc).
- The closed line describing the shape of the chain covering the cylinder (see Fig. 2) is continuous (Fo = Fc) at the point of separation of the chain from the circle.
- This closed line is smooth (F′o = F′c) at the point of separation of the chain from the circle.

In Fig. 8, it is possible to see the graphical representation of the solution of the chain problem on the cylinder for different ratios of the lengths of the closed chain and of the circumference of the cylinder on which the chain is thrown. The calculations were carried out for R = 1m, but the similarity theory can be applied to this problem, and it can be argued that the specific values of R and L do not affect the shape of the chain sag - only their ratio is important here. This statement has so far been proved by the authors only by a series of numerical experiments and requires theoretical confirmation.

C. Finding a Graphic Dependency

The "chain oval" (as the authors propose to name the family of curves shown in Fig. 1, 2 and 8) also has two diameters - small (horizontal d equal to 2R) and large (vertical D equal to R + h – a - see Fig. 2). Let's connect these three parameters of the chain circle (R, L and D), first graphically, and then analytically. Note that two semi-axes are usually distinguished in an ellipse, and not two diameters (double the value of the semi-axes), but diameters will be used in this paper that will not affect the results.

The Find function in the Solve block of Mathcad is capable of not only returning numerical values (see Fig. 7), but also generating functions. This will help us solve the problem of the chain on the cylinder graphically - see Fig. 9.

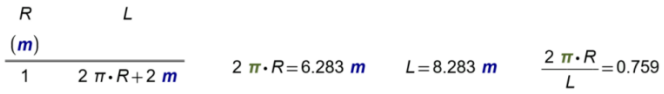


Fig. 6. The Input Data of the Problem of the Chain on a Cylinder.

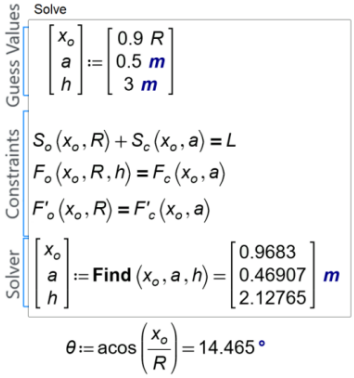


Fig. 7. Solution of the Problem of the Chain on the Cylinder.

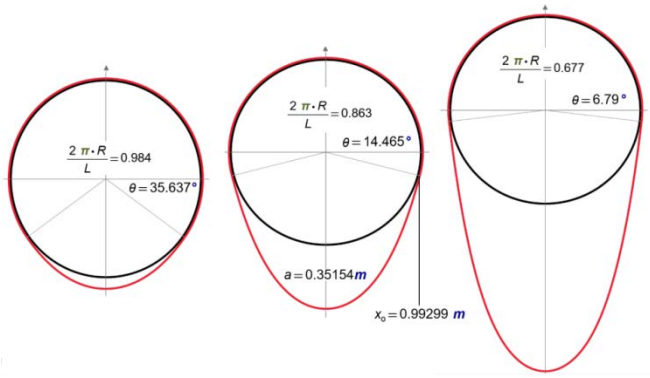


Fig. 8. Graphic Representation of the Solution of the Problem of the Chain on the Cylinder.

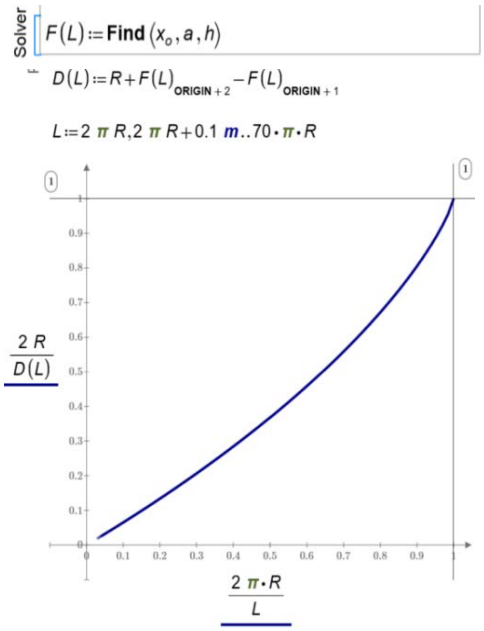


Fig. 9. Graphic Solution to the Problem of a Chain on a Cylinder.

Fig. 9 shows what changes were made to the Solver area of the Solve block in order to create three functions combined into a vector function with name F, whose first element (the element with the ORIGIN number) is the desired function xo(L), the second (ORIGIN + 1) is the desired function a(L), and the third (ORIGIN + 2) is the desired function h(L). The function D(L), composed of the last two with the addition of the radius of the cylinder, returns the desired value of the large diameter of the chain oval. Having such a function at hand, it is not difficult to construct the corresponding dimensionless graph — see Fig. 9. This curve can be considered a key element in the graphical and analytical solution of the problem of the closed chain thrown around the cylinder.

D. Finding Approximation Function

Attempts to analytically solve the chain-to-cylinder problem have been unsuccessful.

Let's solve the chain problem on a cylinder by “Columbus” and find the approximation function.

To do this, let's tabulate the function, whose graph is shown in Fig. 9, using the obtained numerical solution. The obtained data (vectors X and Y) are placed on the website http://zunzun.com, which allows a user to select the least-squares analytical formula that best approximates the numerical dependence. As a result, the formula f(x) = a·x/(b+x)+c·x/(d+x) and the corresponding numerical values of the coefficients a, b, c and d are obtained. For educational demonstration purposes, the coefficients a, b, c and d could be re-calculated using the Mathcad Prime package.

Fig. 10 shows, firstly, the calculation of the coefficients a, b, c and d itself, for which the built-in genfit function is used - general fitting, which requires a first approximation (see the third argument-vector). Secondly, Fig. 10 shows the formation of a function with the name D, which returns the value of the large diameter of the chain oval, depending on the value of its small semiaxis R and the length of the closed chain L. In the same place, it is possible to see the points along which the approximation was carried out, and the smoothing curve itself. The bottom line in Fig. 10 is a calculation of the large diameter of a real chain oval shown in Fig. 1. Its direct measurement (for clarity, graph paper is placed in Fig. 1) gave an answer of 31.4 cm, which is acceptable in accuracy.

The approximating function D with two arguments R and L, shown in Fig. 10, can be considered a “pseudo-analytical” solution to the problem of a chain on a cylinder. This solution can also be considered a kind of “Columbian” - a rough solution.

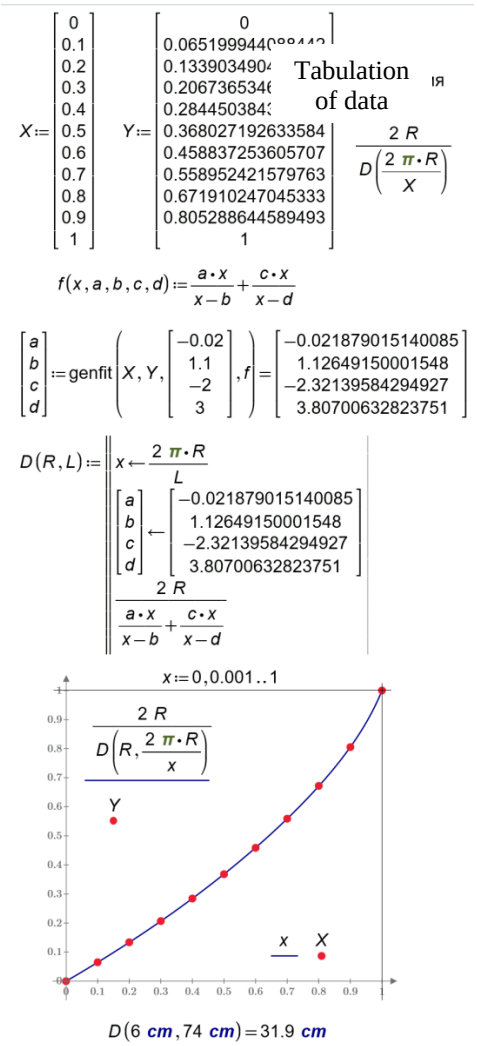


Fig. 10. An Approximating Function for Numerically Solving a Chain Problem on a Cylinder.

III. Chain with a Pendant on a Cylinder

Now let's hang a pendant on the chain and see how it will sag on the cylinder [18]. The problem is solved only numerically. In the new problem, there is not one, but two catenaries shifted to the left and right of the ordinate axis at a distance Δx—see Fig. 11. In addition, the origin is moved to the center of the circle. This caused the Fc function to have one more additional argument h. Without this change, the minimization mechanism applied to this problem [19–21] will not work.

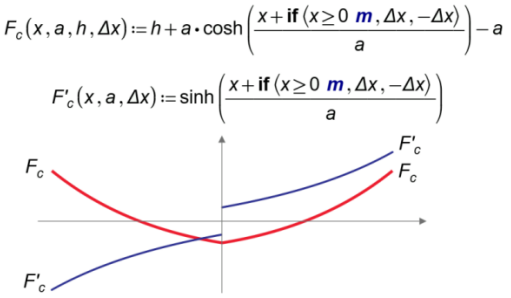


Fig. 11. Catenary Function with a Kink.

To solve this problem, it is needed to add two constants and two functions to the source data:

- a specific (linear) mass of the chain mc;
- a pendant mass m (it is not necessary to enter the value of gravitational acceleration g—this value is built into Mathcad);
- a function Ycg, which returns the ordinate of the center of gravity of the chain with the pendant (the abscissa of this point is zero, because the problem of the chain with the pendant remains symmetric with respect to the ordinate);
- a PE function that returns the potential energy of a chain with a pendant.

Fig. 12 shows these two additional functions.

The Ycg function has two arguments x1 and x2, and not one x (see Fig. 3). Such a change is aimed at developing the problem, at solving other asymmetric problems, which will be discussed below.

In the problem of a chain with a pendant, a fourth unknown quantity Δx appeared (see Fig. 11), and there were three equations. The solution to the problem is to minimize the potential energy of the chain with the pendant, where these three equations derived in the previous section (see Fig. 7) act as constraints [19-21].

This can be done by replacing the Find function in the Solver area of the Solve block with the Minimize function - see Fig. 13.

Fig. 14 shows a graphical representation of the solution of the chain with a pendant thrown over a cylinder for different values of the pendant mass: 1 kg, 20 grams and 0 (chain without a pendant). In Fig. 14, you can see the dashed chain—a continuation of the real chain without a kink. In the left figure, the catenary is almost a straight line: a heavy load pulls the chain into a string. In the middle figure, the minima of the chain lines are visible at x = –0.632 m and x = 0.632 m. The right figure is a repetition of Fig. 6. But the two chain lines do not merge into one due to the limited accuracy of the numerical method for solving the problem. The point on the ordinate axis under the circle is the center of gravity of the hanging part of the chain.

Fig. 14 can also be interpreted as follows: assume some ballast is dropped from a balloon, and then the basket is completely detached from it. This is not a chain line, but a catenoid—a surface formed by the rotation of a catenary. Another analogy is that of a soap film that hangs from a ring, and water has accumulated in its lower part.

The problem of the chain without a pendant is simple because it does not take into account the friction force between the chain and the cylinder. However, the assignments suggested below include cases in which this force must be taken into account and others where it is possible to neglect it.

A cylinder with a chain thrown around it with a pendant begins to rotate around its horizontal axis. Determine the angle of rotation of the cylinder at which the chain begins to slip from it. Additionally, a certain coefficient of friction is set.

The cylinder deviates from the horizontal position. Determine the angle of the cylinder at which the chain begins to slide off it.

$$Y_{cg}(x_1,x_2,a,h,\Delta x) := \frac{\int_{x_1}^{x_2} F_c(x,a,h,\Delta x)\cdot\sqrt{1+F_c'(x,a,\Delta x)^2}\,dx}{S_c(x_1,x_2,a,\Delta x)}$$

$$PE(x,a,h,\Delta x) := g\cdot S_c(-x,x,a,\Delta x)\cdot m_c\cdot Y_{cg}(-x,x,a,h,\Delta x) + g\cdot m\cdot F_c(0\ m,a,h,\Delta x)$$

Fig. 12. Functions of the Potential Energy of the Chain with a Pendant.

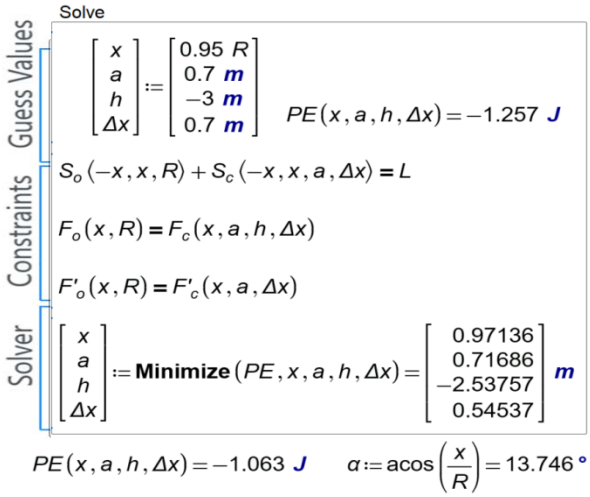


Fig. 13. The Solution to the Problem of the Chain with a Pendant on the Cylinder.

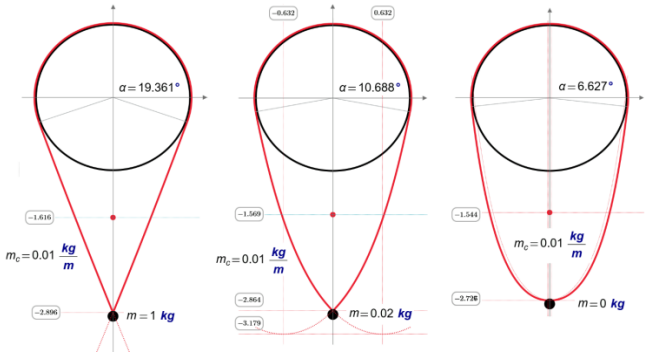


Fig. 14. Graphic Representation of the Solution of the Problem of the Chain with a Pendant on the Cylinder.

The cylinder deviates from the horizontal position, but the chain is fixed on it. How will the contours of this already asymmetric design change?

These problems could form a basis for the further research.

IV. Conclusions

The problem of the parameters of a closed chain (with and without load) thrown over a horizontal cylinder, showing the application of the STEM technology, is considered. Examining together concepts from mathematics, physics and computer science, a numerical solution to the problem is found, a graphical interpretation is obtained, as well as an approximating dependence. The solution of a system of transcendental equations, numerical optimization with constraints and approximation of the numerical solution using the mathematical package Mathcad Prime and online platforms are demonstrated. Additional possible problems for using STEM technology are proposed.

Acknowledgment

This research was funded by NRU "MPEI" and the Ministry of Science and Higher Education of the Russian Federation (unique identifier RFMEFI60719X0323).

References

[1] E.V. Bodrova, and N.B. Golovanova, “Modernization of the higher technical school: historical experience and prospects,” Russian Technological Journal, vol. 5, no. 6, p. 73-97, 2017. DOI: 10.32362/2500-316X-2017-5-6-73-97.

[2] A. Pears, E. Barendsen, V. Dagienė, V. Dolgopolovas, E. Jasutė. “Holistic STEAM Education Through Computational Thinking: A Perspective on Training Future Teachers. In: S. Pozdniakov, V. Dagienė (eds) Informatics in Schools. New Ideas in School Informatics,” ISSEP 2019, Lecture Notes in Computer Science, vol. 11913, Springer, Cham, 2019. DOI:10.1007/978-3-030-33759-9_4.

[3] M. Khine, S. Areepattamannil, “STEAM education: Theory and practice,” Springer, 2019. DOI:10.1007/978-3-030-04003-1.

[4] C. Conradty, S.A. Sotiriou, F.X. Bogner, “How Creativity in STEAM Modules Intervenes with Self-Efficacy and Motivation,” Educ Sci, 2020, 10, 70.

[5] M.-L. How, W.L.D. Hung, “Educing AI-Thinking in Science, Technology, Engineering, Arts, and Mathematics (STEAM) Education,” Educ Sci, 2019, 9, 184.

[6] Y. Li, A.H. Schoenfeld, A.A. di Sessa et al., “On Thinking and STEM Education,” Journal for STEM Educ Res, vol. 2, 2019. DOI: 10.1007/s41979-019-00014-x.

[7] S.E. Shadle, A. Marker and B. Earl, “Faculty drivers and barriers: laying the groundwork for undergraduate STEM education reform in academic departments,” IJ STEM Ed, vol. 4, no. 8, 2017. DOI: 10.1186/s40594-017-0062-7.

[8] B. Yıldırım, and S. Sidekli, “STEM applications in mathematics education: The effect of STEM applications on different dependent variables,” Journal of Baltic Science Education, vol. 17, 2018.

[9] J. Hallström, and K.J. Schönborn, “Models and modelling for authentic STEM education: reinforcing the argument,’ IJ STEM Ed, vol. 6, no. 22, 2019. DOI:10.1186/s40594-019-0178-z.

[10] D.J. Shernoff; S. Sinha, D.M. Bressler et al., “Assessing teacher education and professional development needs for the implementation of integrated approaches to STEM education,” IJ STEM Ed, vol. 4, no. 13, 2017. DOI: 10.1186/s40594-017-0068-1.

[11] V. Ochkov, I. Vasileva, M. Nori, K. Orlov, and E. Nikulchev, “Symbolic Computation to Solving an Irrational Equation on Based Symmetric Polynomials Method,” Computation, vol. 8, no. 2, p. 40, 2020. DOI: 10.3390/computation8020040.

[12] V. Ochkov, “2^5 Problems for STEM Education,” Chapman and Hall/CRC, 2020.

[13] L. Bin, L. Yinghui, and Y. Xuegang, “Dynamic modeling and simulation of flexible cable with large sag,” Appl Math Mech, vol. 21, p. 707, 2000. DOI: 10.1007/BF02460190.

[14] S. Nedev, “The catenary - an ancient problem on a computer screen,” Eur. J. Phys., vol. 21, pp. 451-457, 2000.

[15] B. Maxfield, “Essential PTC Mathcad Prime 3.0. A Guide for New and Current Users,” Academic Press: Cambridge, 2013. DOI: 10.1016/C2012-0-06456-3.

[16] A. Bedford, and W. Fowler, “Engineering Mechanics: Statics,” Addison Wesley, Massachusetts, 1996.

[17] V. Blasjo, “Transcendental Curves in the Leibnizian Calculus. Studies in the History of Mathematical Enquiry,” Academic Press, 2017.

[18] N. Kapustin, and A. Polosin, “On a mixed problem for oscillation of a heavy chain with loads,” AIP Conference Proceedings, 2015. DOI: 10.1063/1.4936721.

[19] R. Almeida, D. Tavares, and D.F.M. Torres. “The Calculus of Variations. In: The Variable-Order Fractional Calculus of Variations. SpringerBriefs in Applied Sciences and Technology,” Springer, Cham, 2019.

[20] Y. Toklu, G. Bekdas, and R. Temur, “Analysis of cable structures through energy minimization. Structural Engineering & Mechanics 2017, 62, pp. 749-758. DOI:10.12989/sem.2017.62.6.749.

[21] W.-J. Chen, G.-S. Du, and X.-Q. Ren, “Minimization iteration procedure of potential energy and structural analysis for the tension cable-strut structure,” Chinese Journal of Computational Mechanics, vol. 27, pp. 1001-1005, 2010.

Recursive Least Square: RLS Method-Based Time Series Data Prediction for Many Missing Data

Kohei Arai[1]
Graduate School of Science and Engineering
Saga University, Saga City, Japan

Kaname Seto[2]
Former Student
Saga University, Saga City, Japan

***Abstract*—Prediction methods for time series data with many missing data based on Recursive Least Square (RLS) method are proposed. There are two parameter tuning algorithms, time update and measurement update algorithms for parameter estimation of Kalman filter. Two learning methods for parameter estimation of Kalman filter are proposed based on RLS method. One is the method without measurement update algorithm (RLS-1). The other one is the method without both time and measurement update algorithms (RLS-2). The methods are applied to the time series data of Defense Meteorological Satellite Program (DMSP) / Special Sensor Microwave/Imager (SSM/I) data with a plenty of missing data. It is found that the proposed RLS-2 method shows smooth and fast convergence in learning process in comparison to the RLS-1.**

Keywords*—Special Sensor Microwave/Imager (SSM/I); Defense Meteorological Satellite Program (DMSP); Kalman filter; Recursive Least Square (RLS) method; missing data; parameter estimation

I. Introduction

In general, earth observation satellites observe arbitrary points on the earth at unequal time intervals based on their orbital conditions. When this observation data is regarded as time-series data at equal time intervals, it can be regarded as time-series data including many unobserved and missing data.

One of the purposes of the time series analysis is to improve prediction accuracy of future data with the past data for the time series of data with a plenty of missing data. There is the famous method, so called, Kalman filter for future data prediction with the previously observed time series of data. There are the parameters for Kalman filter. It, however, is difficult to estimate the parameters.

Kalman filter is composed of an algorithm that updates the state with time (time update algorithm) and an algorithm that updates the observation process (observation update algorithm). Here, as a dynamic characteristic extraction method, we consider both the time series state and the observation process, and examine a method based on time series analysis using the Kalman filter [1], which is widely used because of its relatively high estimation accuracy.

As the parameter estimation method for adaptive filtering, the sequential least squares method (RLS method) [2], [3], which performs sequential learning on the assumption that the target time series is linearly stationary, is generally used.

Matsuoka and Tateishi reported on reflectance correction for remote sensing data including missing data actually observed using a time series model (BRDF model; Bidirectional Reflectance Distribution Function model) [a priori knowledge] [4]. However, when the a priori knowledge cannot be introduced when extracting the dynamic characteristics of the target time series, or when the a priori knowledge is used and the residual time series is used to improve the accuracy, the target is used. It is also conceivable to try to extract the dynamic characteristics. In such a case, a method for extracting the dynamic characteristics from only the time series data is required. In addition, there is no qualitative study on the method for extracting the dynamic characteristics from only the remote sensing data including the observed missing data.

The method proposed in this paper aims to estimate the data at an arbitrary time only from the time series data including such a large amount of missing data. In order to make this purpose possible, some method for extracting the dynamic characteristics of the target time series is required.

The following section describes research background followed by related research works. Then the proposed method is described followed by experiment. After that, conclusion is described together with some discussions.

II. Research Background

As an example, Fig. 1 shows the SSM / 1 (Special Sensor of Microwave / Imager) microwave radiometer mounted on the DMSP (Defense Meteorological Satellite Program) satellite [quasi-return orbit] on June 2, 1998 (Japan time).

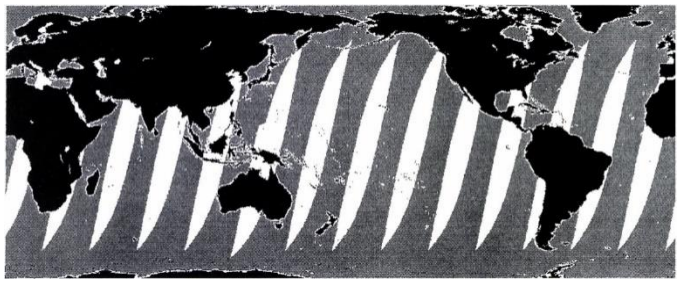


Fig. 1. An Example of the Observed Area with SSM/l Acquired on June 2 1998. The Land Areas are Colored in Black While the Ocean Areas are Colored in Gray and White. The Gray Colored Areas show the Observed Areas While the White Colored Areas show the Non-Observed Areas.

The obtained global observation area is shown. The black area is the land area, and the white / gray area is the sea area.

The gray area is the area observed in the sea area, and the white area is the area not observed in the sea area (missing area; missing data). Although there are missing areas in the land area, this time the land area was classified as the land area. From Fig. 1, it can be seen that there are many missing areas (missing data) in a wide area.

III. Related Research Works

As for the time series analysis, prediction method for time series of imagery data in eigen space is proposed [5]. Meanwhile, Geography Markup Language: GML based representation of time series of assimilation data and its application to animation content creation and representations is proposed [6]. On the other hand, recovering method of missing data based on the proposed modified Kalman filter when time series of mean data is known is proposed [7]. Time series analysis for shortened labor mean interval of dairy cattle with the data of BCS, RFS, Weight, Amount of Milk and Outlook is conducted [8].

Meanwhile, as for the Kalman filter related research, detecting algorithm for rainfall area movement based on Kalman filtering is proposed [9]. Rain flagging with SSM/I based on Kalman filtering with new parameter estimation is proposed [10] together with rain flagging for NSCAT with SSM/I through gap filling based on Kalman filter [11]. Also, rain flagging method with Kalman filtering for ADEOS/NSCAT is proposed [12]. On the other hand, comparative study on image prediction methods between the proposed morphing utilized method and Kalman filtering method is conducted [13].

On the other hand, time series analysis based on Kalman filter with parameter estimation considering non-linearity of system is conducted [14]. Then handling of missing data in parameter estimation of Kalman filter by RLS method is also conducted [15]. Furthermore, recovering method of missing data based on the proposed modified Kalman filter when time series of mean data is known is proposed [16].

Moreover, detecting algorithm for rainfall area movement based on Kalman filtering is proposed [17] together with rain flagging with SSM/I based on Kalman filtering with new parameter estimation [18]. Then rain flagging for NSCAT with SSM/I through gap filling based on Kalman filter is proposed [19] together with rain flagging method with Kalman filtering [20].

IV. Proposed Method

A. Theoretical Background: Kalman Filter

The Kalman filter was proposed by Kalman in 1960. This is an extension of the previous Wiener filter theory so that it can be applied even when time series data including signals and noise are described in a non-stationary process [14], [15]. Below, x represents the mean of x and x^T represents the transpose of x.

In the Kalman filter, the process of time series data is generally expressed by equations (1) and (2).

$$x_{t+1} = F_t x_t + u_t \quad (1)$$

$$y_t = H_t x_t + w_t \quad (2)$$

where, x_t indicates the state at time t, u_t is the system noise, y_t is the observation data at time t, w_t is the observation noise, F_t is the state transition matrix, and H_t is the observation matrix. Equations (1) and (2) are called state equations and observation equations, respectively, and the Kalman filter can be constructed by determining the state transitions of both equations and the elements of the observation matrix. It is possible to estimate missing data.

The learning algorithm for estimating the state x_t of the Kalman filter (Kalman filter learning algorithm) consists of the time update algorithm and the observation update algorithm of Eq. (3), (4).

$$\widehat{x_{t+1|t}} = F_t \widehat{x_{t|t}} \quad (3)$$

$$P_{t+1|t} = F_t P_{t|t} F^T_t + Q_t \quad (4)$$

$$K_t == P_{t|t-1} H^T_t (H_t P_{t|t-1} H^T_t + R_t)^+ \quad (5)$$

$$\widehat{x_{t|t}} = \widehat{x_{t|t-1}} + K_t(y_t - H_t \widehat{x_{t|t-1}}) \quad (6)$$

$$P_{t|t} = P_{t|t-1} - K_t H_t P_{t|t-1} \quad (7)$$

where, x represents the estimated value of x, and $(H_t P_{t|t-1} H^T_t + R_t)^+$ represents the Moore-Penrose generalized inverse matrix of $(H_t P_{t|t-1} H^T_t + R_t)$. t | t-1 means a variable that transitions from t-1 to t. However, the following equations (8) to (14) are assumed.

$$E[u_k] = \overline{u_k} = 0 \quad (8)$$

$$E[(u_k - \overline{u_k})(u_l - \overline{u_l})^T] = \delta_{k,l} Q_k \quad (9)$$

$$E[w_k] = \overline{w_k} = 0 \quad (10)$$

$$E[(w_k - \overline{w_k})(w_l - \overline{w_l})^T] = \delta_{k,l} R_k \quad (11)$$

$$E[x_0] = \overline{x_0} \quad (12)$$

$$E[(x_0 - \overline{x_0})(x_0 - \overline{x_0})^T] = P_0 \quad (13)$$

$$\delta_{k,l} = \begin{cases} 1 & k = l \\ 0 & k \neq l \end{cases} \quad (14)$$

B. Autoregressive Model and Sequential Least Squares Method (RLS Method)

Consider the n-dimensional autoregressive model of Eq. (15).

$$y(t) = \sum_{j=1}^{d} A_j \, y(t-j) + v(t) \;\; (t = 0,1, \dots) \quad (15)$$

When estimating the coefficient matrix A_j (j = 1,2,…, d) in Eq. (15), the Kalman filter can be constructed as follows.

$$z(t+1) = z(t) \quad (16)$$

$$y(t)^T = H_t z(t) + v(t)^T \quad (17)$$

However, the following equations (18) and (19) are used.

$$z(t) = \begin{bmatrix} A_1^T \\ A_2^T \\ \vdots \\ A_{d-1}^T \\ A_d^T \end{bmatrix} \quad (18)$$

$$H_t = [y\ (t\text{-}1)^{T}, y\ (t\text{-}2)^{T}, \ldots, y\ (t\text{-}d)^{T}] \quad (19)$$

Also, the fact that the equations (20), (21), and (22) are used is used.

$$y\ (t)\ T = (y1\ (t), y2\ (t), \ldots, yn\ (t)) \quad (20)$$

$$z\ (t) == [x\ (t, 1), x\ (t, 2), \ldots, x + (t, n)] \quad (21)$$

$$v\ (t)\ T = (v1\ (t), v2\ (t), \ldots, vn\ (t)) \quad (22)$$

Therefore, it can be expressed as Eqs. (23) and (24).

$$x\ (t+1, i) = x\ (t, i) \quad (23)$$

$$y_i\ (t) = H_t x\ (t, i) + v_i\ (t) \quad (24)$$

where, $i = 1,2, \ldots, n$. The RLS method applies the Kalman filter learning algorithm to Eqs. (16) and (17).

C. The Proposed Method

Fig. 2 shows the configuration of the Kalman filter. Adaptive filtering is a method of estimating the parameters used in these algorithms (time update algorithm, observation update algorithm) to realize the Kalman filter.

There is. However, when the RLS method is applied as a dynamic characteristic extraction method for time-series data containing a large amount of missing data, the RLS method performs sequential learning for the target time series, so the missing data in the sequential learning process, the coping method becomes a problem. Here, as a countermeasure for missing data in the learning process of the RLS method, we propose a method that uses only the time update algorithm without using the observation update algorithm and a method that does not use both the observation update algorithm and the time update algorithm. The former will be called RLS method 1 and the latter will be called RLS method 2. Fig. 3 and 4 show the difference between RLS method 1 and RLS method 2.

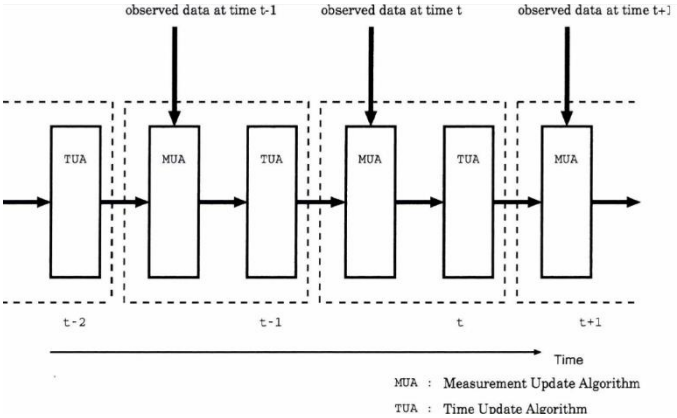


Fig. 2. Kalman Filter Consists of the Measurement Update Algorithm [Equation (5), (6), and (7)] and the Time Update Algorithm [Equation (3), and (4)].

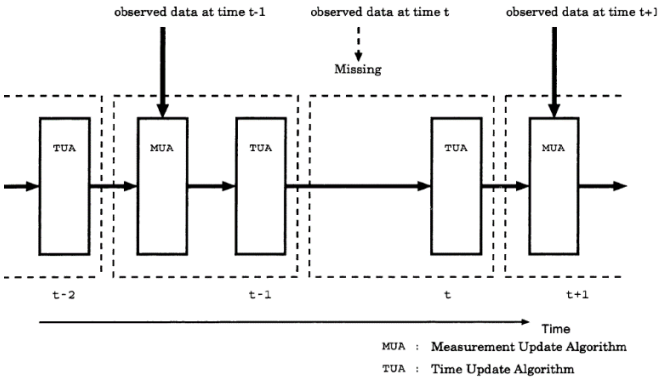


Fig. 3. The RLS Method #1 (MUA is not Applied to the Missing Data).

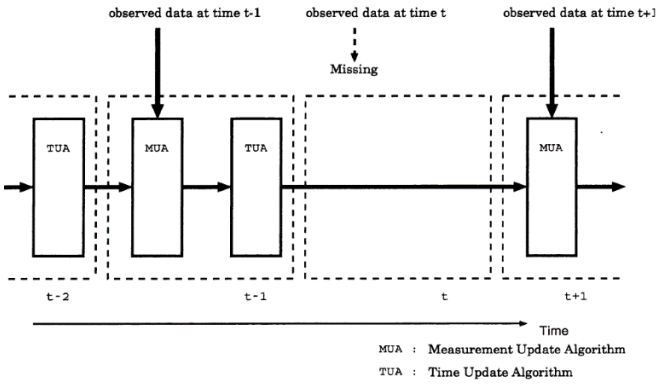


Fig. 4. The RLS Method #2 (Both MUA and TUA are not Applied to the Missing Data).

In this paper, we compare the learning process for past observation data and the prediction behavior for future data in the Kalman filter with parameter estimation based on those proposed methods (RLS method 1, RLS method 2), SSM / 1 observation brightness temperature.

The data (actual observation data) was used. By making these qualitative comparisons, it is possible to investigate the measures and application limits when the RLS method is used as a dynamic characteristic extraction method for time series containing a large amount of missing data. For time series containing many missing data, RLS method 2. The superiority of the proposed method is confirmed.

V. Experiment

A. The Data Used

Here, a simulation experiment is performed by imitating the SSM / I 19.3GHz vertical polarization observation luminance data as an n-dimensional time series. Experiments will be conducted with the time axis in the traveling direction of the DMSP satellite equipped with SSM / 1 and the dimensional axis in the scanning direction. In other words, the experiment is performed using the path-shaped observation luminance data.

SSM / I is a passive microwave radiometer launched by the United States for earth observation, and has 4 frequencies and 7 channels. The breakdown is 19.3GHz vertical polarization, 19.3GHz horizontal polarization, 22.235GHz vertical polarization, 37.0GHz vertical polarization, 37.0GHz

horizontal polarization, 85.5GHz vertical polarization, 85.5GHz horizontal polarization.

B. Creation of Missing Data

For the above n-dimensional time series, generate a uniform random number l (k) with an integer from 0 to q, and use it as missing data for the unit time of the number l (k) that appears (k-1, 2). , 3, ...). In addition, we will introduce an operation to move (shift) in the positive direction in the time axis direction for a unit time by ρ for an n-dimensional time series for time series data generation. In other words, p is the period of continuous observation [the length of discrete time, and q is the period of non-continuous observation (missing)).

Represents the maximum length of [discrete time]. It is considered that it is possible to generate time-series data including arbitrary missing data by introducing these (p,). Examples of time-series data including the missing data generated in this way are shown in Fig. 5 and 6. As shown in Fig. 5, multidimensional time series data including missing data was displayed using the time series of images. Fig. 6 helps to grasp the time axis and dimension axis in this experiment. When p = ∞ and q = 0, the time series does not include missing data (complete observation time series).

C. Evaluation Function

The evaluation function for past data (t <t0) and future data (t> t.) with the current time as t0 is as follows. As for the past data, Eq. (25) and (26) are used as evaluations to examine the learning rate for past observation data.

$$J1(t) = \sqrt{\frac{1}{N_t}\sum_{i=1}^{t} I(y(i))\frac{1}{n}\|y^*(i) - y(i)\|^2} \quad (25)$$

$$I(y(i)) \equiv \begin{cases} 1\ y(i)\ is\ observed\ data \\ 0\ y(i)\ is\ missing\ data \end{cases} \quad (26)$$

However, t is an arbitrary time, and Nt is the number of observed data up to time t. Further, y * (i) is the estimated data (n-dimensional) at time i, and y (i) is the correct answer data (n-dimensional) at time i. Since missing data cannot be evaluated, the value of J1 (t) is negative.

In general, the value of J1 (t) is expected to decrease as the time t increases.

On the other hand, as for the future data, in order to verify the result of learning using the past data, the future data is predicted using the estimated parameters, and the behavior of the prediction accuracy is investigated. Equation (27) is used for evaluation.

$$J2(m) = \sqrt{\frac{1}{n}\|y^*(t_0 + m) - y(t_0 + m)\|^2} \quad (m>0) \quad (27)$$

Further, y^* (i) is the estimated data (n-dimensional) at time i, and y (i) is the correct answer data (n-dimensional) at time i.

D. Experimental Results

In this experiment, we assume an autoregressive model with a degree of 1, and change p and q with respect to the SSM / I19.3GHz vertical polarization observation brightness data with dimension n = 5 and current time t0-100. We compared the learning process for past observation data and the prediction behavior for future data between RLS method 1 and RLS method 2.

By changing p and q from one multidimensional time series, it is possible to generate multiple multidimensional time series including arbitrary missing data. This time, we generated four types of five-dimensional time series: (p, q) = (5,5), (5,10), (10,5), (10,10). The following similar experiments were performed on these four 5-dimensional time series. Based on these experimental results, it is possible to make a qualitative comparison between RLS method 1 and RLS method 2. RLS method 1 and RLS method 2 were used to train until the current time t0, respectively. J1 (t) was used as the evaluation function. Then, in order to verify the learning results of RLS method 1 and RLS method 2, future data was predicted using each parameter estimated at the current time t0, and the behavior of prediction accuracy was investigated. J2 (m) was used as the evaluation function. The prediction was made up to 50 years ahead (m = 50).

This experiment is for qualitative comparative verification of prediction methods (RLS method 1, RLS method 2) based on the RLS method for time series data including missing data. Therefore, it is considered that only 19.3 GHz vertical polarization observation luminance data (1 channel) is enough, and it is sufficient to consider an autoregressive model having an order of 1.

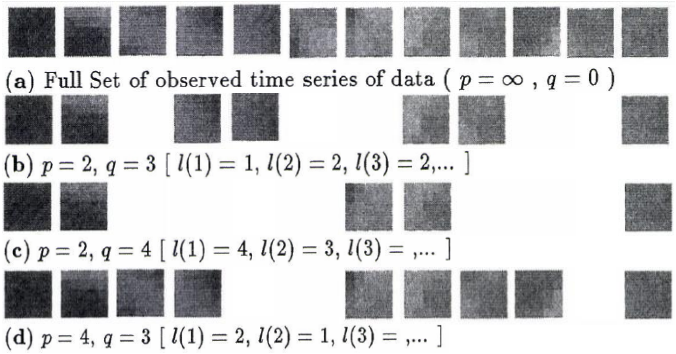


Fig. 5. Examples of the Time Series Including the Missing Data. The Parameter p is the Length of the Period when it is Observed. The Value $l(k)(k$=1, 2, 3, …) is the Length of the Non-Observed Period. ($0<l(k)<q$).

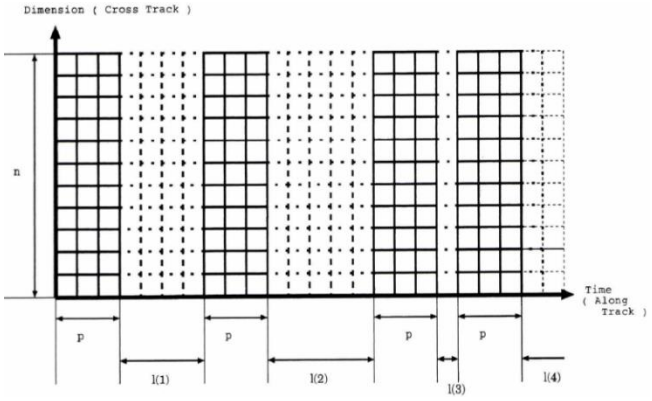
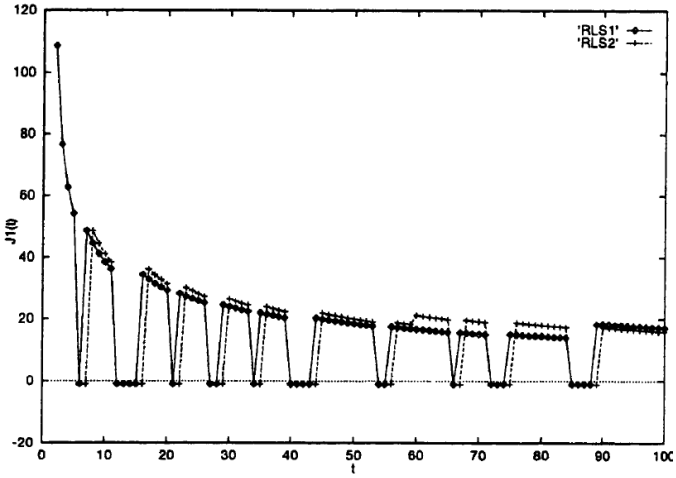
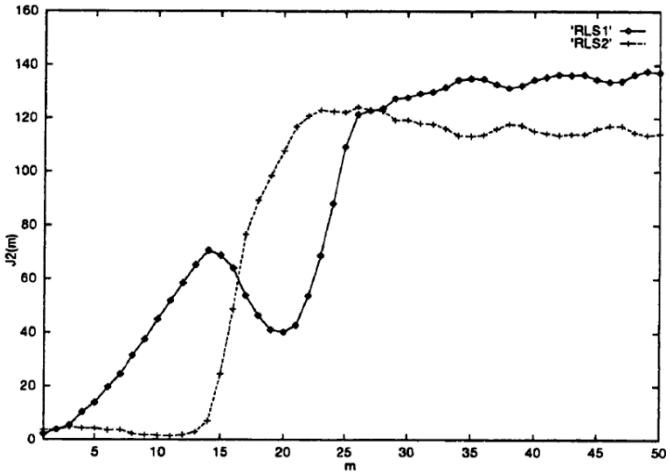


Fig. 6. The Representation of the Observed and the Non-Observed Data. (The Horizontal Axis shows Time Slots While the Vertical Axis shows the Number of Dimensionality which are Corresponding to the Along Track and the Cross Track Directions.) In this Example, Data are Observed for three Time Slots (p=3).

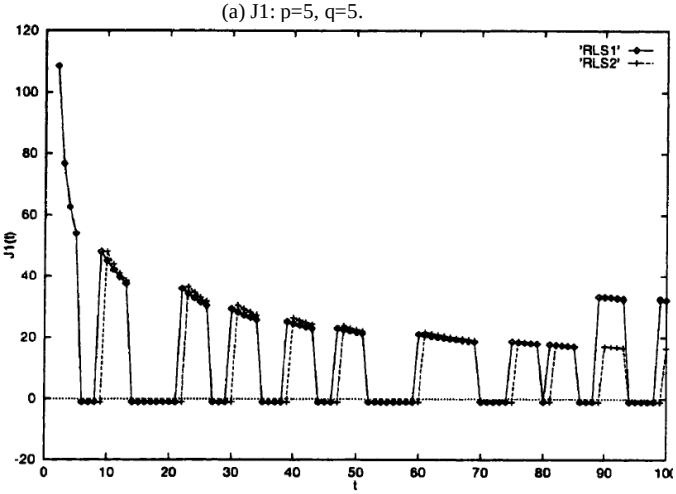
Fig. 7 shows the experimental results when p and q are changed, respectively. The figure on the left shows the transition of J1 (t), and the figure on the right shows the transition of J2 (m). In the figure on the left, the horizontal axis is displayed in the range of [0,100] and the vertical axis is displayed in the range of [-20,120]. In the figure on the right, the horizontal axis is displayed in the range of [1,50] and the vertical axis is displayed in the range of [0,160].



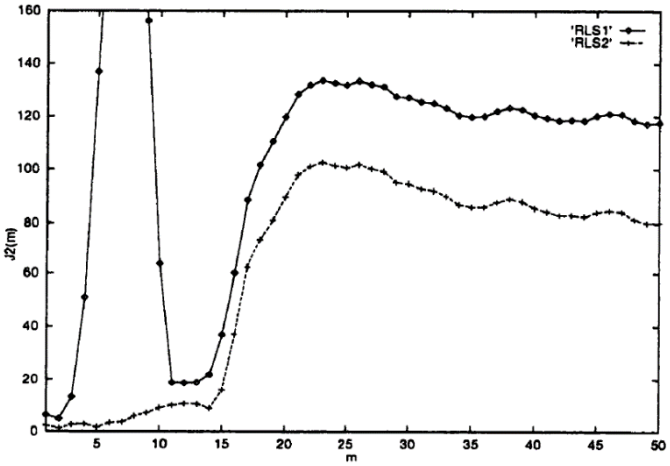
(a) J1: p=5, q=5.



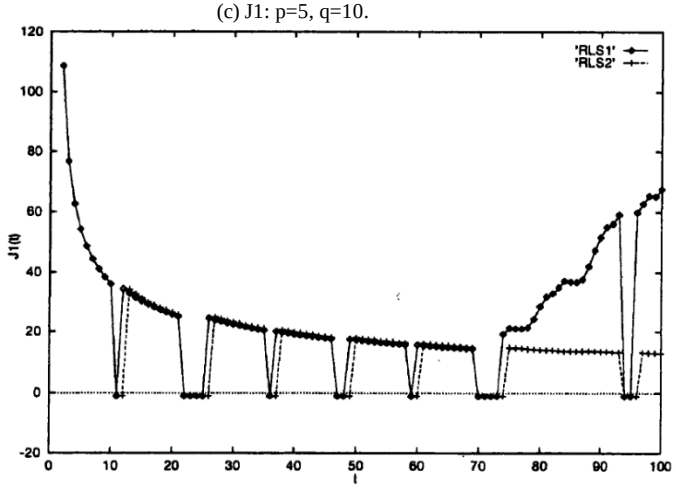
(b) J2: p=5, q=5.



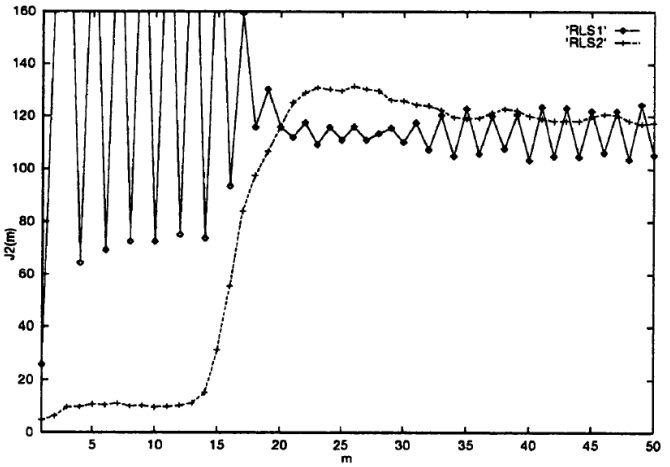
(c) J1: p=5, q=10.



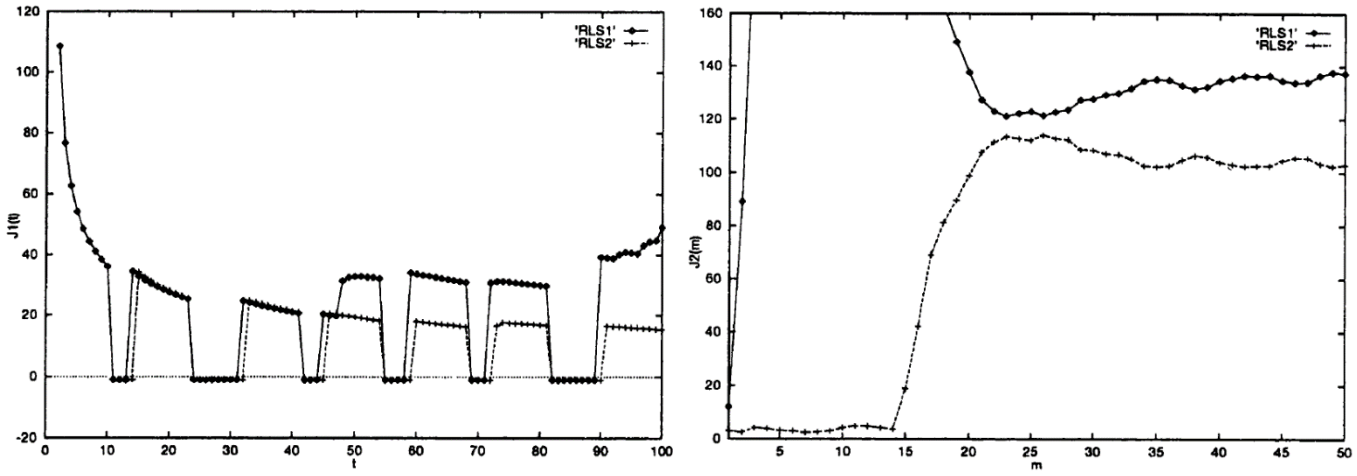
(d) J2: p=5, q=10.



(e) J1: p=10, q=5.



(f) J2: p=10, q=5.



(g) J1: p=10, q=10. (h) J2: p=10, q=10.

Fig. 7. The Transient Responses of J1(t) and J2(m) with the RLS Method #1 and the RLS Method #2 for (p, q)=(5,5), (5, 10), (10, 5), (10,10).

E. Remarks

In the comparison of the evaluation function J1 (t) between RLS method 1 and RLS method 2, the value of J1 (t) in RLS method 2 decreased in all four cases as t increased, whereas RLS decreased. Since the value of J1 (t) in method 1 may increase in some cases, the superiority of RLS method 2 can be confirmed.

Evaluation between RLS method 1 and RLS method 2 In the comparison of J2 (m), RLS method 2 shows almost the same characteristics of J2(m) with respect to rn in four cases, whereas the RLS method In some cases, 1 has drawbacks such as the behavior of J2 (m) with respect to rn causing vibration.

The reason why the behavior of J2 (m) with respect to m in RLS method 1 causes vibration is that the value of the evaluation relation J1 (t) of the learning process for the past observation data increased. For the above reasons, it can be said that RLS method 2 is superior to RLS method 1.

VI. Conclusion

In this paper, we proposed a method based on the Kalman filter as a method for estimating missing data from a multidimensional time series including missing data, or for estimating data at any time. At that time, we proposed two transformation methods of the sequential least squares method (RLS method), which has been widely used in the past as a parameter estimation learning method for the Kalman filter, and compared the two methods (RLS method 1, RLS method 2). The learning process for the past observation data and the prediction behavior for the future data were compared using the observation brightness temperature data by SSM / 1. In order to generate a multidimensional time series including arbitrary missing data and perform qualitative comparison verification between RLS method 1 and RLS method 2, the time axis is set in the direction of travel of the satellite with respect to the path-shaped observation data. The experiment was conducted with the dimension axis in the scanning direction.

In RLS method 1, learning is performed while performing a time update algorithm on past missing data, so it is difficult to perform effective learning when there are many missing data in the past. It is considered that there are cases.

In RLS method 2, learning is performed on past missing data without performing a time update algorithm, so it is possible to prevent a rapid increase in error due to estimation of many past missing data. Conceivable.

It was confirmed that RLS method 2 is suitable as a learning method by RLS method for a time series in which many missing data are continuously present.

VII. Future Research Works

Further research works are required for the other missing data consideration. There are some interpolation and extrapolation methods of the alternative methods for the proposed Kalman filter based method with RLS method.

Acknowledgment

The author would like to thank Prof. Dr. Hiroshi Okumura and Prof. Dr. Osamu Fukuda of Saga University for their valuable comments and suggestions.

References

[1] R.E. Kalman: A New Approach to Linear Filtering and Prediction Problems, Trans. ASME, J. Basic Eng., 82D-1, 35/45, (1960).

[2] Toru Katayama: “Applied Kalman Builta”, Asakura Shoten, (1983).

[3] Eweda Eweda: Comparison of RLS, LMS, and Sign Algorithms for Tracking Randomly Time-Varying Channels, IEEE Trans. Signal Processing, 42-11, 2937/2944, (1994).

[4] Shinyo Matsuoka, Ryutaro Tateishi: “Study on Correction of Reflectance in AVHRR Data Using Time Series Model”, Photogrammetry and Remote Sensing, 37-2, 4/14, (1998).

[5] Kohei Arai Prediction method for time series of imagery data in eigen space, International Journal of Advanced Research in Artificial Intelligence, 2, 1, 12-19, (2013).

[6] Kohei Arai, Geography Markup Language: GML based representation of time series of assimilation data and its application to animation content creation and representations, International Journal of Advanced Research in Artificial Intelligence, 2, 4, 18-22, 2013.

[7] Kohei Arai, Recovering method of missing data based on the proposed modified Kalman filter when time series of mean data is known, International Journal of Advanced Research in Artificial Intelligence, 2, 7, 18-23, 2013.

[8] Kohei Arai, Osamu Fukuda, Hiroshi Okumura, Kenji Endo, Kenichi Yamashita, Time Series Analysis for Shortened Labor Mean Interval od Dairy Cattle with the Data of BCS, RFS, Weight, Amount of Milk and Outlook, IJACSA, 9, 7, 108-115, 2018.

[9] Kohei Arai, Detecting Algorithm for Rainfall Area Movement based on Kalman Filtering, Proceedings of the NSAT/SWT Symposium, Kyoto, Nov. 1995.

[10] Kohei Arai, Rain Flagging with SSM/I Based on Kalman Filtering with New Parameter Estimation, Proceedings of the NSCAT Science Workshop, (1997).

[11] Kohei Arai, Rain Flagging for NSCAT with SSM/I Through Gap Filling Based on Kalman Filter, Proc. of the NASA Scatterometer Science Symposium, pp. 161-163, 1998.

[12] Kohei Arai and Kaname Seto Rain Flagging Method with Kalman Filtering for ADEOS/NSCAT, Proc. of the COSPAR Congress, A0.1-0009, 1998.

[13] Kohei Arai, Comparative Study on Image Prediction Methods between the Proposed Morphing Utilized Method and Kalman Filtering Method, International Journal of Research and Reviews in Computer Science (IJRRCS) Vol. 3, No. 6, 1875-1880, December 2012, ISSN: 2079-2557, 2012.

[14] Kohei Arai, Kaname Seto: "Time series analysis based on Kalman filter with parameter estimation considering non-linearity of system", Proceedings of the 24th Annual Meeting of the Remote Sensing Society of Japan, (1998).

[15] Kohei Arai, Kaname Seto: "Handling of missing data in parameter estimation of Kalman filter by RLS method", Proceedings of the 26th Annual Meeting of the Remote Sensing Society of Japan, (1999).

[16] Kohei Arai, Recovering method of missing data based on the proposed modified Kalman filter when time series of mean data is known, International Journal of Advanced Research in Artificial Intelligence, 2, 7, 18-23, 2013.

[17] Arai,K., Detecting Algorithm for Rainfall Area Movement based on Kalman Filtering, Proceedings of the NSAT/SWT Symposium, Kyoto, Nov. 1995.

[18] Kohei Arai, Rain Flagging with SSM/I Based on Kalman Filtering with New Parameter Estimation, Proceedings of the NSCAT Science Workshop, (1997).

[19] Kohei Arai, Rain Flagging for NSCAT with SSM/I Through Gap Filling Based on Kalman Filter, Proc. of the NASA Scatterometer Science Symposium, pp. 161-163, 1998.

[20] Kohei Arai and Kaname Seto Rain Flagging Method with Kalman Filterinmg for ADEOS/NSCAT, Proc. of the COSPAR Congress, A0.1-0009, 1998.

AUTHOR'S PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 55 books and published 620 journal papers as well as 450 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.html

Mapping Linguistic Variations in Colloquial Arabic through Twitter

A Centroid-based Lexical Clustering Approach

Abdulfattah Omar[1]*
Department of English
College of Sciences and Humanities
Prince Sattam Bin Abdulaziz University
Department of English, Faculty of Arts
Port Said University

Hamza Ethleb[2]
Translation Department
Faculty of Languages
University of Tripoli
Tripoli, Libya

Mohamed Elarabawy Hashem[3]
Department of English
College of Science and Arts in Tabarjal,
Jouf University, KSA
Faculty of Languages and Translation
Cairo, Al-Azhar University, Egypt

***Abstract*—The recent years have witnessed the development of different computational approaches to the study of linguistic variations and regional dialectology in different languages including English, German, Spanish and Chinese. These approaches have proved effective in dealing with large corpora and making reliable generalizations about the data. In Arabic, however, much of the work on regional dialectology is so far based on traditional methods; therefore, it is difficult to provide a comprehensive mapping of the dialectal variations of all the colloquial dialects of Arabic. As thus, this study is concerned with proposing a computational statistical model for mapping the linguistic variation and regional dialectology in Colloquial Arabic through Twitter based on the lexical choices of speakers. The aim is to explore the lexical patterns for generating regional dialect maps as derived from Twitter users. The study is based on a corpus of 1597348 geolocated Twitter posts. Using principal component analysis (PCA), data were classified into distinct classes and the lexical features of each class were identified. Results indicate that lexical choices of Twitter users can be usefully used for mapping the regional dialect variation in Colloquial Arabic.**

Keywords*—Colloquial Arabic; computational statistical model; lexical patterns; linguistic mapping; principal component analysis (PCA)

I. Introduction

Sociolinguists have studied lexical variation and correlated the process through which speaker groups choose their vocabulary with a bundle of variables, such as gender, context, social status, topic [1-4]. More recently, researchers have focused on dialect geography in social media, due to the advances in technology and the unprecedented development of communication channels and networks [5-7]. It is true that these communication channels and networks provide good opportunities for researchers and sociolinguists to study and explore linguistic variation among different speaker groups. Interestingly, the study of linguistic variation through social media networks has been parallel to computational methods. These methods have the potential of dealing with big data and investigating linguistic variation on a larger scale which have positive implications to the generalizability and reliability issues [8-10]. In Arabic, however, much of the work on regional dialectology is so far based on traditional methods; therefore, it is difficult to provide a comprehensive mapping of the dialect variation of all the colloquial dialects of Arabic. As thus, this study is concerned with proposing a computational model for mapping the linguistic variation and regional dialectology in Colloquial Arabic through Twitter based on the lexical choices of speakers. The purpose of the study is to explore the lexical patterns for generating regional dialect maps as derived from Twitter users. In order to map the linguistic variation of Colloquial Arabic dialects, cluster analysis methods were used. This is a clustering method where each class or group has distinct features that make it different from other groups. In dialectology, speakers who share the same linguistic features should be grouped together. This should serve as a basis for exploring the distinctive features of each speaker group. The remainder of the article is organized as follows. Section 2 is a brief survey of the literature on linguistic mapping through social media networks. Section 3 describes the methods and procedures. Section 4 presents the lexical features of dialectal variations among Arab speakers. Section 5 is conclusion.

II. Literature Review

Many advances have been made in the recent years in representing the world's linguistic diversity or what is referred to as language mapping, also referred to as linguistic cartography [11-13]. This is defined as "the visualization of linguistic and language-related data in geographic space and, hence, the representation of correlations between geographic and linguistic facts" [14]. Numerous research projects have been concerned with the regional classification of languages based on several parameters including phonetic and lexical variables. The premise is that there is a significant correlation between geographic location and the linguistic facts. It is argued that linguistic maps can be drawn or generated based on the lexical and phonetic variables as they still carry unique features that can distinguish speakers of the same language. Although the classification of languages and dialects is an established tradition in linguistic studies, the widespread of social media networks and platforms as well as electronic/digital databases has provided researchers with rich and untraditional resources to data. In fact, the social networks

*Corresponding Author
Paper Submission Date: October 26, 2020
Acceptance Notification Date: November 11, 2020

and platforms have become an integral part in people's daily lives and created virtual speech communities which should not be ignored in sociolinguistic studies.

The unprecedented development of social media networks which have been parallel to the development of computational and statistical methodological frameworks have made it possible for researchers to investigate the issue of linguistic mapping on a larger scale. Today, computational approaches provide researchers with the potentials of processing big data in a fast and efficient way. In this regard, numerous studies have been developed using the potentials of computational systems in dealing with big data [15-17]. Studies of the kind are generally based on large corpora for investigating the correlation between lexical patterns and regional dialects [18]. The premise is that correlation between linguistics on the one hand and geography and population on the other hand can be best investigated through statistical and computational methods for their effectiveness in dealing with big corpora that are thought to have good implications to data representativeness and generalizability of the results. Another advantage of the use of computational models and systems in linguistic mapping is that they provide researchers with clear visualizations of the linguistic maps. Today, three-dimensional representation systems are used for the visualization of the geography of linguistic features [10, 19].

Moisl [20] argues that the integration of computational methods into linguistic mapping has significantly contributed to the literature. Traditionally, linguistic mapping of dialect variation was based on single linguistic (mainly phonetic and lexical) features. These maps were also normally limited to small bundles of dialects within the same language. Due to the capabilities of computational systems, maps of regional dialects can be based on multiple linguistic features. They can also be based on many dialects within the same language. An obvious example is the Atlas of German Dialects [21-23]. The Atlas provides a detailed classification of German dialects even beyond the political borders of Germany. The project records and documents the remaining German dialects which were spoken in Northern Moravia. This Atlas is different from traditional linguistic mapping projects of the German dialects which were based on partial explorations and lacked holistic view. The newly developed Atlas is determined and confirmed by multiple linguistic features and unified methods [24]. Computational approaches have also been used in mapping the dialect variation of different languages including English, German, and Chinese [14, 25-28].

Despite the recent advances in the classification of regional dialects using computational and statistical approaches, the studies of Arabic dialects have not been fully explored yet. Much of the work on the classification of the regional dialects of Arabic has been mainly based on comparing a small number of dialects using a limited set of linguistic variables. Although Mulki, et al. [29] suggested the use of recent clustering technologies and systems in the classification of social media language in Arabic, so far there is no holistic view of the regional dialects in Colloquial Arabic. This study seeks to address this gap in the literature through proposing a computational model for the classification of the regional dialects in Colloquial Arabic.

III. Methods and Procedures

For the purposes of the study, a corpus of 1597348 geolocated Twitter posts by 650847 users was designed. Selected tweets are limited to those written in Arabic. However, posts written in Arabizi or Phranco-Arabic are included. The rationale is that such alphabets are very popular today especially on social media networks and therefore should not be disregarded. Data were collected during December 2019 on the most important trends in the Arab world, according to the BBC News Arabic survey that included representatives from almost all the Arab countries. These topics included atheism, women's rights, refugees, honor killing, LGTB, and the Arab-Israeli conflict. Hashtags on these topics were selected and data were derived.

As an initial task, the tweets/posts were converted into what is known as bag of words. Tweets were represented as strings of lexical types. This is because the study is concerned with the lexical properties only. It asks whether lexical choices can map the linguistic variation of Colloquial Arabic dialects. This had the effect of having a corpus of 12778784 words. These were mathematically represented in a vector space matrix, henceforth referred to as colloquial_arabic_dial_ corpus. The matrix is built out of rows and vectors. The rows represent the number of speakers (650847 Twitter users) and the vectors represent all the lexical types included in this study (12778784 words).

One main problem with this corpus is that it is too big for any clustering system to handle. This problem is referred to as high-dimensionality of data. In cases of such kind, it is very challenging to identify the most distinctive lexical features within the corpus. In order to address the problem, (Term frequency-versus-document frequency) analysis TF-IDF was used. In term weighting applications, it is normally assumed that variables with the highest TF-IDF values are to be the most important. Given that hypothesis, variables 1-250 (representing the highest TF-IDF values) were only retained. This had the effect of reducing the matrix into just 250 lexical variables.

For validity purposes, principal component analysis (PCA) was used. PCA is one of the most reliable data reduction methods. It was revealed that the highest 217 lexical variables were the most important. In order to make sure that the corpus now includes only the most distinctive lexical variables, only the recurring or repeated variables in both TF-IDF and PCA tests were finally selected. This had the effect of reducing the matrix to only 113 lexical variables or words as shown in Table I.

Using Cluster Analysis, the 650847 speakers were classified into four main clusters, as shown in Fig. 1.

Referring to the personal information of users, it was found out that the clustering was not based on any geographic or regional grounds. The most distinctive lexical features of each cluster were thus investigated. It was clear then that clustering was based on thematic grounds. Accordingly, thematic words as well as proper names including murder, honor, killing, Israel, Palestine, and Trump were all deleted. This had the effect of reducing the corpus into just 68

variables. The assumption now is that any grouping of the tweets and users will not be based on thematic grounds.

Once again, cluster analysis was carried out for the Matrix colloquial_arabic_dial_corpus (650847, 68) where the former represents the number of users and the latter the number of lexical variables. Results are shown in Fig. 2.

TABLE I. EXTRACTED LEXICAL VARIABLES THROUGH THE EXECUTION OF PCA AND TF-IDF

عراقي	سوري	لاجئ	بلدنا	وظايفنا
Iraqi	Syrian	refugee	our country	Our jobs
عار	فلسطين	ترامب	اسرائيل	شرف
shame	Palestine	Trump	Israel	honor
دمار	قتل	مخنث	ميسي	شغل
destruction	murder	gay	Messi	job
مصاري	فلوس	بقد	حق	مطاعم
money	money	Really!		restaurants
بالحق	انتخابات	غزة	نتنياهو	عن جد
Really!	elections	Gaza	Netanyahu	Really!

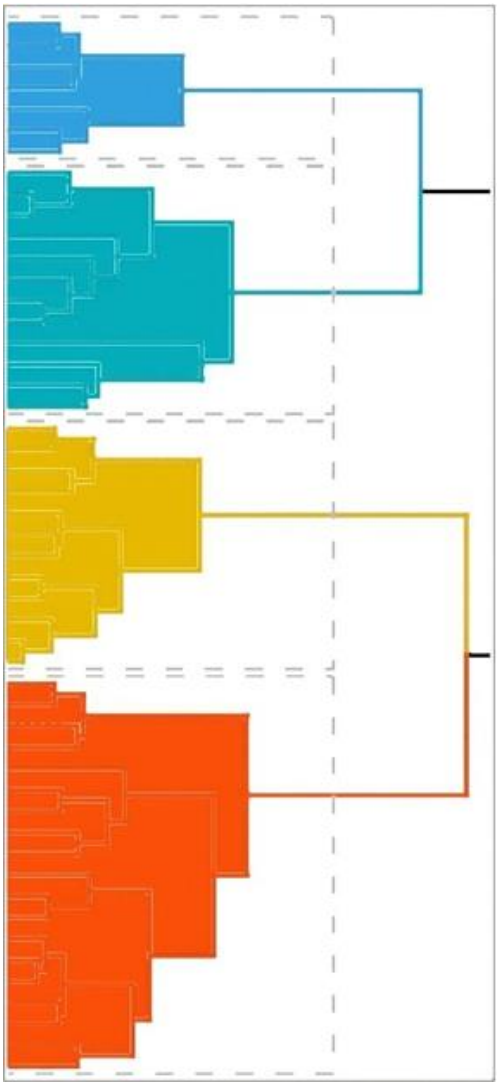


Fig. 1. A Cluster Analysis of the 650847, 113 Matrix.

The Matrix rows were assigned to nine clusters as shown in Fig. 2. Comparing the results of the clustering structures to the personal information available about the users, it was quite obvious that clustering was based on regional basis. These can be referred to as Groups 1-9. Thus, it can be claimed that clustering is based on regional basis.

Interestingly, more than 80% of the retained lexical variables which are considered the most distinctive features are best described as intensifiers and expressions of surprise. This may be due to the fact that such expressions are spontaneous in nature and frequently used in informal Arabic versions. In this regard, they (intensifiers and expressions of surprise) can be good indicators or predictors for mapping the linguistic variation in Colloquial Arabic. This will be the focus of the next section. The distinctive lexical features of each group are discussed.

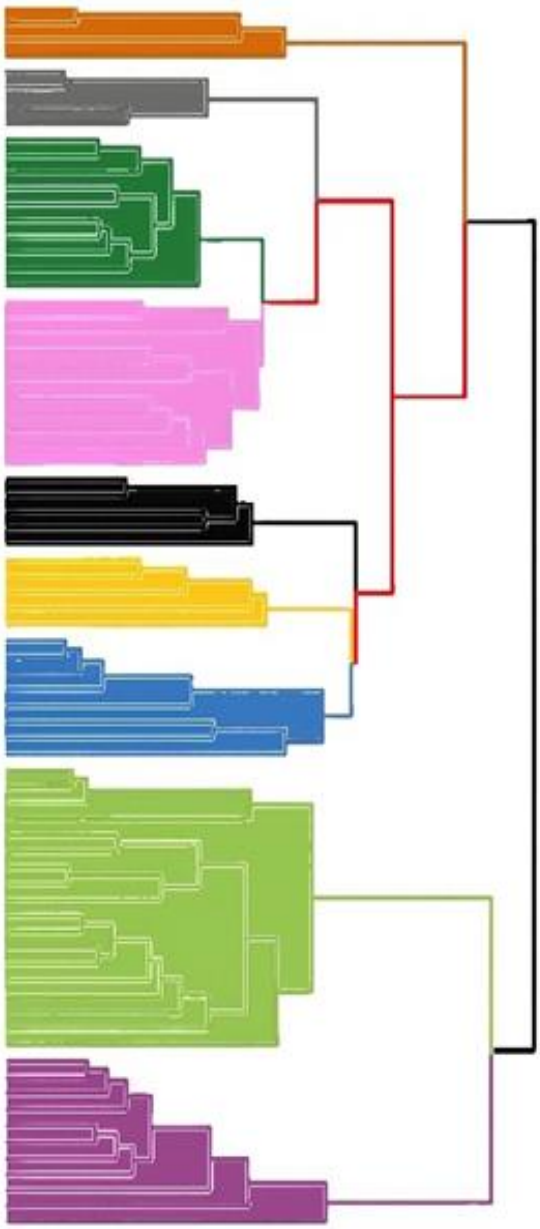


Fig. 2. A Cluster Analysis of the 650847, 68 Matrix.

IV. Analysis and Discussions

The clustering structure shown in Fig. 2 indicates that the datasets fall into nine distinct groups. It serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster. In our case, each group is distinctive from other groups based on its lexical profile. Based on this clustering structure and the centroid analysis of the lexical features of each group or cluster, mapping the regional dialects of Colloquial Arabic can be useful.

It is obvious that intensifiers and expressions of surprise are the most distinctive lexical features of each cluster or speaker group. These expressions are normally known as degree words or intensifiers as they show a degree of intensity by the use of varied word classes [30-33]. According to Peters [34], intensifiers play a significant role in the social interaction and emotional expressions among language users. Based on these lexical features, the main regional dialects of Colloquial Arabic can be mapped through Iraq, the Arab Gulf, Levantine, Egypt, Tunisia, Morocco, Algeria and Libya, as shown in Fig. 3.

Interestingly, the four regional dialects of Algerian, Moroccan, Tunisian, and Libyan Arabic are traditionally classified under just one dialect known as Maghrebi Arabic. In our case, however, there are distinctive linguistic differences among these four dialects, as shown in Table II.

It is noticeable that Libyan and Tunisian dialects in the situation of expressing a surprise or joy are somehow close to each other. They both appear to include the use of 'حق' (hag), 'حقا' (haga) and 'بالحق' (balhag) interchangeably, depending on the mode of the speech. In fact, the word 'بالحق/حق/حقا' (haga/hag/balhag) is probably the most popular intensifying expressions that are employed by speakers of most Arabic dialects. This is due to its close derivational form from the Standard word 'حقيقة' (hagigatn) (truth; truly).

Algeria and Morocco primarily use different lexical choices in terms of showing joyful surprising news. The Algerians say 'منيتنك' (menitik) and the Moroccans use 'واش بصح' (wash'bishah) (what, is it true?) as first choice. These two different dialects sound arguably heavy to the ears of the Gulf and Levantine dialects speakers. Another choice that is popular in the Maghreb dialect is 'قول والله' (qul'walla) (swear by Allah), which received huge employment by most of the dialects speakers of the Arab region, but with huge difference in pronunciation, in terms of stress and pitch.

In Libyan Arabic, for example, a speaker would react to discomfort news as 'متقولهاش' (don't say it) or 'لا ياراجل' (No O man!), i.e. an intensifying phrase that vocally harmonizes with other Maghrebi and Egyptian dialects, and not peculiar to the ears of the Levantine and Mesopotamian dialects. Given the close geographical distance between Tunisia and Algeria, Morocco and Libya, the Tunisian expression of surprise 'يزي عاد' (that can be translated literally to 'stop it' and communicatively to 'really') has more tendency to be comprehended by speakers of those countries and easily identified as in the case of the Tunisian dialect.

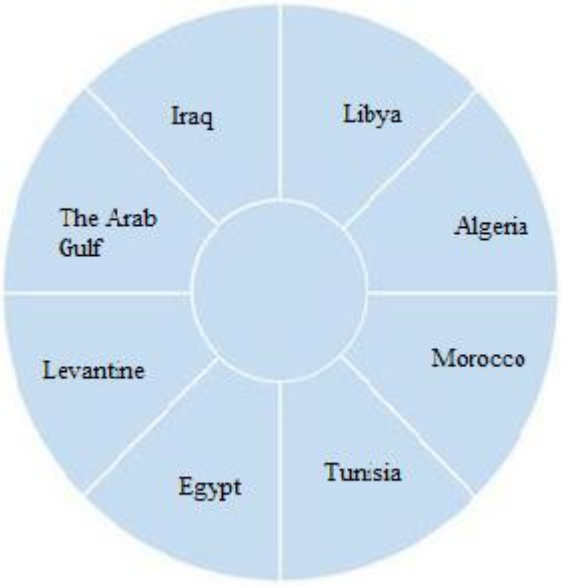


Fig. 3. Geographic Regions of Colloqual Arabic.

TABLE II. Maghreb's Distinctive Features of Intensifiers and Expressions of Surprise

Country	Libya	Algeria	Morocco	Tunisia
Expressions of surprise	Showing comfort			
	حق	منيتك	واش بصح	بالحق
Transliteration	haq	menitik	wash'bishah	balhag
Literal meaning	True	Kidding	What true	True
Communicative meaning	Really	Really	Really	Really
Expressions of surprise	Showing discomfort			
	متقلهاش	قول والله	بصح	يزي عاد
Transliteration	matqulhash	qul'walla	Besah	yazi'aad
Literal meaning	Don't say it	Swear by Allah	Correct	Stop it
Communicative meaning	Really	Really	Really	Really

Table III shows some of the distinctive features of dialectal intensifiers that are used by Maghrebi speakers as a form of intensives or downtoners– showing maximization and minimization. These of course are not the only ones used but as appeared in the methodology, they are more frequently used in the Maghreb region. Bolinger [33] refers to terminologies that are classified as 'adverbs' as amplifiers and downtoners. The latter are adverbs that usually reflect a small amount of quantity. The standard Arabic word for this is 'قليل' (qalil) which seems to have slight dialectal variations among Arabic dialects. However, the varieties of languages used in the Maghrebi dialect seem to influence its peripheries. The intensifier of maximization 'بزاف' is used in Algeria and Morocco. It is understood by the Tunisians and Libyans very well. This rings true to the argument of Harrat *et al.* [35] that dialects are morphologically and syntactically simplified, especially in the regions where one dialect coincides with one another.

TABLE III. MAGHREBI INTENSIFIERS

Country	Libya	Algeria	Morocco	Tunisia
Intensifiers	Showing maximization			
	هلبة	بزاف	بزاف	برشة
Transliteration	Halba	bezaf	bezaf	barsh
Meaning	A lot/too much/too many	A lot/too much/too many	A lot/too much/too many	A lot/too much/too many
Intensifiers	Showing minimization			
	شوي	شوية	شوية	شوية
Transliteration	Shwi	shwia	shwia	shwia
Meaning	A little/a few	A little/a few	A little/a few	A little/a few

This study explores a variety of distinctive lexical variations that are employed by dialect speakers of the Arab world. For instance, the Maghrebi and Egyptian dialect speakers appear to be having heated discussions on Twitter posts that handle political and cultural issues with respect to their individual countries. It is normal to see comments by different Maghrebi dialect users on a geopolitical topic with different lexical variations showing intensifiers and expressions of discomfort. This interaction on Twitter and other social media platforms has undoubtedly expanded the dialectological repertoire of speakers of Arabic dialects across the region.

It can be seen that Table IV shows a variety of lexical choices in expressing the concepts of showing surprise in positive and negative manners. It indicates that speakers' reflection of expression shows their positive attitude in one word at a time of the hearing of a particular piece of information that brings joy to their situation. In such case, the Egyptian speaker will say 'بقد' (beggad) (sure). This form can also be expressed by بجد (bejad) (sure). On the other hand, they would use 'ياخبر' to express discomfort of unwanted news, as shown in Table V.

As it has been previously mentioned, intensifiers are amplifiers and have the function of intensifying or maximizing a certain quantity [36]. The Egyptian dialect is widely spread among other Arabic dialects and this is probably due to its TV industry and famous civilization. The Egyptian people use 'أوي' (*awii*) when they intend to amplify certain situations in conversations. On the contrary, the use of the downtoners 'اليل' (alil) that is derived from 'قليل' (qalil) to express a small amount of quantity or not giving much importance to a quality. This is similar to the distinctive lexical item or 'حتة' (hita). It can be seen that the Egyptian dialect has some vocabularies that are inspired from other Arabic dialects, but with phonological alterations. In fact, there is an evident variation in the vocalization of most of the vocabulary in Egyptian dialect, especial where 'ق' is pronounced as 'أ' (a).

The Levantine group of dialects includes Lebanese, Syrian, Jordanian, and Palestinian dialects. They are quite similar in their ways of expressing intensification. For example, speakers from Lebanon use the expression 'عن جد' (aan'jed) or combined 'عنجد' (aanjed) – as in the case of Syrian and Jordan – to react to happy news. Similarly, in Palestine, they say 'بجد' (bejad), altering the first letter from 'ع' (a) into 'بـ' (b). Those expressions of surprise are close to the Egyptian expression with a slight alteration in the way letter 'ق' is uttered. Although the Levantine expressions of surprise are understood in the Maghreb region, they are rarely or never used by the speakers of the Maghreb region dialects.

The lexical items in Table VI are a result of mapping expressions of a speaker receiving sudden discomfort news of a given social phenomenon. It shows the differences among Arabic dialects, as shown elsewhere in this research. It has to be expressively evident that context is a determining factor in deciding the tone of the speaker; whether their reaction is positive or negative towards a particular intake. For example, the phrase 'قول والله' (swear by Allah), explained above, can be used in both contexts – expressions of happiness and sadness by most of speakers of Arabic dialects. The intensification use is, in fact, not restricted to certain dialects. It is used by dialects of Arabic in colloquial contextualized situations with different intonations and stresses on certain syllables. In this respect, Díaz-Campos and Navarro-Galisteo [37] suggest that the geographical factor in relation to dialects can play a significant role in recognizing lexical variations. This rings true with regard to the Levantine 'عن جد' (sure?) expression in relation to Lebanon, Syria, Jordan, and Palestine, as shown in Table VII.

TABLE IV. DISTINCTIVE FEATURES OF INTENSIFIERS AND EXPRESSIONS OF SURPRISE IN EGYPTIAN ARABIC

Country	Egypt
Expressions of surprise	Showing comfort
	بجد
Transliteration	beggad
Literal meaning	Sure
Communicative meaning	Really
Expressions of surprise	Showing discomfort
	ياخبر
Transliteration	ya'khber
Literal meaning	What a news
Communicative meaning	Really

TABLE V. EGYPTIAN INTENSIFIERS

Country	Egypt
Intensifiers	Showing maximization
	أوي
Transliteration	awii
Meaning	A lot/too much/too many
Intensifiers	Showing minimization
	شوي/اليل/حتى
Transliteration	Shwi/alil/hita
Meaning	A little/a few

TABLE VI. LEVANTINE'S DISTINCTIVE FEATURES OF INTENSIFIERS AND EXPRESSIONS OF SURPRISE

Country	Lebanon	Syria	Jordan	Palestine
Expressions of surprise	Showing comfort			
	عن جد	عنجد	عنجد	بجد
Transliteration	Aan'jed	Aanjed	Aanjed	bejad
Literal meaning	Sure	Sure	Sure	Sure
Communicative meaning	Really	Really	Really	Really
Expressions of surprise	Showing discomfort			
	والله عنجد	شو	والله	ايش
Transliteration	walla'anjed	shoo	walla	Ish
Literal meaning	Swear by Allah it is sure	what	Swear by Allah	what
Communicative meaning	Really	Really	Really	Really

TABLE VII. LEVANTINE INTENSIFIERS

Country	Lebanon	Syria	Jordan	Palestine
Intensifiers	Showing maximization			
	كثير	اكتير/خيرات الله	كوم/كوميات	كتير
Transliteration	kathir	Ekti/khirat allah	kom/komiat	ektir
Meaning	A lot/too much/too many	A lot/too much/too many	A lot/too much/too many	A lot/too much/too many
Intensifiers	Showing minimization			
	شوي	شوي	شوي	شوي
Transliteration	shwi	shwi	shwi	shwi
Meaning	A little/a few	A little/a few	A little/a few	A little/a few

The Levantine intensifiers are almost the same, especially those of downtoners. Data shows that to express a small amount of something, or to explain that certain quality or quantity is low, the minimizing term used is 'شوي' (shwi) or 'شوية' (shwia) depending on the contextual and other singular/plural, masculine/feminine factors. However, the utterance is sometimes colloquially-based and is not subject to syntactic structure. On the other hand, Levantine dialects, as seen in the data, employ the intensifier 'كتير' (ktir) to express maximization of quality or quantity. The morphological process in uttering the term involves stressing and stretching one sound more than the others, such as the case of Syria and Palestine, where the first syllable is stressed. Therefore, difference among Levantine dialects appears only in the surface and can hardly establish a distinction among its speakers. It can be argued that dialects in the Levantine region, and other regions in the Arab world, borrow many words from each other and use them interchangeably. However, the Jordanian intensifier 'كوم' or 'كوميات' (kom/komiat) is properly the most marked and distinctive one appeared in the data.

In the Gulf dialect, and Saudi Arabia in particular, speakers use 'أكيد' (akid). This is another form that is close to Standard Arabic and is usually used in most of the other mapped dialects. However, in the Gulf region, 'أكيد' (akid) appears to eclipse all the other existing intensifiers that express surprise. The data shows that it appears to be widely used in this geographical area. It is, in fact, a standardized form that can be understood by all Arab speakers. It is also used in Iraq in parallel with the intensifier 'صُدك' (sudek), the first syllable receives the primary stress. Expressions of surprise in Arabic dialects vary tremendously, but speakers of neighboring dialects may feel more comfortable where their lexical features coincide with other countries that have boundaries with. On the other hand, the expression 'ابد' (abad) (no/never?) is somehow a distinctive lexical item that expresses discomfort in disturbing situations. It enjoys a high frequency of usage in the Gulf region. It is important to stress here that these are not the only expressions of surprise that Gulf dialect speakers use, but these are the most frequent as per our data, as shown in Table VIII.

The intensifier 'واجد' (wajed) appears to be widely used in the Gulf Arabic, as shown in Table IX. Omar and Alotaibi [38] indicate that the term is used to state that something is provided in plenty in Saudi context. It is worth noting that this utterance has also been mapped in other countries in the regions, such as Libya, Tunisia and Sudan. To a certain degree, the lexical intensifier 'واجد' (wajed) is sometimes substituted by 'كتير' (katir), which is pronounced differently from the Levantine dialect. Contrary to maximization, the minimizer 'شوي' is also frequently used in the Gulf as is the case with almost all Arabic dialects.

According to Ito and Tagliamonte [39], the use of intensifiers is linked to colloquial usage and dialectal varieties. This argument is supported by the findings of this study, where the expressions collated from the data present a degree of difference cross the Arab countries surveyed in this study. The Iraqi dialects exhibit the most distinctive lexical features of the use of intensification, shown in Table X. In fact, some intensifiers still compete in occupying the first place or what really comes subconsciously to mind of speakers of a particular dialect. For instance, in Iraq, the use of intensifiers in expressing surprise in a positive manner is 'صدك' (sudek), but this could not be the case in other sub-regions in Iraq, where 'حقا' (haga) takes over. 'صدك' (sudek), with stressing the first syllable, is derived from the word 'صدق' (truth) and is used here to say that 'is this true?'; in other words, 'are you sure?'. On the contrary, the Iraqis generally react to discomfort news by 'جذب' (jedeb), meaning 'كذب' (lie) – 'you are lying', pronounced in a rising intonation to form a question, as shown in Table XI.

The intensifying utterance in Iraqi dialect is also a distinctive one that subscribes to the Standard Arabic word 'كثير' (kathir), but with a change of the first sound to 'j'. In terms of using minimization, the Iraqi downtoner is not different from the other Arabic dialects surveyed in this study. Quirk, et al. [40] state that downtoners are minimizing items that lessen the degree of intensity of something, they lower the efficacy to a degree of small extent. They always offer a downwards scale to things. Of course, they can be expressed

by using different words classes and structure, but as our data is limited to lexical items that are usually of one morpheme, rather than using 'chucks' or 'fixed phrases', the data show that the most frequent minimizer among Arabic dialects is 'شوي' (shwi), shown in Table XII.

TABLE VIII. GULF'S DISTINCTIVE FEATURES OF INTENSIFIERS AND EXPRESSIONS OF SURPRISE

Country	**The Arab Gulf**
Expressions of surprise	Showing comfort
	أكيد
Transliteration	Akid
Literal meaning	Sure
Communicative meaning	Really
Expressions of surprise	Showing discomfort
	ابد
Transliteration	abad
Literal meaning	No/Never
Communicative meaning	Really

TABLE IX. GULF INTENSIFIERS

Country	**Saudi Arabia**
Intensifiers	Showing maximization
	واجد/كتير
Transliteration	Wajed/katir
Meaning	A lot/too much/too many
Intensifiers	Showing minimization
	شوي
Transliteration	shwi
Meaning	A little/a few

TABLE X. IRAQ'S DISTINCTIVE FEATURES OF INTENSIFIERS AND EXPRESSIONS OF SURPRISE

Country	**Iraq**
Expressions of surprise	Showing comfort
	صدك
Transliteration	sudek
Literal meaning	Truth
Communicative meaning	Really
Expressions of surprise	Showing discomfort
	جذب
Transliteration	jedeb
Literal meaning	You lie
Communicative meaning	Really

TABLE XI. IRAQI INTENSIFIERS

Country	**Iraq**
Intensifiers	Showing maximization
	جثير
Transliteration	jethir
Meaning	A lot/too much/too many
Intensifiers	Showing minimization
	شوية
Transliteration	shwia
Meaning	A little/a few

TABLE XII. OTHER REGIONS' DISTINCTIVE FEATURES OF INTENSIFIERS AND EXPRESSIONS OF SURPRISE

Country	**Sudan**
Expressions of surprise	Showing comfort
	صحي
Transliteration	sahi
Literal meaning	Sure
Communicative meaning	Really
Expressions of surprise	Showing discomfort
	جدي
Transliteration	jedi
Literal meaning	No/Never
Communicative meaning	Really

Watson [41] deduces that the tendencies of using different lexical choices in colloquial interactions, in fact, appear to be a unification factor to the Arabic dialects. She attributes this to the regional tendencies of language usage among Arabs as well as to the "predictable phonological processes" that Standard Arabic goes through in deriving and uttering words [41]. The use of expressions of surprise and intensifiers travels across Arabic dialects and some of them give way to others in different regions. For example, the use of 'قول والله' (qul'walla), discussed above, or its sister variations "أحلف' (ahlef) (swearing as promising), 'والله' (walahi), and 'بالله عليك' (balhi alik) can be used both ways: to express welcomed and unwelcomed effects, depending on the context. It is worth noting that such intensifiers bear religious nuances in their conceptual structure. This expressive feature of showing surprise is more of idiosyncratic nature in colloquial usage of intensifiers.

Although the religious nuances appeared in the data at hand can be found across the Arabic dialects, Sudanese speakers would probably use both صحي (sahi) or جدي (jedi) to express surprise. The latter, which is used to express uninvited news, is closer to the Egyptian 'beggad'. Further to such distinctive lexical items, the Sudanese expression of discomfort 'ما تهظر' (ma'tehadher) (Don't joke) could pose difficulty to the Levantine users to comprehend, especially when uttered in decontextualized conversations. This is not due to the meaning of the expression as the words cause no challenge, but due to

its phonetic features as it is pronounced in a way that make it troublesome to the ears of other Arabic dialect users. It shows a change in the morphological structures of the consonants that are somehow close to the Egyptian 'ما تهزر' (ma tehazar) (Don't joke). Khrisat and Harthy [42] attribute such changes of morphological structure to the 'economy of effort' and 'ease of articulation' that speakers of certain dialects adopt.

TABLE XIII. SUDANESE INTENSIFIERS

Country	**Sudan**
Intensifiers	Showing maximization
	تُف
Transliteration	tuf
Meaning	A lot/too much/too many
Intensifiers	Showing minimization
	شوية/حبة
Transliteration	shwia/haba
Meaning	A little/a few

Unsurprisingly, Twitter provides exposure to all dialects of Arabic languages. It is a place where a variety of linguistic experience is gained by interacting with other dialect users. In fact, many Arabs tend to form friendships with people from other Arab countries who speak a different dialect. Such interaction has developed the linguistic reservoir among them in terms of comprehension. The minimizing downtoner conveys no difference to the above dialects. It is unmarked lexical item that seems to be rigorously used among Arab speakers. However, a more distinctive downtoner item is 'حبة' (haba). The Sudanese dialect lends similarity to the Egyptian dialect in expressing lexical items that feature minimization. According to [35], differences in lexical items among dialects are marked with variations in form. This indicates that even though dialects differ in their morphological structure, they still represent the same meaning. Further, the Sudanese dialect is marked with the distinctive use of 'تُف' (tuf) or 'فُل' (ful) to express intensity, as shown in Table XIII. Arguably, those phrases are challenging to people of other Arabic dialects, especially the Levantine and Iraqi.

V. CONCLUSION

In this paper, we have mapped the most frequent linguistic variations in most popular six dialects in Arabic language. For the purpose of limitation of such tremendous amount of data, this study proposed a computational statistical model for mapping regional dialectology in Colloquial Arabic through Twitter based on the lexical choices of speakers in relation to use of intensifiers and expressions of surprise. Using Twitter corpus of 12778784 words, we had to use a mathematically vector space matrix, henceforth referred to as colloquial_arabic_dial_corpus. The matrix is built out of rows and vectors. With such high-dimensionality of data, we had to adopt a term-frequency-versus-document frequency analysis (TF-IDF). This had the effect of reducing the matrix colloquial_arabic_dial_corpus into just 250 lexical variables. However, and for validity purposes, a reduction method of principal component analysis (PCA) was also employed. This had the effect of reducing the matrix to only 113 lexical variables or words. With this result of corpus reflecting the most frequently used distinctive dialects in colloquial Arabic in relation to intensifiers and expressions of surprise, we began to explore shapes of similarities and differences among Colloquial Arabic dialects.

Hence, twitter corpus was applied to Colloquial Arabic. It showed that Arab people use different lexical items in expressing their surprise. In fact, synonymous occurrences of every expression exist in bounty in every dialect in the study. This would permeate more linguistic variations and colloquial choices. Such variations could be in the morphological or phonological structures of the pattern. Most of the dialects of the Arab region exhibit huge difference in pronunciation in terms of intonation, stress and pitch. Further, the work is consistent with other studies claiming that speakers of neighboring dialects would have more tendencies to understand each other and can easily identify their dialects. Furthermore, it can be claimed that social media platforms such as Twitter is a reservoir of different dialects and a presentation mirror of lexical variations. This has been shown in the discussions of political and cultural issues with respect to their individual countries. Twitter users express geopolitical topics with lexical variations showing a degree of differences of intensifiers and expressions of discomfort. Those utterances sometimes depict a slight alteration in their orthographic structure.

Our findings show that speakers of different Colloquial Arabic dialectal varieties can distinguish speakers of other countries by the vocalization they use. More importantly, the data shows that the minimizer 'شوي' is almost used by all Arabic dialects with intonational differences. Our work could be expanded by examining other linguistic concepts rather than intensifiers and expressions of surprise as we come across the fact that regular contact with other dialects, especially through social media, contributes to the findings of aspects of similarities and differences among Arab speakers. We suggest further research in exploring emerging words and origin of terms that travel across regions in the Arab world. This would give a clearer pattern of their origin and more information on dialect variations in Arabic language.

ACKNOWLEDGMENTS

This publication was supported by the Deanship of Scientific Research at Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia and the Research, Consulting and Training Center at the University of Tripoli, Tripoli, Libya. Authors would like also to thank Dr. Samira Farhat for her helpful comments and insightful suggestions.

REFERENCES

[1] P. Eckert, Meaning and Linguistic Variation: The Third Wave in Sociolinguistics. Cambridge: Cambridge University Press, 2018.

[2] J. K. Chambers and N. Schilling, The Handbook of Language Variation and Change. Wiley, 2018.

[3] D. Brouwer, Gender Variation in Dutch: A Sociolinguistic Study of Amsterdam Speech. Berlin: De Gruyter, 2011.

[4] J. Coates, Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language. Taylor & Francis, 2015.

[5] J. King and S. Sessarego, Language Variation and Contact-Induced Change: Spanish across space and time. John Benjamins Publishing Company, 2018.

[6] A. Georgakopoulou and T. Spilioti, The Routledge Handbook of Language and Digital Communication. Taylor & Francis, 2015.

[7] E. Teich, Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts. Berlin:De Gruyter, 2012.

[8] M. Krug and J. Schlüter, Research Methods in Language Variation and Change. Cambridge: Cambridge University Press, 2013.

[9] G. Parodi, Working with Spanish Corpora. Bloomsbury Publishing, 2007.

[10] W. Maguire and A. McMahon, Analysing Variation in English. Cambridge Cambridge University Press, 2011.

[11] C. Boberg, J. A. Nerbonne, and D. Watt, The Handbook of Dialectology. Blackwell, 2018.

[12] Z. Cao, Linguistic Atlas of Chinese Dialects. Beijing: The Commercial Press, 2008.

[13] H. Goebl, "Dialectometry and quantitative mapping," in Language and Space. An International Handbook of Linguistic Variation, vol. 2, R. K. Lameli, & S. Rabanus Ed. (Language Mapping, Berlin/New York: De Gruyter Mouton, 2010, pp. 433–457.

[14] S. Rabanus, "Language Mapping Worldwide: Methods and Traditions," in Handbook of the Changing World Language Map, K. R. Brunn S., Ed. Cham: Springer, 2020.

[15] P. Rácz, Salience in Sociolinguistics: A Quantitative Approach. Berlin: De Gruyter, 2013.

[16] E. Benmamoun and R. Bassiouney, The Routledge Handbook of Arabic Linguistics. Taylor & Francis, 2017.

[17] J. M. Hernández-Campoy and J. C. Conde-Silvestre, The Handbook of Historical Sociolinguistics. Wiley, 2012.

[18] J. Grieve, A. Nini, and D. Guo, "Mapping Lexical Innovation on American Social Media," Journal of English Linguistics, vol. 46, no. 4, pp. 293 –319, 2018.

[19] H. Moisl and W. Maguire, "Identifying the Main Determinants of Phonetic Variation in the Newcastle Electronic Corpus of Tyneside English," Journal of Quantitative Linguistics, vol. 15, no. 1, pp. 46-69, 2008.

[20] H. Moisl, "Statistical corpus exploitation," in Handbook of Corpus Phonology, J. Durand, Gut, U., Kristofferson, G., Ed. Oxford: Oxford University Press, 2010.

[21] M. Muzikant, "Verkleinerungsformen in den deutschen Dialekten Mährens und Schlesiens - eine historische Reminiszenz (Diminutives in German Dialects in Moravia and Silesia - Historical Reminiscence)," in Sprachen verbinden Beiträge der 24. Linguistik- und Literaturtage, Brno/Tschechien, 2018.

[22] M. Wese and M. Muzikant, Atlas der deutschen Mundarten in Tschechien. Band III Lautlehre 2: Langvokale und Diphthonge (Atlas of german dialects in Czech republic. Volumen III Phonetics 2: Long vowels and diphthongs). Tübingen: Narr/Francke. 2016.

[23] K. Simet and M. muzikant, Atlas der deutschen Mundarten in Tschechien. Band IV Lautlehre 3: Konsonanten (Atlas of german dialects in Czech republic. Volumen IV Phonetics 3: Consonants). Tübingen: Narr/Francke, 2016.

[24] M. Rosenhammer, A. Dicklberger, D. Nuzel, and M. Muzikant, Atlas der deutschen Mundarten in Tschechien. Band II Lautlehre 1: Kurzvokale. (Atlas of german dialects in Czech Republic. Tuebingen: Narr/Francke, 2014.

[25] C. Xu and L. Mao, "The sociolinguistic meanings of syllable contraction in Chinese: A study using perceptual maps," Asia-Pacific Language Variation, vol. 3, no. 2, pp. 160-199, 2017.

[26] G. Roche and H. Suzuki, "Mapping the Minority Languages of the Eastern Tibetosphere," Studies in Asian Geolinguistics, vol. 6, pp. 28-42, 2017.

[27] S. Cullotta and G. Barbera, "Mapping traditional cultural landscapes in the Mediterranean area using a combined multidisciplinary approach: Method and application to Mount Etna (Sicily; Italy)," Landscape and urban planning, vol. 100, no. 1-2, pp. 98-108, 2011.

[28] M. Barni and G. Extra, Mapping linguistic diversity in multicultural contexts. Walter de Gruyter, 2008.

[29] H. Mulki, H. Haddad, M. Gridach, and I. Babaoglu, "Tw-StAR at SemEval-2017 Task 4: Sentiment Classification of Arabic Tweets," in Proceedings of the 11th International Workshop on Semantic Evaluations, Vancouver, Canada, 2017, pp. 664–669: Association for Computational Linguistics.

[30] O. Zaidan and C. Callison-Burch, "Arabic dialect identification," Computational Linguistics, vol. 52, no. 1, pp. 1-36, 2012.

[31] W. Dressler and L. M. Barbaresi, Morphopragmatics: diminutives and intensifiers in Italian, German, and other languages. Berlin, New York: Walter de Gruyter, 1994.

[32] C. Paradis, Degree Modifiers of Adjectives in Spoken British English. Lund: Lund University Press, 1997.

[33] D. Bolinger, Degree Words. The Hague & Paris: Mouton, 1972.

[34] H. Peters, "Degree adverbs in early modem English," in Studies in Early Modern English, D. Kastovsky, Ed. Berlin & New York: Walter de Gruyter, 1994, pp. 269-288.

[35] S. Harrat, K. Meftouh, M. Abbas, W.-K. Hidouci, and K. Smaïli, "An Algerian dialect: Study and Resources," International Journal of Advanced Computer Science and Applications, vol. 7, no. 3, pp. 384-396, 2016.

[36] R. Plo Alastrué and C. Pérez-Llantada, English as a Scientific and Research Language: Debates and Discourses, English in Europe. Berlin, Geramny; Boston, MA, USA: De Gruyter, 2015.

[37] M. Díaz-Campos and I. Navarro-Galisteo, "Perceptual Categorization of Dialect Variation in Spanish," in Selected Proceedings of the 11th Hispanic Linguistics Symposium, J. e. a. Collentine, Ed. Somerville, MA: Cascadilla Proceedings Project, 2009, pp. 179-195.

[38] A. Omar and M. Alotaibi, "Geographic Location and Linguistic Diversity: The Use of Intensifiers in Egyptian and Saudi Arabic," International Journal of English Linguistics, vol. 7, no. 4, pp. 220-229, 2017.

[39] R. Ito and S. Tagliamonte, "Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers," Language in Society, vol. 32, pp. 257-279, 2003.

[40] R. Quirk, S. , G. Greenbaum, G. Leech, and J. Svartvik, A Comprehensive Grammar of the English Language. London: Longman, 1987.

[41] J. Watson, "Arabic dialects " in The Semitic Languages: An internationalHandbook, S. Weninger, Khan, G, Streck, M and Watson, J, Ed. (Handbooks of Linguistics and Communication Science, Berlin Walter de Gruyter, 2011, pp. 851-896.

[42] A. A. Khrisat and Z. A. Harthy, " Arabic dialects and Classical Arabic Language," Advances in Social Sciences Research Journal, vol. 2, no. 3, pp. 254-260, 2015.

Implementation of Text Base Information Retrieval Technique

Syed Ali Jafar Zaidi[1], Safdar Hussain[2]
Department of Computer Science
Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan

Samir Brahim Belhaouari[3]*
Division of Information and Computing Technology
College of Science and Engineering
Hamad Bin Khalifa University, Qatar

***Abstract*—Everyone is in the need of accurate and efficient information retrieval in no time. Search engines are the main source to extract the required information, when a user search a query and wants to generate the results. Different search engines provide different Application Programming Interface (API) and Libraries to the researchers and the programmers to access the data that has been stored in servers of the search engines. When a researcher or programmer search's a query by using API, it returns a Java Script Orientation Notation (JSON) file. In this JSON file, information is encapsulated where scraping techniques are used to filter out the text. The aim of this paper is to propose a different approach to effectively and efficiently filter out the queries based on text which has been searched by the search engines and return the most appropriate results to the users after matching the searched text because the previous techniques which are used are not enough efficient. We use different comparison techniques, i.e. Sequence Matcher Method and then compare the results of this technique with relevance feedback and in the end we found that our proposed technique is providing much better results.**

***Keywords*—Information retrieval; sequence matcher method; relevance feedback**

I. Introduction

Well before the invention of the internet it was so much tough to keep in touch with the world. But, with the dramatic growth of internet (see Fig. 1), it is so much easier for the people to remain in touch with each other's and to spread the information over the world using internet. Approximately, 80% data available on internet is in textual form and is highly unstructured. So, during last two decades, the websites, web blogs and other informative material contain such a massive amount of textual and unstructured data [1]. When anyone uses internet he usually deals with text because text is the main source of information and communication on the internet, majority of internet searches are text-based. This expanding availability of text has demanded lots of research in this area [13].

Internet is the connection of two or more than two computers which can communicate with each other's. Billions of people are connected on internet and are the source of generating text over the internet [1]. As the use of text is increasing day by day and further there is growth in technology is noticed as well, this was not the case in past as in previous decades the only source to use the internet facility was our desktop computers. These desktop computers used internet with wired net but now user can interact with the internet with their laptops, tablets, smart phones and even by using their smart watches with 3G, 4G and 5G technologies [8][13]. With this fast increase in usage of electronic media technology, the speed of internet has also increased such as 3G, 4G and 5G, as these are the technologies which are intended to deliver the required information in the no time [2].

Increased text over the internet is a result of rapid increase in the internet usage (see Fig. 1). So, the chance of irrelevant data extraction fist also increased, and the filtration of irrelevant data is so much necessary in this scenario [3]. Keeping in mind the defined issues, this paper addresses the issue of fetching the most appropriate and relevant text by using the text-based queries in efficient and effective manners. From our expertise in the normal world, we noticed that people are in touch with the text of different types regularly for different reasons [12].

People in this world are in touch with newspapers, televisions, radio, graphical representations and different advisory services to understand, to learn and more specifically to remain in touch with the revolution and change in the world which is occurring in daily life. The main feature for all such activities is to judge how useful and meaningful this text can be about their requirements [11]. By association, we tend to imply that, in such exercises, people don't seem to be recently upstage beneficiaries of messages, however instead dynamic searchers of writings, and therefore the active constructors of importance from these writings. They search around for writings of potential intrigue, so as to support individuals in their information [5]. In this study we have focused on the problem of relevant, real, and precise text retrieval which guarantees correct and precise results to entertain the user needs [22].

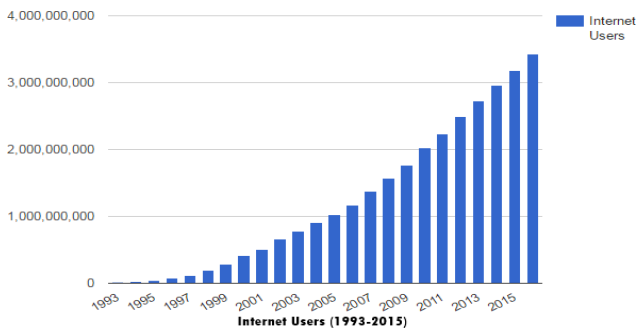


Fig. 1. The Increase of Internet users forms 1993-2015.

*Corresponding Author

II. Material and Methods

A. Approach

In this research, Bing API is used to retrieve the text by applying text-based search query. If we talk about Bing API, it is one of the very useful tools to fetch the information in the form of text or in the form of multimedia information from the server [21]. The results that are returned by this API are in the form of Java Script Orientation Notation file, we then further process the collected results to fetch the text of the information. The gathered information in maximum of the cases were not the wanted ones, thus, we have to create a system that will be helpful in retrieval of the most accurate information from the data.

B. Information Retrieval and Information Filtering

Some Modern information and text retrieval techniques are used to fetch the data in minimum possible time. The basic aim of Information retrieval model is to "discover the relevant knowledge-based information" or a document that fulfill user needs [4]. In modern days the term used is Information Retrieval (IR) rather than Information Gathering. This information can be in any form e.g. Image, Sound, Video, Text or anything which can be used in meaningful purpose [16]. When someone will search anything with the help of Search Engine it will return all the possible results which will match that query [7].

Nowadays, search engines are not the only source to retrieve information (Fig. 2). In parallel, the social media sites are also the big source of the Information Retrieval anyway [10][19]. As compared to past, the Information Retrieval is not only becoming fast but also becoming more accurate nowadays by implementing more reliable and more efficient algorithms [14]. But, in contrast, the chances of Irrelevant Information Retrieval cannot be eliminated [9]. So, to make it more efficient and more reliable, different we have implemented an information retrieval algorithm on actual data which is retrieved.

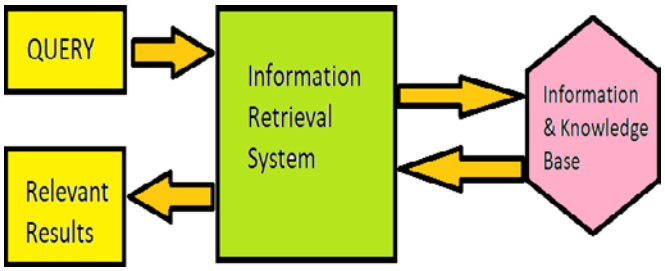


Fig. 2. Information Retrieval System.

III. Implementation

Whenever a user wants to search something on computer or any electronic device, he uses search engine(s) or social media site(s) to write a query and, as a result, receives the relevant results against the query. Relevancy of the query can be judged by different methods which are not sufficient to identify the best relevant text of the information because there are many other words and stop words that can minimize the relevancy of the relevant text. Besides this technique, as it seems not suitable or aligned sufficient, our study aims to find out the relevancy of a text, we use the Sequence Matcher Method which could be more helpful to retrieve appropriate and effective information. But to find out the efficiency of our algorithm we took the relevancy feedback from the user and then compared the precision of relevance feedback with precision Sequence Matcher Method. Where precision refers to a term that how close a result is close to the overall results which has been retrieved.

A. Relevance Feedback

The Relevance feedback refers to the feedback from the user has been taken to identify that how the returned data from the server against the query is relevant for the user [18]. The basic aim of relevance feedback is to check the relevant results of retrieval systems. The basic procedure of relevance feedback is that user will search the query, then the user will tell that which data is relevant and which one is not relevant.

B. Sequence Matcher

Sequence Matcher is basically one of the module used in python programming language, because of this module it is very easy to find out the comparison of the strings, to find out the sequence of the text and then relevancy which is computed with the help of that sequence [19][20]. Sequence Matcher Method uses equation for sequence matching which is given below.

$$D_{ro} = \frac{2*k_m}{|S_1||S_2|},$$

where, k_m denotes the number of same characters in sequence whereas |S1| & |S2| specifies the length of these both strings correspondingly.

The longest substring that is common in both S1 and S2 strings is called anchor. The right and the left part of the string must be examined once again because it will be considered as a new string and this process is repeated again and again until all the characters of S1 and S2 are examined [6][15].

C. Implementation of Sequence Matcher

To implement Sequence Matcher Method, consider the strings Waiting and Main String (Table I).

The length of the string S_2 is 7 whereas the length of string S_1 is 11. Now

$|S_1| = |11|$ *and* $|S_2| = |7|$

In S1and S2 the longest common substring between them is ING therefore ING is an anchor hence now.

$K_m = |ING| = 3$

So, in (Table II) now there is only a substring on the left side of the Km (anchor) of both strings and no substring on the right side of the anchor now the next sequence which is the longest one between them is (AI) which is the second largest sequence. So, AI is new anchor and the Km value will be

$K_m = 3 + |AI|$

$K_m = 3 + 2 = 5$

TABLE I. COMPARISON OF TWO STRINGS

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
S1	M	A	I	N		G	E	T	I	N	G
S2	W	A	I	T	I	N	G				

TABLE II. LARGEST COMMON SEQUENCE IN TWO STRINGS

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
S1	M	A	I	N		G	E	T	I	N	G
S2	W	A	I	T	I	N	G				

TABLE III. ALL COMMON SEQUENCES IN TWO STRINGS

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
S1	M	A	I	N		G	E	T	I	N	G
S2	W	A	I	T	I	N	G				

As we can see now in (Table III) now the AI is in the both strings S1 and S2 so there is no common string on the left side of both strings. And on the right of AI there is another common substring in both strings ING so the value of Km will be 2+3 = 5. Till now we have obtained all the values which are needed to calculate the final score. Now the calculation of string MAIN STRING and WAITING is given below by using equation number.

$$D_{ro} = \frac{2 * 5}{11 + 7}$$

$$D_{ro} = \frac{10}{18} = 0.555$$

IV. RESULTS AND DISCUSSIONS

To check the efficiency and accuracy of the system, we have tested our information filtering system in contrast to the Bing API or Relevance Feedback. As a result, we got different results as compared to Bing API, this has been tested over different queries and then relevance feedback from the user has taken and then compared to Sequence Matcher Method. We have retrieved only the first twenty results against several queries using precision; although Bing API returns almost thirty-five results. Precision is the total number of relevant information over total information, the equation is given below.

$$Precision = \frac{relevant\ data}{retrieved\ data}$$

Precision is the fraction of a query which is searched by the user that is related to the particular query. Its calculation is needed to find out the relevant and non-relevant results in the evaluated documents [17]. Hence the whole results of different tests of different queries has been shown in (Table IV) which is given below.

After applying different experiments, it is clearly seen that the precision value keeps on changing for each and every result, so it shows that our system is effective and efficient. As we can see in the table that the precision of sequence matcher technique is relatively better than the precision of relevancy feedback in maximum cases.

TABLE IV. RESULTS COMPARISONS OF SEQUENCE MATCHER PRECISION WITH RELEVANCE FEEDBACK PRECISION

	Searched Query	**Relevance Feed Back**	**Relevant Sequence Order**	**Precision of Relevance Feedback**	**Precision of Relevance Sequence Order**
1	Jinnah International Airport	13	15	0.87	1.00
2	England Flag	6	12	0.40	0.80
3	Hello	10	10	0.67	0.67
4	Tom Cat Cartoon	11	11	0.73	0.73
5	Pakistan Super League	14	12	0.93	0.80
6	Fakhar Zaman Cricketer	12	12	0.80	0.80
7	Suzuki Cars	14	15	0.93	1.00
8	Indus River	13	12	0.87	0.80
9	David Miller	12	10	0.80	0.67
10	Qamar JavaidBajwa	13	14	0.87	0.93
11	Urdu Alphabets	12	13	0.80	0.87
12	Micheal Hussey	15	15	1.00	1.00
13	George Bush	14	14	0.93	0.93
14	Salman bin Abdulaziz	11	10	0.73	0.67
15	Blue Line Shark	11	13	0.73	0.87
16	National Mosque of Pakistan	12	14	0.80	0.93
17	Lawn Tennis	12	11	0.80	0.73
18	Mehdi Hassan	14	15	0.93	1.00
19	ABC Alphabets	15	15	1.00	1.00
20	Jerry Cartoon	15	15	1.00	1.00

V. CONCLUSION

In this era, data has been growing on daily basis over the internet and one of the important things is to gather data in no time. In this research, we have developed a technique of information retrieval and information filtration processes. There are several algorithms of information filtration and information retrieval that currently have been developed, but the key issue which is still to be addressed is the design of an accurate information retrieval.

In this research, a novel technique is applied which can be considered helpful and may outperform in creation of results more precisely and efficiently as compared to the original information retrieval and filtration systems. Different tests against different queries have been done and the precision of each tested query is calculated to find out the result for future use in average precision calculation of all the techniques. After comparing the results of our proposed system with relevance feedback on different queries, we have found that our proposed system has been able to improve the original Bing API collected results. Our developed algorithm provided more accurate and precise results in fetching the more relevant information as compared to the Bing API.

This study could also be helpful in the development of text-based information retrieval and filtration systems in the near future. We have used the Sequence Matcher Method individually in this paper which could be combined with relevance feedback to explore the accuracy in future work. The idea of comparing sequence matcher method with Relevance feedback is worth trying and we are sure that it would be more effective and precise.

ACKNOWLEGMENT

The authors would like to thank Qatar National Library (QNL) for supporting in publishing the paper.

REFRENCES

[1] Mok, et al: Did distance matter before the Internet? Interpersonal contact and supporting the 1970s, Social Networks, (2007).

[2] Nandha kumar Pandti1, Mohsin Nargund: Mobile Communication -Past, Present and Future: A REVIEW, (2018).

[3] Saad Farooq: Aversarial Information Retrievalon the Web (2018).

[4] Balwinder Siani, Vikram Singh and Satish Kumar. "Information Retrieval Model and Searching." International Journal of Advance Foundation and Research in Science & Engineering (IJAFRSE), (2014).

[5] Vicente Ordonez, Xufeng Han, Polina Kuznetsova, Margaret Mitchell. "Large Scale Retrieval and Generation of Image Descriptions." Springer Science,2015.

[6] Ilyankou, Ilya. "Comparison of Jaro-Winkler and Ratcliff/Obershelp algorithms in spell check." IB Extended Essay, 2014.

[7] Bhakar Mitra, Nick Craswell, Neural Model of Information Retrieval (2017).

[8] K.Kumaravel. "Comparative Study of 3G and 4G in Mobile Technology." IJCSI International Journal of Computer Science Issues, 2011.

[9] W.F Du, G.X. Chen. "Analysis and Research of Several Problems of Bad Short Message Filtering System." International Conference on Computer Information Systems and Industrial Applications, 2015.

[10] M. Rami Ghorab, Dong Zhou, Alexander O'Connor, Vincent, "Personalised Information Retrieval: survey and classification." © Springer Science, 2012.

[11] Chu-Xu Zhang, Zi-Ke Zhang, Lu Yu, Chuang Liu, Hoa Liu, Xiao-Yong Yan. "Information filtering via collaborative user clustering modeling." Elsevier, 2011.

[12] Gourav Bathla, Rajni Jindal. "Similarity Measures of Research Papers and Patents using Adaptive and Parameter Free Threshold." International Journal of Computer Applications (0975 – 8887), 2011.

[13] Diana Mok, Barry Wellman, "Did distance matter before the Internet?" Elsevier, 2007.

[14] SAJ Zaidi, A Buriro, M Riaz, A Mahboob, MN Riaz - Implementation and Comparison of text base Image Retrieval Scheme International Journal of Advanced Computer Science …, (2019).

[15] Alan F. Smeaton, Edel O'Connor, FionaRegan. "Multimedia information retrieval and environmental monitoring: Shared perspectives on data fusion." Elsevier, 2013.

[16] Christopher D. Manning, Parbhakar Raghavan, Hinrich Schutze. An introduction of Information Retrieval. Cambridge: Cambridge University Press, England, 2009.

[17] Sharan Narang, Gregory Diamos, Erich Elsen, Paulius Micikevicius, Jonah Alben, David Garcia, Boris Ginsburg, Michael Houston, Mixed Precision Training, 2017.

[18] Upendra Shardanand, Pattie Maes, Social Information Filtering Algorithms for Automating "Word of Mouth" May 1995.

[19] Nicholas J.Belkin, W. Bruce Croft. Information filtering and information retrieval: two sides of the same coin? Communications of the ACM 1992.

[20] Tanmoy Mondal, Nicolas Ragot, Jean-Yves Ramel, Umapada Pal, Flexible Sequence Matching Technique: Application to Word Spotting in Degraded Documents 2014 14th International Conference on Frontiers in Handwriting Recognition.

[21] Mike Thelwall, Pardeep Sud, "WeboMetric Research with the Bing Search API 2.0" Journal of Informetrics Pages 44-52 2012.

[22] MM Eltoukhy, I Faye, BB Samir: Breast Cancer Diagnosis Based on Texture Feature Extraction Using Curvelet Transform, International Congress on Instrumentation and Applied Sciences. Kuala Lumpur, Malaysia 2010.

Single Modality-Based Event Detection Framework for Complex Videos

Sheeraz Arif[1], Adnan Ahmed Siddiqui[2], Rajesh Kumar[3]
Avinash Maheshwari[4], Komal Maheshwari[5], Muhammad Imran Saeed[6]
Department of IT, Barrett Hodgson University, Karachi, Pakistan[1]
Department of Computing, Hamdard University, Karachi, Pakistan[2, 3, 5, 6]
Department Applied Mathematics and Data science, Hochschule Mitteida University, Mitteida, Germany[4]

***Abstract*—Event detection of rare and complex events in large video datasets or in unconstrained user-uploaded videos on internet is a challenging task. The presence of irregular camera movement, viewpoint changes, illumination variations and significant changes in the background make extremely difficult to capture underlying motion in videos. In addition, extraction of features using different modalities (single streams) may offer computational complexities and cause abstraction of confusing and irrelevant spatial and semantic features. To address this problem, we present a single stream (RGB only) based on feature of spatial and semantic features extracted by modified 3D Residual Convulsion Network. We combine the spatial and semantic features based on this assumption that difference between both types of features can discover the accurate and relevant features. Moreover, introduction of temporal encoding builds the relationship in consecutive video frames to explore discriminative long-term motion patterns. We conduct extensive experiments on prominent publically available datasets. The obtained results demonstrate the great power of our proposed model and improved accuracy compared with existing state-of-the-art methods.**

***Keywords*—Event detection; single-stream; feature fusion; temporal encoding**

I. Introduction

Detection of events in complex and untrimmed videos has been the topic of great concern for many years. Furthermore, it is imperative for many real-world applications such as video indexing, video retrieval, and video surveillance. However, event detection in videos became very challenging due to the different environmental and video recording factors. Video captured from different devices show lots of variations such as variations in environment and variations in recording setting. Variations in environment are due to the occlusion, confusing background, rapid changes in background in video scene, camera motion, noise and viewpoint changes. Variations in video recording also cause different kinds of noise in different lighting conditions. In addition, video low resolution and its high dimensionality may also degrade accurate detection of complex events. Moreover, existing available event detection datasets are too complex and large amounts of uploaded videos on internet are captured in unconstrained conditions. To combat these challenges, there is an immense need for effective and robust activity recognition system to achieve best performance.

In contrast to the simple human action recognition, event detection is a semantic composition of many atomic concepts and there may be involvement of various objects and actors with their different locations and appearances. In addition, videos related to event detection may be of longer duration with multiple scenes and mostly focus on real-world scenarios. For example, the event of "wedding ceremony" in which there are so many related sub-activities with different actors and objects, which can infer the event with a high probability.

Over the past decade, several low level and high-level representations have been proposed to address the issues in context of video event detection. Early attempts are an extension of static image-based representations and pattern recognition. Initially, trajectory-based representations [1-3] have been introduced and obtained satisfactory results. These models utilized Gaussian mixture and Hidden Markov models for the extraction of trajectories and work well for detection of deviant trajectories in less crowded scenes. However, these trajectory-based methods are occlusion-sensitive and not ideal for crowded scenes. These issues are well addressed by hand-crafted methods for example, Histogram of arranged angle (HOG) [4], Histogram of Optical Flow (HOF) [5] and Motion Boundary Histogram (MBH) [6]. These models construct the template behavior and model the background, shape, appearance, and motion and yielded remarkable results. However, these models are only specific to the simple events and do not link between local patterns. Many methods followed Bag of Visual Words (BoVW) by applying dense sampling or detecting spatiotemporal interest points. However, these methods ignore the intrinsic difference between video volumes.

Recently, deep learning achieved a remarkable breakthrough in the image domain and many researchers start applying those learning Spatio-temporal clues by extending deep 2D Convolution Network with 3D Convolution Network [7-13]. These deep learning methods providing high discriminative power and have produced promising results for action recognition. However, CNN based strategy just concentrates visual appearance highlights and comes up short on the capacity to long-run worldly displaying. Most of the researchers implement temporal modeling by introducing two stream-based CNN learning models by applying an extra input stream known as stacked multi-frame dense optical flow along with raw RGB stream.

However, these two stream approaches are not able to capture the motion and semantic changes accurately and only limited to short-term temporal modeling. In light of the above discussion, this research paper proposes lightweight event detection framework by considering only RGB data and address the disadvantages of optical flow in complex and unconstrained videos. Based on assumption that motion can be represented in series of video frames and temporal dynamics of an actor/object can be observed by computing the difference between appearance and semantics. First, we extract two kinds of video features mainly spatial and semantic features by using convolutional and fully connected layers respectively by utilizing by modified 3D Residual Conv Network. Next, we join both low-level (spatial) and high-level information (semantic). To weaken the effect of semantic gap, we add extra learnable filters on the output of different layers. Then, frame-level representation is achieved by employing global average pooling. We also design attention model to take deep insight within the neural network to find important parts in video and ignoring the redundant features and background noise effect for finding the temporal discriminative patterns, temporal encoder is introduced to achieve clip-level representation. Finally, the specific classifier is used to identify the event. The main contributions of this research are listed as under:

1) We only consider RGB stream to extract both spatial and semantic information and extract the motion of the object via the changing of both features and take aside the use of optical flow.

2) We introduce global averaging pooling to represent frame-level representation along with attention mechanisms to learn temporal focus of action.

3) The temporal encoder is applied to detect motion in series of frames.

4) The proposed model experimentally demonstrates the super performance when evaluated on publically benchmark datasets and obtained state-of-the-art results.

The rest of article is organized as follows: Section 2 provides the high level of related works. In Section 3, we present our approach in detail. In Section 4, we demonstrate the experimental evaluation. Finally, conclusion is drawn in Section 5.

II. Related Work

For video analysis, many previous methods adopted a similar approach to image analysis. The video domain is different and complex from the image domain due to the ever-changing motion patterns with target actors/objects and their appearances in different scenes. For the accurate and robust video event detection motion resides in temporal dimension plays crucial role. Many spatiotemporal representation methods such as HOG, HOF, HOG3D [14] and SIFT3D [15] have been proposed to present the motion in a video sequence. In these models extracted features are encoded or pooled in hierarchal structure before feeding to the Support Vector Machine (SVM) classifier. To take the full advantage of motion features dense-point trajectory model [16] has been proposed. These all hand-crafted features models have shown remarkable performance, however, there are several weaknesses are present. These models are computationally expensive and do not consider the changes in semantic clues along the temporal dimension. In addition, features extracted by these schemes are not very discriminative and limited to only simple event detection. Recently, deep convolutional neural networks (DCNN) have achieved great success in many research areas such as object/action detection, classification, recognition. These networks have great potential to learn features automatically from a large datasets. Most of these networks are the natural extension from 2D CNN which are now using in time dimension to represent motion using 3D sensitive filters. More recently, Kinetic 3D networks such as ResNet-3D [16] and I3D [17] obtained great success in the area of action analysis and event detection. However, simultaneously learning appearance and motion brings complexity in the process. Most of the models adopted optical stream as an additional information methodology to catch movement portrayal, for example, [9]. In this model author utilized stacked of dense optical flow as extra stream along with RGB stream to extract static and motion features respectively. The phenomena of optical flow introduced computational complexity and also optical flow may not very robust and accurate capturing semantic and motion changes. In addition, this practice is not ideal for real-world untrimmed and unconstraint videos due to the irregular camera movement. Extra computation by optical flow may degrade the efficiency of event detection framework. Based on this analysis, it is required to re-think the capturing process of motion for complex event detection. This research study represents a single stream model for spatial and temporal feature extraction by considering only RGB frames. RGB frame represented by high dimensional features such as background, objects, and actors. RGB single frame usually encodes static information; however, this study observes the object motion by analyzing the difference in both extracted features i.e. appearance and semantic features. We follow the modified version of ResNet and utilized convolution and fully-connected layers to extract spatial and semantic features respectively. We combine these two features to obtain frame-level representation by using global average pooling. We also introduce a special type of encoding scheme i.e. temporal encoding to achieve clip-level representation.

III. Proposed Framework

Before this section provides a detailed description of our framework, which inputs untrimmed RGB frames of video and detects the event accordingly. The overall flowchart of our method is demonstrated in Fig. 1. Our framework mainly comprises of feature extractor, feature encoder, and classifier. We explain the detailed description of each step in the following sub-section.

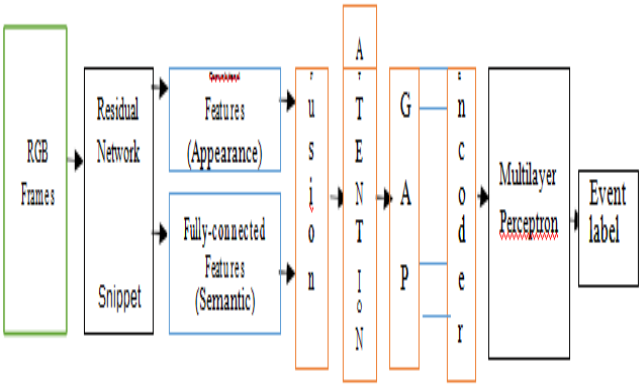


Fig. 1. Feature Extraction and Encoding.

A. Frame-level Representation

For the extraction of appearance and semantic feature representation the pre-processed RGB data is used as input. Deep Residual networks [18, 19] with deep layers are followed as our base network. We combine the low level and higher-level features by using fusion scheme and global average pooling (GAP) is adopted to achieve frame-level representation of video.

1) Residual network: Residual networks are developed similar to VGG networks [20] in which layers are arranged to learn residual function with respect to the given input. Residual networks play an important role to avoid the information bottleneck by introducing residual units. This practice allows skipping connection which permits direct signal propagation between the first and final layers of the network. This network is pre-trained on ImageNet as basic architecture. Residual networks comprise of small 3 x 3 spatial filters with 1 x 1 filters for learned dimensionality expansion and reduction. The network takes input of size 224 x 224 which is the reduced size by stride 2. To prevent the direct fitting of underlying mapping P(l), a residual mapping G(l) = P(l) – l is introduced by training deep network. We can represent the residual unit as follows:

$$l_{n\,1} \quad (l_n \quad G(l_{\,n}\,; Wn)) \qquad (1)$$

Where l_n and $l_{n\,1}$ are the input and output of the n^{th} layer of the network, $G(l_n\,; Wn)$ is a nonlinear mapping based on residual given by Weights of convolutional filters.

$Wn\ \{Wn,\ s\ |1\ s\ S\}$ with $^{S}\ \{2,3\}$, and represents the ReLU function. This practice can achieve direct propagation across all layers of network. In addition, the problem of gradient explosion and disappearance can be avoided. Another advantage is that short connection does not introduce extra computational complexity and parameters. Moreover, ResNet follows the batch normalization (BN) before the activation layer which not only addresses the issues of covariant shift but also speeds up the performance of network.

2) Extraction of appearance and semantic information: We utilize the RGB data as input to the ResNet and to perform the sampling on the data, we adopt two different sampling strategies i.e. dense sampling and sparse sampling. For thick testing, every video is partitioned into T cuts with length of 1-2 seconds and afterward we haphazardly select a picture/outline from each clasp and organize them in an arrangement {N1, N2… NT}. For meager examining, we select three edges of equivalent span from video arrangement and receive setting rules given in [21]. As referenced before, movement of any article/entertainer can be investigated by means of the distinction of both appearance and semantic highlights. The yield of the profundity layers for the most part gives the elevated level (semantic) highlights. In our base organization, the yield highlights of both convolution and the completely associated layers are extraordinary. The output of convolution layer is appearance features (outline, shapes, etc.), while fully-connected layer provides semantic features (rotation invariance, and location invariance). Our baseline CNN generates two feature maps for the nth frame. The last pooling layer of the network generates feature maps fcln and fully-connected layers outputs feature maps f fcln. Both feature maps having the dimension (W x H x D, C.), representing width, height, temporal depth and number of feature channels respectively. The matrix representation of both feature maps for the video length of duration T can be given as:

$$\text{fcln} \quad [\ \text{fcl1}\ ,\ldots,\ \text{fclt}\ ,\ldots,\ \text{fclT}\ ]\ ¡ \qquad \text{WHTC} \qquad (2)$$

$$\text{f fcln} \quad [\ \text{f fcl1}\ ,\ \text{f fcl2}\ ,\ldots,\ \text{f fclt}\ ,\ldots,\ \text{f fclT} \qquad]\ ¡\text{DC} \qquad (3)$$

3) Fusion and attention mechanism: In the previous sub-section, we obtain two feature maps produced by the pooling layer and FC layer. Next, we perform weighted linear fusion scheme to integrate both appearance and semantic features by employing pixel-wise operation. After fusion, we again obtain a frame blended with both spatial and semantic properties. Then, we apply attention mechanism by computing the weights of both appearance and semantic features. The purpose of attention model is to decide important frames in a video for event recognition. The attention mechanism is very close to human visual model as humans always concentrate and focus on moving objects instead of whole frame or static background. In addition, it plays important role to eliminate the effect of background noise and adds a dimension of interpretation ability. If we assume that W is the weight mapping of both appearance and semantic information of tth element of frame and N is the number of frames then the probability of informative frame can be represented as follow:

$$= \frac{\exp(W^{T})}{(W_{t}^{T})^{\,\mathrm{N}}} \qquad (4)$$

Where t is the probability with which the corresponding frame is considered an informative frame. Finally, we get the vector representation of each selected (attention mechanism) frame by using global average pooling. The function of the global average pooling (GAP) layer is to average the feature values of the respective pixels in each chosen frame, and the average value is taken as the probability value of each feature. After applying this pooling scheme, a video can be represented as a sequence of vector V = {v1, … vM }of M clip of input video. Each vm¡ d is the expression of M video sequence i.e. Sm.

B. Video Level Representation

As we mentioned earlier that temporal information is presented in sequence of video frames. To model the relationship between video frames we introduce a temporal encoder E. We can combine the different features from the entire video sequence into powerful and compact clip level representation. If V = {v1, … vM }is the input to the encoder then clip-level representation can be obtained by applying simple function or neural network. This research work, apply and compare the three different encoders mainly Average Encoder, Max Encoder, and LSTM Encoder. All these encoders take the sequence of vectors of M video clips and generate video representation as a single vector Z such as Z ¡ d. Where Z is the vector representation of video integrating the high-level semantic and low-level appearance features along with temporal relationship. We can define working of each the encoder as follows: Average encoder: This encoder performs the element-wise addition on the feature vectors and compute the single feature vector using the length of M video clips as:

$$Z \quad \frac{1}{M} \quad (v \quad v \quad v \ldots v \quad) \qquad (5)$$

1) Max encoder: This encoder represents a video by a single vector using maximum feature value (highly weighted) from the list of finite values and can be given as under:

$$Z \max(v1, v2, v3 \ldots vM) \qquad (6)$$

2) LSTM encoder: This encoder outputs the feature vector Z using the hidden state of the LSTM hj at time step j and feature vector vj.

$$h_j \quad LSTM(v_j, h_{j-1}), 0 \quad j \quad M \qquad (7)$$

C. Event Classification

We require a prediction function F(Z) to detect the event category for the given video. We adopt a multi-layer perceptron as classifier which comprises of FC-Dropout-FC pipeline. The dropout option is used to prevent the framework from overfitting. If yˆ is the prediction of classier and y is the ground-truth label of the video then final loss can be formulated as:

$$L(\hat{y}, y) \sum_{i=1}^{C} y_i (\hat{y} \quad \log \quad \exp y_j) \qquad (8)$$

In addition, if our temporal encoder E is differentiable, so our network can also be differentiable. We can utilize the multiple frames to jointly optimize the model parameter W with the standard back-propagation scheme. We can compute the loss P by using the chain rule using gradient W as follows:

$$\frac{P(\hat{y}, y)}{W} \quad \frac{P}{\hat{y}} \frac{\hat{y}}{Z} \sum_{i=1}^{M} \frac{Z}{v_i} \frac{v_i}{W} \qquad (9)$$

IV. Experiments and Analysis

Evaluate the performance of our proposed framework, we carried out several experiments on challenging publically available datasets. We analyze our introduced method using different aspects. The portrayal of datasets with their approval plans, trial arrangement, results, and similar examination are introduced in ensuing segments.

A. Datasets

For our experiments, we use well-known event detection datasets, namely YLI-MED [22], MEDTest-14 [23] and Columbia Consumer Video (CCV) [24].

YLI-MED: This dataset comprises of 1823 videos and each of them classified is classified into 10 event categories. These videos are divided into training (1000 videos) and testing (823 videos). The Videos length duration is variable, which makes event detection more challenging. We measure the accuracy of the test set for all experiments. Columbia Consumer Video (CCV): The Columbia Consumer Video dataset contains 9,317 videos in total from 20 semantic categories, including events like “parade” and “baseball”.

B. Implementation Setup and Details

We extract the RGB frames from the original video by using the guidelines given in FFMPEG [25]. For the training of our model, we augment the extracted images to reduce the effect of overfitting. We horizontally flip input images with 55% probability and then crop them by resizing of 320 x 240. We scale the height and width of cropping rectangle by a randomly selected factor of in the range of 0.8 ~1. We utilize ResNet50 which network weights are initialized by pre-training on ImageNet. We replace the final classifier with a two-layer perceptron. The unit number of FC layer is set to 512. The dropout ratio is set to 0.8. For LSTM encoders, we set up one hidden layer with 512 units. The momentum of stochastic gradient is selected as 0.9 for optimizing the model. All experiments are conducted on a single GPU with weight decay of 1x10-4 and mini-batch of size 16. The initial learning rate is set as 0.003 and decreased to 12% at 150 epochs. The whole training procedure is stopped at 300 epochs.

C. Experiments and Discussion

We direct broad tests to assess the exhibition of our proposed technique. In this part, we introduced significant trial results and execution investigation. We direct broad tests to assess the exhibition of our proposed technique. In this part, we introduced significant trial results and execution investigation.

1) Exploration results: First, we tested our model by employing different exploration aspects. We conducted our experiments on YLI-MED and (CCV) datasets and use all videos associated with 10 event categories of YLI-MED dataset and 20 event categories of CCV. We explore the performance of our proposed method by using convolution features, semantic features and fusion of both features with and without using attention mechanism and obtained results are demonstrated in Table I and Table II.

TABLE I. MAP(%) ON YLI-MED DATASET ON DIFFERENT FEATURE INFORMATION

Feature Information	With attention	Without attention
Convolutional	76.2	74.5
Fully-Connected	78.9	77.1
Fusion (Both)	**82.2**	81.1

TABLE II. MAP(%) ON CCV DATASET ON DIFFERENT FEATURE INFORMATION

Feature Information	With attention		Without attention	
Convolutional	71.2			69.5
Fully-Connected	74.9			72.1
Fusion (Both)	**78.2**			75.1

It can be viewed from the obtained results, the introduced method (Fusion of appearance and semantic features) performs better than using convolution (appearance) and fully-connected (semantic) features separately. This illustrates that it is necessary to combine both appearance and semantic features in the temporal domain. This practice can discover more useful information for robust and accurate event detection. It can be also observed that introduction of attention mechanism yields improved performance especially in YLI-MED dataset and performs better than without attention model in both datasets. This attention mechanism provides insight for finding important parts of the video and prevents the background noise, thus, play important role to achieve better event recognition accuracy.

2) Effect of sampling strategies and encoders: We also carried out some experiments to analyze the effect of sampling strategies and encoders on our proposed model. Our model takes a series of frame sequence and we use two different sampling strategies: dense sampling and sparse sampling. For dense sampling, each video is divided into T clips with duration of 1-2 seconds and then we randomly select an image/frame from each clip and arrange them in a sequence. For sparse sampling, we select 3 frames of equal duration from video sequence and adopt setting guidelines given in [21]. We use YLI-MED and Columbia Consumer Video (CCV) dataset for these experiments and consider all events categories in both datasets. We explore the capacity of both sampling strategies using three encoders i.e. Max encoder, Average encoder, and LSTM encoder. The obtained results can be shown in Fig. 2 and Fig. 3. It can be seen from results that dense sampling achieves better performance than sparse sampling in the presence of max encoder. The possible reason is that sparse sampling may miss more crucial and important frames of the video sequence as compared to dense sampling and there may be loss of some important semantic features. We also analyze the performance of three different encoders in this experiment in the presence of both sampling strategies. According to the result, max encoder obtains the best performance against average and LSTM encoder. The underlying reason is that max encoder strengthens the features which are useful for specific event over a long-range. Both average and LSTM encoders perform similarly on both datasets. We can observe that performance of both encoders is relatively moderate. One reason is that both encoders are complex encoders as they have more parameters may lead to over fitting problem. We will adopt dense sampling with max encoder for our framework for all remaining experiments.

3) Class-Wise accuracy for event classification: We further investigate the event classification accuracy of our method by constructing the confusion matrix of two datasets i.e. YLI-MED and CCV datasets. The confusion matrixes of our introduced approach on both datasets can be depicted in Table III and Fig. 4. The confusion matrix indicating the accuracy of each action and correspondence between the target classes along x-axis (true label) and output classes (predicted label) along y-axis. We consider 10 event categories from YLI-MED dataset and 16 event categories from CCV dataset to conduct our experiment. Table III demonstrates the accuracy of each action category in the form of confusion matrix. The intensity of the true score is high (diagonal) for each category, and our method achieves 83% for all 10 event categories. It is interesting to note that some of categories with similar actions are more easily confused with each other, such as Birthday Party (Event-1), Wedding Ceremony (Event-9) and grooming an animal (Event-6), hand-feeding an animal (Event-7); these classifications meddle with one another and yield low scores. A potential purpose behind this is the comparability of the highlights and portrayals among activities. Be that as it may, our proposed approach actually performs well with the majority of function classes.

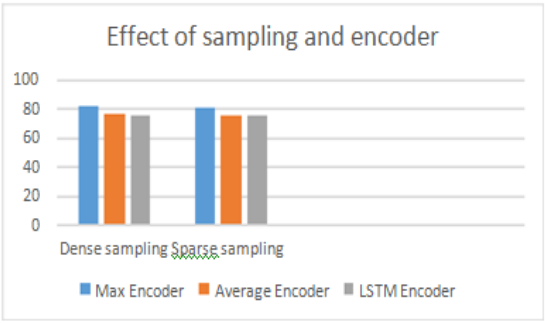


Fig. 2. Comparison of different Sampling Strategies and Encoders on YLI-MED Dataset.

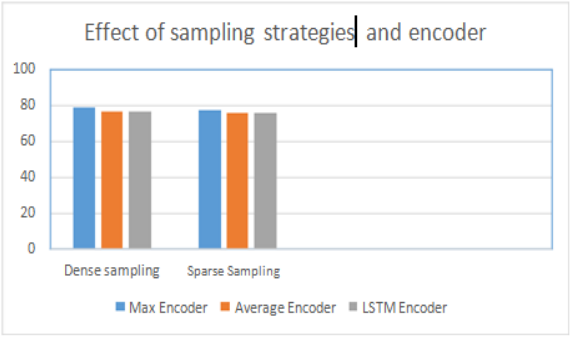


Fig. 3. Comparison of different Sampling Strategies and Encoders on CCV Dataset.

TABLE III. CONFUSION MATRIX ON THE YLI-MED DATASET USING OUR MODEL

Categories	Event-1	Event-2	Event-3	Event-4	Event-5	Event-6	Event-7	Event-8	Event-9
Event-1		0	0.01	0	0	0.01	0.02	0	0.04
Event-2	0		0	0.17	0	0	0.02	0	0.04
Event-3	0	0.17		0.02	0.05	0	0.02	0.04	0
Event-4	0	0.16	0.01		0	0	0	0	0
Event-5	0.01	0.01	0.02	0.01		0	0	0.02	0
Event-6	0.02	0	0	0	0		0.09	0	0
Event-7	0.03	0.02	0.01	0	0.01	0.16		0.04	0
Event-8	0	0	0.03	0	0.05	0	0.03		0
Event-9	0.11	0.02	0	0	0	0	0.03	0	
Event-10	0	0.01	0.02	0	0.02	0.02	0	0.03	0.05
Average									
Accuracy									

In addition, we also investigate the class-wise recognition accuracy of our method by constructing confusion matrices of CCV datasets. We consider 16 event classes from this dataset. The confusion matrices are given in Fig. 4. In this figure, the x-axis represents the classified labels of action classes whereas y-axis denotes the ground truth label. The accuracies in the diagonal cells are indicated by different colors and yellow cells show the 100% accuracy achieved for the particular action class. From the results, it can be seen that both of the confusion matrices are well diagonal zed. However, some of the action classes are giving low prediction scores by giving different colors of cells other than yellow it means few categories are mixed up when classifying. The possible reasons for interfering and misclassification are the motion similarity in actions or the same background, objects and appearance and motion-based features. However, most of the scores are well diagonal zed.

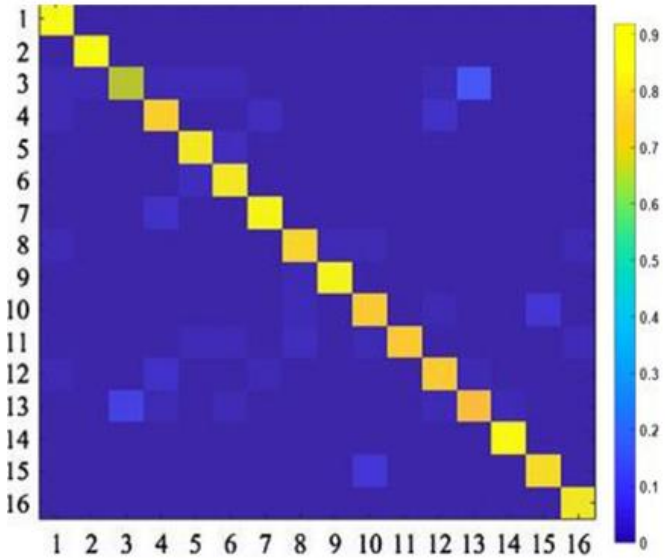


Fig. 4. Confusion Matrix on the CCV Dataset using our Model.

4) Visualization of feature embedding: Furthermore, we investigate the discriminative power of our learned fused features for human activity recognition. We consider the 10 different event categories (Birthday Party, Flash Mob, Vehicle unstuck, Parade, Board trick, grooming animal, Feeding animal, Landing a fish, Wedding Ceremony, Woodworking Project) from YLI-MED datasets. For each of the event category, we utilize 30 video clips of each event class for our experiment. Each video clip can be viewed by a single color point and we used the same color for al videos related same action class. For successful recognition of these action classes, an action recognition framework must possess high discriminative power. We adopt the method of t-SNE visualization [34] and show the visualization of feature representation embedding extracted by our introduced approach in Fig. 5. It can be observed from results that our method provides the better-separated clusters and clip-level features are semantically well separated as compared to the other existing prominent methods (two-stream model, and C3D). Thus, we can conclude that our proposed method can integrate both appearance and semantic features and possesses high discriminative information.

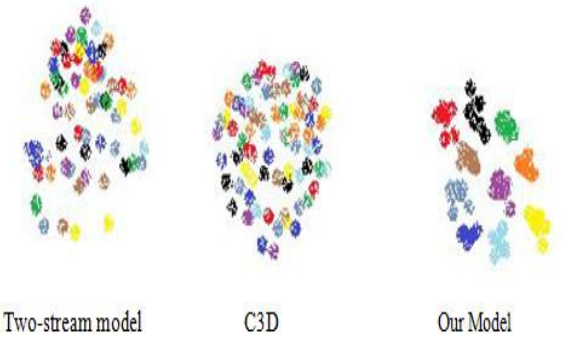


Fig. 5. Visualization of Clip-Level Features Embedding Learned by Two-Stream Model, C3D and Proposed Model.

5) Comparison with the state-of-the-art models: In this part, we further check the viability and plausibility of our model, we contrast our proposed approach with various existing cutting edge human activity acknowledgment draws near on both YLI-MED and Columbia Consumer Video (CCV) datasets for all videos of each event category. The comparison results are reported in Table IV in which average detection accuracies (%) are reported for both datasets. We consider two-stream models such as two-stream ConvNets [9], Two Stream 3D-Nets [26] and two-stream Fusion [27]. We also consider existing state-of-the-art hybrid model-based techniques such as TDD-IDT [28] and MTC3D-IDT [29] and P3D+IDT [30]. All of these models follow improved dense trajectories (iDT) for trajectories extraction and adopt higher-order encoding scheme i.e. Fisher Vector (FV) to encode the hand-crafted features.

TABLE IV. COMPARISON OF PROPOSED METHOD WITH THE STATE-OF-THE-ARTS APPROACHES ON YLI-MED AND CCV

Modality	Method	Input	YLI-MED (%)	ccv (CCV) (%)
Two-Streams	Two-stream ConNet [9]	RGB	69.9	58.9
	Two-stream ConNet [9]	O.F	53.4	49.2
	Two-Stream 3D-Nets	RGB	71.9	66.4
	[26]	O.F	64.3	61.7
	Two-Stream 3D-Nets	RGB	75.2	66.9
	[26]	O.F	67.3	61.5
	Two-stream Fusion [27]			
	Two-stream Fusion [27]			
Hybrid	TDD-IDT [28]	RGB	77.2	74.3
	MTC3D-IDT [29]	RGB	76.2	73.9
	P3D+IDT [30]	RGB	79.3	72.7
Very deep ConvNet	C3D [31]	RGB	65.6	63.2
	3D-ResNet [32]	RGB	72.6	69.0
	TSN [33]	RGB	74.5	70.0
Ours	SM-AB	RGB	**82.2**	78.2
	SM-AB	O.F	69.1	66.8

For the two-stream models, we analyze their performance on both stream i.e. RGB and optical flow and we can notice that performance of the optical flow is worst against the RGB images. This phenomenon verifies the assumption that the optical flow (O.F) in less flexible and inaccurate to capture motion of object due to the movement of camera and large-scale perspective transformation in complex videos. We also analyze the performance of some hybrid-features model in which features from both domains i.e. deep learning and hand-crafted features (improved dense trajectories) are incorporated and obtained competitive results, however, our approach outperforms them by fair margin on both datasets. We also compare our model with existing prominent and successful 3D convolution based methods such as C3D [31] and 3D-ResNet [32] and Temporal Segment Network model TSN [33]. Our approach possesses higher discriminative power and our system to be on par with the state-of-the-art. We compare the performance of our proposed model on both modalities i.e. RGB and Optical flow data and we achieve far better results when using only RGB frames so obtained results suggest that temporal long term dynamics can be capture from RGB frames. Thus, from results we can say that our model in the presence of only RGB data explores more relationships between video clips and semantic features and introduction of max encoder works well by capturing the long-term dependencies and successful for the detection of complex events.

V. CONCLUSION

This paper proposes a new lightweight framework for video event detection, which comprises CNN, features fusion, attention mechanism, and global average pooling. This framework obtains high representational power and finds the discriminative patterns in complex videos for event detection. We just use the RGB data to extract appearance and semantic features for each frame of video using convolution and fully-connected layers. This practice avoids the additional computational power required by optical flow. We explore the motion by computing the difference between both semantic and appearance features. We also employ the attention mechanism to concentrate and focus on key frames keeping motion information and avoiding the redundant effect of static background. Furthermore, we utilize temporal encoder to establish temporal relationships between frames and explore discriminative long-term motion patterns. The introduced model achieved promising performance when tested on two widely used challenging datasets. In future work, we will try to improve the sampling strategy or may modify the pooling or fusion layers in the network.

ACKNOWLEDGMENT

The authors would like to thank the editor and reviewers for their work on this manuscript.

REFERENCES

[1] F. Jiang, J. Yuan, S.A, Tsaftaris and A.K Katsaggelos. “Anomalous video event detection using spatiotemporal context”. Comput. Vis. Image Underst. pp. 323–333, vol. 115, 201.

[2] B.T Morris and M.M Trivedi.”Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach”. IEEE Trans. Pattern Anal. Mach. Intell. pp. 2287–2301, vol. 33, 2011.

[3] S. Calderara, U. Heinemann, A. Prati, R. Cucchiara and N. Tishby.”Detecting anomalies in people’strajectories using spectral graph analysis”. Comput. Vis. Image Underst. pp. 1099–1111. Vol. 115, 2011.

[4] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 886–893, June 20-25, 2005.

[5] H. Wang, A. Klaser A and C. Schimid, “Dense trajectories and motion boundary descriptor for action recognition,” in proceeding international journal of computer vision, vol. 103, pp. 60-79, March, 2013.

[6] Wang H, Ullah MM, Kl¨aser A, Laptev I, Schmid C (2009) Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conference.

[7] G.W Taylor, R. Fergus and Y. LeCun, "Convolutional learning of spatio-temporal features," in proceeding of 11th European conference on Computer vision, pp. 140-153, September 5-11, 2010. Article (CrossRef Link).

[8] Ji Si, Xu W, Yang M, et al., "3d convolutional neural networks for human action recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.35, no.1, pp.221–231, January, 2013.

[9] D. Tran, L. Bourdev and Fergus, "Learning spatiotemporal features with 3d convolutional networks," In proceeding of IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp. 4489–4497, December 7-13, 2015. Article (CrossRef Link).

[10] Limin Wang, Yuanjun, Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaou Tang. Temporal segment network: Towards good practices for deep action recognition," in ECCV, 2016.

[11] Julia bernd, Damian Borth, Benjamin Elizade, et al. YL1-Med corpus: characteristics, procedures, and plans, in arXiV: 1503.04250, 2015.

[12] A. Diba, A. M. Pazandeh, and L. V. Gool, "Efficient two-stream motion and ap-pearance 3D CNN for video classification," in Proceedings of European Conferenceon Computer Vision, 2016, pp. 1- 4.

[13] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition", in Poc. IEEE Conf. Comp. Vis. Pattern Recognit., Jun. 2016, pp-1933-1941.

[14] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in Proceeding of IEEE international conference on Computer Vision and Pattern Recognition, 2015, pp. 4305-4314.

[15] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang and M. Shah. High-Level Event Recognition in Unconstrained Videos. In IJMIR, 2012.

[16] X. Lu, H. Yao, and S. Zhao, "Action recognition with multi-scale trajectory-pooled 3D convolutional descriptors," Multimedia Tools and Applications, 2017 pp.1-17.

[17] Z. Qiu, t. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks, "in proc. IEEE Intl. Conf. Comut. Vis., Oct. 2017, pp. 5533-5541.

[18] A. Karpathy, G. Toderici, S. Shetty and T. Leung, "Large-scale video classification with convolutional neural networks," in proceeding IEEE conference on computer vision and pattern recognition, pp. 1725 – 1732, June 23-28, 2014.

[19] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. arXiv preprint arXiv:1507.02159, 2015

[20] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In Proc. ECCV, 2014.

[21] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In Proc. CVPR, 2015.

[22] S. Venugopalan, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence video to text. In Proc. ICCV, 2015.

[23] N. Srivastava, and R. Salakhutdinov. Multimodal Learning with Deep Boltzmann Machines. In NIPS'12, 25, pages 2231-2239.

[24] L. Sun, K. Jia, and D. Yeung, "Human action recognition using factorized spatio-temporal convolutional networks," in proceeding of IEEE International Conference on computer vision (ICCV), pp. 4597 – 4605, December 7-13, 2015.

[25] A. Klaser, M. Marszalek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in proceeding of 19th British Machine Vision Conference, British Machine Vision Association: Leeds, United Kingdom, pp.1–10, September, 2008. Article (CrossRef Link).

[26] P. Scovanner, S. Ali and M. Shah, "A 3-Dimensional SIFT Descriptor and its Application to Action recognition," in Proceedings of the 15th International Conference on Multimedia, pp. 357–360,September 25 – 29, 2007.

[27] H. Wang and C. Schmid, "Action recognition with improved trajectories," in proceeding of IEEE International conference on computer vision, pp. 3551-3558, December 1-8, 2013.

[28] Joao Carreira and Andrew Zisserman, action recognition? A new model and the kinetics dataset, in CVPR, 2017.

[29] Kaming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual for image recognition. arXiv preprint arXiv: 1512/03385, 2015.

[30] Kaming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mapping in deep.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networksfor large-scale image recognition. In proc. ICLR, 2014.

[32] Limin Wang, Yuanjun, Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaou Tang.Temporal segment network: Towards good practices for deep action recognition," in ECCV, 2016.

[33] N. Srivastava, and K. Salakhutdinov. Multimodal Learning with Deep Boltzmann Machines. In NIPS'12, 25, pages 2231-2239.

An Efficient Domain-Adaptation Method using GAN for Fraud Detection

Jeonghyun Hwang[1], Kangseok Kim[2]*
Department of Artificial Intelligence and Data Science
Graduate School of Ajou University, Suwon, Korea[1, 2]
Department. of Cyber Security, Ajou University, Suwon, Korea[2]

***Abstract*—In this paper, an efficient domain-adaptation method is proposed for fraud detection. The proposed method employs the discriminative characteristics used in feature maps and generative adversarial networks (GANs), to minimize the deviation that occurs when a common feature is shifted between two domains. To solve class imbalance problem and increase the model's detection accuracy, new data samples are generated by applying a minority class data augmentation method, which uses a GAN. We evaluate the classification performance of the proposed domain-adaption model by comparing it against support vector machine (SVM) and convolutional neural network (CNN) models, using classification performance evaluation indicators. The experimental results indicated that the proposed model is applicable to both test datasets; furthermore, it requires less time for learning. Although the SVM offers a better detection performance than the CNN and proposed domain-adaptation model, its learning time exceeds those of the other two models when dataset increases. Also, although the detection performance of the CNN-based model is similar to that of the proposed domain-adaptation model, its learning process is longer. In addition, although the GAN used to solve the class imbalance problem of the two datasets requires slightly more time than SMOTE (synthetic minority oversampling technique), it shows a better classification performance and is effective for datasets featuring class imbalances.**

Keywords*—Fraud detection; domain adaptation; data augmentation; deep learning; GAN

I. Introduction

With the rapid development of information technology, the existing financial industry paradigm is changing; the new paradigm, following the evolution of smartphones and mobile technologies, is creating new forms of electronic financial services, increasing the number of non-face-to-face transactions (through the use of various devices and communications technologies), and simplifying and diversifying payment methods. However, alongside these developments, concerns over security incidents (e.g., cyber threats involving the leakage and hacking of financial and personal information) are also increasing, owing to the new approaches facilitated by the Internet, device diversity, transaction simplicity, and ease of data flow. Therefore, the performances of fraud detection systems (FDS) must be improved, to actively respond to these diversified and intelligent cyber threats. Accordingly, machine- and deep-learning based technologies, which learn large quantities of data to improve prediction and classification accuracies, have recently been developed; thus, research incorporating these technologies has increased accordingly, to improve the performances of FDSs. However, the existing FDS's abnormal-transaction-detection method which combines machine- and deep-learning techniques to identify abnormal transactions in large quantities of real-time data is time-consuming and computationally expensive. Therefore, this study presents a faster-learning abnormal-transaction-detection model, by training a model suitable for data across different domains and utilizing the common features and information thereby found. The proposed model to detect anomalies between different domains is constructed using domain adaptation method [1] which is one of transfer learning [2], a machine-learning method that utilizes pre-learned domain information from similar domains when a specific task or domain is changed. The datasets employed in the proposed domain-adaptation method are generally used in research relating to abnormal-transaction detection; in particular, they are benchmark datasets for fraud detection in credit card [3] and financial [4] datasets. However, because both datasets feature an unbalanced ratio between the normal transactions and fraudulent or anomalous ones, the classes must be balanced to improve the machine learning performance and ensure smoothly learning. Then, a data augmentation method can be used to increase the total number of data when datasets are insufficient; this method is applied to the minority class using a generative adversarial networks (GANs) [5]; the augmented data are used for training/test data of the proposed domain-adaptation model, and the results are compared with those of SMOTE (Synthetic Minority Oversampling Technique) [6] which is one of oversampling methods. Therefore, in this study, a GAN and SMOTE are used to solve the class imbalance problem for credit-card and financial-transaction fraud datasets; then, the domain-adaptation method is used to implement a model for detecting abnormal transactions in the two datasets; finally, the method's effectiveness is verified through a comparison of its classification performance against those of support vector machine (SVM) [7] and convolutional neural network (CNN) [8] based methods. The remainder of the paper is organized as follows: In Section II, the background and related research are described; in Section III, the model and datasets employed are described in detail; in Section IV, the experimental environment, learning method, and hyperparameters are described; in Section V, the classification performance of the model is compared and analyzed against those of the SVM- and CNN-based models; and in Section VI, the conclusions and limitations of the research are described, and future research directions are considered.

*Corresponding Author.

II. Related Works

This section describes existing fraud detection methods, data augmentation approaches, and domain adaptation methods.

A. Fraud Detection

Abnormal transaction detection is a data mining approach used to detect transactions that differ from normal transaction patterns. The detection results are divided into two transaction classes: normal and abnormal. A variety of detection technologies are constantly being studied to minimize the risks posed to users by fraudulent transactions. Studies for abnormal transaction detection include the development of procedures for classification (a field of supervised learning), clustering (a field of unsupervised learning), deep learning and so on. In the existing research on classification-model-based abnormal transaction detection approaches, [9] proposed *Very Fast Decision Tree*, which can manage unbalanced data using decision trees; [10] employed a hidden Markov model (HMM) to learn a normal credit card transaction, and they classified transactions that were not accepted by the HMM as abnormal; [11] detected abnormal transactions using *k-Nearest Neighbors*, which offers reduced memory consumption compared to other machine learning methods. Furthermore, [12] proposed a model to detect abnormal transactions and money laundering, by applying an SVM. In addition, deep learning models have been applied to abnormal transaction detection using auto-encoders or GANs as a solution for data unbalancing [13, 14]. In addition, a significant number of abnormal detection models have been proposed to increase the accurate detection rate of FDS. In our study, for the classification performance of fraud detection, the proposed domain-adaptation model was evaluated by comparing it with the SVM and CNN models, which are supervised learning-based analytical models.

B. Oversampling

Approaches to solving the data imbalance problem can be divided into four categories: sampling-based, cost-based, kernel-based, and active-learning-based methods [15]. The approach of changing the distribution between the majority and minority classes in unbalanced datasets is a sampling-based method; the distribution balance can be adjusted to reduce the number of data samples in the majority class (undersampling) or to increase the number in the minority class (oversampling). SMOTE is an oversampling method: it generates data between the minority class' data samples by connecting a straight line between them. Majority Weighted Minority Oversampling Technique [16] identifies minority class data and assigns weights according to the Euclidean distance between them and the nearest data samples in the majority class; then, a clustering approach generates data between the weighted minority class data in the same way as SMOTE. Meanwhile, the Random Oversampling Examples (ROSE) [17] method generates new minority data based on the existing kernel-density estimate; robROSE [18] is an oversampling method that overcomes the shortcomings of ROSE (which can deviate under the influence of outliers). Of the above methods, we used SMOTE to solve the class imbalance problem, because it is easier to implement and understand than other methods and offers excellent performance characteristics.

C. Data Augmentation

Data augmentation, which was first introduced in [19], is a popular method for processing image data; it generates noise whilst preserving the amount of information in the data. GANs are suitable models for performing data augmentation; it consists of two artificial neural networks (ANNs) that learn by competing against each other: one is a generator that receives random noise as an input and processes it to resemble the distribution of the original data; the other is a discriminator that distinguishes the original data from those created by the generator. The generator seeks to make the data it produces indistinguishable from the original data as much as possible, and the discriminator tries to classify the two types of data with the highest possible probability, in opposition to the generator. As a result, data that pass through a network consisting of generators and discriminators are generated with a distribution similar to that of the original data. By varying the structures and purposes of GANs, researchers have successfully applied them to various fields; in particular, the field of image-data-related research [5, 20] has found considerable use for them, and models for increasing their performance and generating new image data have been proposed. Among them, deep convolutional GANs [21] provided guidelines for stable learning, and the *Wasserstein GAN* (WGAN) [22] improved the stability by attributing unsuccessful learning to the limit of the Kullback–Leibler (KL) divergence and redefining the loss function. Most of studies (e.g., [23, 24, 25]) have aimed to improve the network performance for image data. However, some studies have attempted to solve the data imbalance problem using GAN. In particular, the study [25] applied numerical data, not image data, to GAN. However, since GANs learn via the gradient descent method, learning problems can occur due to the loss functions [22]. Therefore, in this study, data augmentation was performed for the minority class data samples of each dataset, by applying the loss function of WGAN to alleviate the GAN's limitations and generate datasets more closely resembling the original data. Because the GAN-based minority class data-augmentation method is similar to the oversampling method, it is applied by integrating it with oversampling techniques rather than data augmentation. Therefore, in this study, we use the terms "data oversampling" and "data augmentation" interchangeably.

D. Domain Adaptation

A transfer learning is a machine-learning method that utilizes pre-learned domain information from similar domains when a specific task or domain is changed. The area in which the transfer learning model previously worked is referred to as the source domain, and the new one is referred to as the target domain; transfer learning, depending on the presence or absence of labels in the domain, is primarily divided into multi-task learning [26], in which the class exists only in the target domain; self-taught learning [27], in which the class exists in the source domain but no classes exist in the target domain; and domain adaptation [1], in which the class exists in both domains. In this study, we consider a domain adaptation model to detect anomalies between different domains. Regarding domain adaptation [28], several previous studies [29, 30, 31] have focused on minimizing the differences between the source and target domain feature-map distributions; most of these have

used the maximum mean discrepancy [32] loss function. *Deep Correlation Alignment* [29] matches the mean and covariance of the two distributions. In [31], the addition of a fully connected layer to the domain adaptation model was proposed, and a method was derived to determine the resulting value of the binary label and approximate the uniform distribution via the domain confusion loss. *ReverseGrad* [30], a gradient-reversal algorithm, calculates the gradient in the reverse direction when deriving the loss function in the network; it has exhibited a faster learning performance than comparable methods. In addition to [30], a study investigating methods of reconstructing images in the target domain was also presented in [31]. In [33], probabilities were used to learn the distribution between the two domains, and the distance between data within the same class across the two domains was expressed as a probability; learning was conducted to maximize this probability. *Adversarial Discriminative Domain Adaptation* (ADDA) [34] applied the loss function used in discriminator of GAN to match the distributions between the two domains, thereby enabling more effective learning. This method has the advantage of being able to interact with other domain-adaptation models. In this study, to facilitate interactions between similar domains, considering the advantages of ADDA, it was applied to the abnormal transaction detection model.

III. Methodology

This section describes a set of approaches conducted for fraud detection in FDSs. Section A describes the experimental dataset used in this study. Section B and C describe data augmentation to solve class-imbalance problems in learning. GAN model was used for data augmentation of minority class through the creation of new samples. It was compared to SMOTE used for data oversampling as well. Section D presents the proposed domain adaptation method, which is capable of evaluating classification performances on two datasets of similar domains. Fig. 1 shows the simplified overall structure of the model proposed in this study, and Fig. 2 illustrates the flow of this structure.

A. Dataset

The credit card dataset here employed consists of data collected by the Machine Learning Group [3] and Worldline. The dataset contains a total of 284,315 normal and 492 abnormal transaction data samples. For the data, owing to security issues (e.g., financial and personal information leaks); the test was conducted using a total of 30 variables. Similarly, the financial transaction dataset is an artificial (owing to security issues) dataset based on actual data. This dataset contains simulation results obtained through PaySim [4], using real financial transaction samples taken over a period of one month; it consists of a total of 11 variables and includes 6,354,407 normal and 8,213 abnormal transaction data samples. Unlike the credit card fraud dataset, this dataset was processed via min-max normalization before being used as input data in this work.

B. Data Oversampling

SMOTE oversamples the minority class data when class imbalances occur; in this study, it was adopted as the oversampling method because it delivers a strong performance whilst also being theoretically simple and easy to implement. First, SMOTE takes the data of a minority class and then finds the k-nearest neighbors of these data. Next, the differences between the current sample and these k neighbors are obtained, multiplied by a random value (between 0 and 1) to generate data, and combined with the original sample. It also shifts the existing data slightly, to account for the neighbors it adds. In this study, SMOTE was implemented using the imbalanced-learn Python library [35]. The oversampled data were tested with ratios of 0.3:1, 0.5:1, 0.7:1 and 1:1 between the minority and majority classes, respectively.

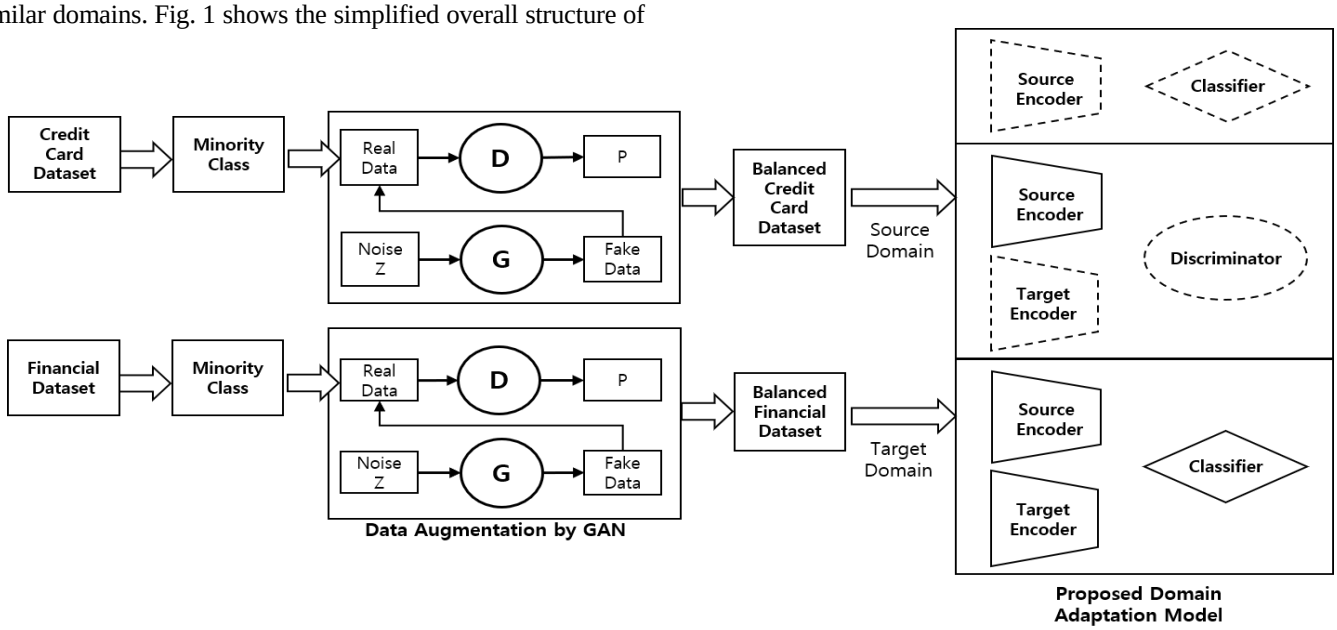


Fig. 1. Simplified Overview of the Proposed Methodology.

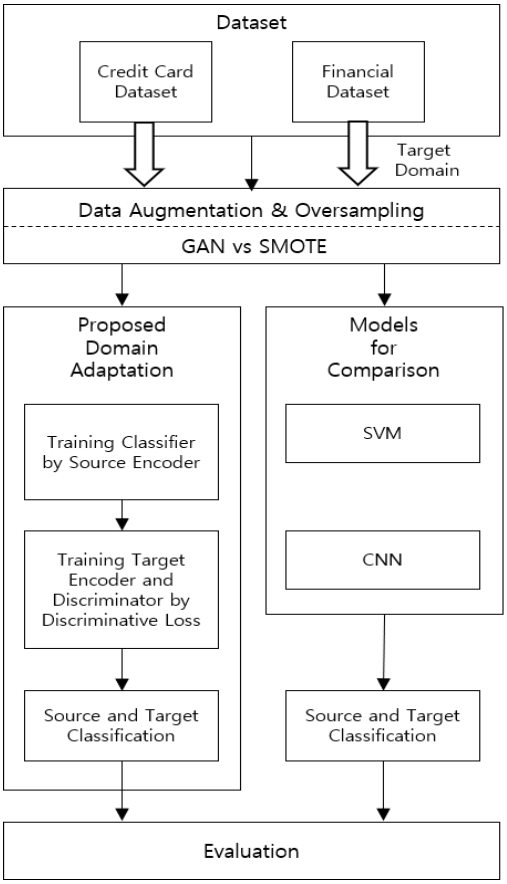


Fig. 2. An Overview Flow Chart of the Proposed Methodology.

C. Data Augmentation using GANs

In existing GANs, several problems can arise when training the GAN via the gradient descent method [22]. First, if the discriminator makes an incorrect judgment, the generator does not receive accurate feedback, and the loss function cannot learn properly. Second, if the discriminator makes a very accurate judgment, the gradient of the loss function quickly converges to 0, resulting in a significant delay or disturbance to the learning speed. Because of these two problems, existing GANs are limited. WGANs compensate for these GAN shortcomings; in them, the KL divergence, which is used to define the loss function in existing GANs, is redefined using the Wasserstein distance (also referred to as the Earth mover's distance); this is an index that measures the distance between the two probability distributions. Under KL divergence, the distance value is 0 when the two distributions overlap each other, and it is infinite or constant when they do not overlap, showing an extreme distance value. The Wasserstein distance can be readily applied in training because a constant value is maintained regardless of whether the distributions overlap. Therefore, WGANs redefine the loss function using this Wasserstein distance, to smoothly train and improve the data such that it resembles the existing data as much as possible. Therefore, in this study, oversampling was performed using the WGAN loss function within a general GAN model and inputting the minority class of the original data. The structure of the GAN-based data oversampling model is as shown in Fig. 3. Although it has an identical structure to the general GAN, the potential problems of the existing GAN have been resolved by applying the WGAN theory and loss function. For each epoch, a random noise z is fed into the generator to generate fake data, and the fake data are merged with the abnormal transaction data (the minority class) from the original dataset. The random noise is expressed as a vector of the size to be generated, and the combined data are input to the discriminator, which attempts to distinguish the original data from the fake data (generated by the generator) and classify them as either real (1) or fake (0). Using the discriminator's classification results, the generator applies loss function to minimize the classification probability and the discriminator seeks to maximize it. The loss function is expressed as.

$$\nabla_{\omega} \frac{1}{m} \sum_{1=1}^{m} \left[f\left(x^{(i)}\right) - f\left(G\left(z^{(i)}\right)\right) \right] \quad (1)$$

$$\nabla_{\theta} \frac{1}{m} \sum_{1=1}^{m} \left[f\left(G\left(z^{(i)}\right)\right) \right] \quad (2)$$

and (2) are the loss functions applied to the discriminator and generator, respectively. Above, ω is the parameter of the discriminator, and ∇ω is the gradient descent for ω. Also, θ is the parameter of the generator and ∇_{θ} is the gradient descent for θ. x is the original data, z is the random noise and G is the generator. These loss functions differ from that of existing GANs, and the purpose of the discriminator also differs therefrom. Instead of using a direct criterion for identifying the fake data generated by the generator, the discriminator learns the K-Lipschitz continuous function, which is used to calculate the Wasserstein distance. In this process, as the loss function decreases, the Wasserstein distance becomes smaller and the fake data generated by the generator approach the actual data distribution [22].

For oversampling, the WGAN loss function was applied in a GAN. Only the data in the minority classes were selected and input to the model; the random noise followed the distribution of the input data through the interaction of the generator and discriminator. Finally, when the probability of distinguishing between the input and generated data converged to 0.5, the model was terminated, and the generated data combined with input data to resolve the original data imbalance. The proportions of generated data and random noise were determined by adjusting the ratio according to the quantity of original data. For the data oversampled through SMOTE, the amount of minority class data was determined according to the sampling strategy of the original data. If the sampling strategy was 1, the [minority class: majority class] ratio became [1:1]; if the sampling strategy was 0.5, it became [0.5:1]. Therefore, to generate GAN oversampling results similar to the data processed through SMOTE, the amount of random noise z was set to (0.3, 0.5, 0.7, 1) times the size of the majority class.

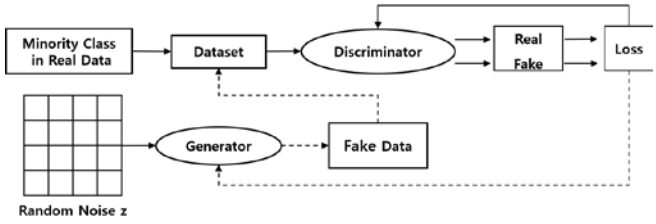


Fig. 3. Architecture for GAN-based Data Augmentation.

D. Domain Adaptation for Fraud Detection

Detecting abnormal transactions is a time-consuming and expensive process when using different models for two datasets of similar domains. Therefore, to develop a single model capable of detecting abnormal transactions from two datasets, we applied a domain-adaptation method which employs the discriminative characteristics of GANs, such as those used in ADDA [34]. While the ADDA was applied to image datasets, the proposed domain-adaptation method was applied to text datasets. Also in our study, the text datasets were augmented to avoid class imbalance problems. The domain-adaptation model used in this study was composed of source and target encoders that employed CNNs as shown in Fig. 4 and 5, respectively. Each encoder consisted of a 1D convolution layer (Conv1d), max pooling, and a fully connected layer. The convolution layer was used because it can readily extract feature maps and does not require any further layer (e.g., recurrent neural networks) for time-independent datasets. In addition, a CNN was used because these networks outperform ANNs in terms of time and performance efficiency. Two convolutional layers and two max pooling layers were used to prevent unsmooth learning or overfitting from occurring when adjusting the hyperparameters to match the feature maps. The model first learned a source encoder and classifier using the credit card fraud dataset (source domain). The loss function applied to the source encoder is expressed as follows:

$$\min L_C\left(C\left(f_s(X_S)\right), Y_S\right) \tag{3}$$

Here, C is the classifier, f_S is the source encoder, X_S is the credit card dataset, and Y_S is the credit card dataset class. Next, the financial transaction fraud dataset (target domain) was input to the CNN-based target encoder. The learning proceeded by labeling the output of the target encoder as 1 and inputting it to the discriminator. Expressed otherwise, when the discriminator receives the output of the target encoder, the learning proceeds in the direction in which the result value becomes 1. The target encoder's loss function is expressed as

$$\min L_t\left(D\left(f_t(X_t)\right), 1\right) \tag{4}$$

where D is the identifier, f_t is the target encoder, and X_t is the financial transaction dataset. The discriminator learns the distribution by labeling the output value of the source encoder as 1 (real) and the output value of the target encoder as 0 (fake), to properly distinguish between normal and fraudulent data; then, it applies a loss function. The loss function applied to the discriminator is expressed as follows:

$$\min L_D\left(D\left(f_s(X_S)\right), 1\right)$$
$$\min L_D\left(D\left(f_t(X_t)\right), 0\right) \tag{5}$$

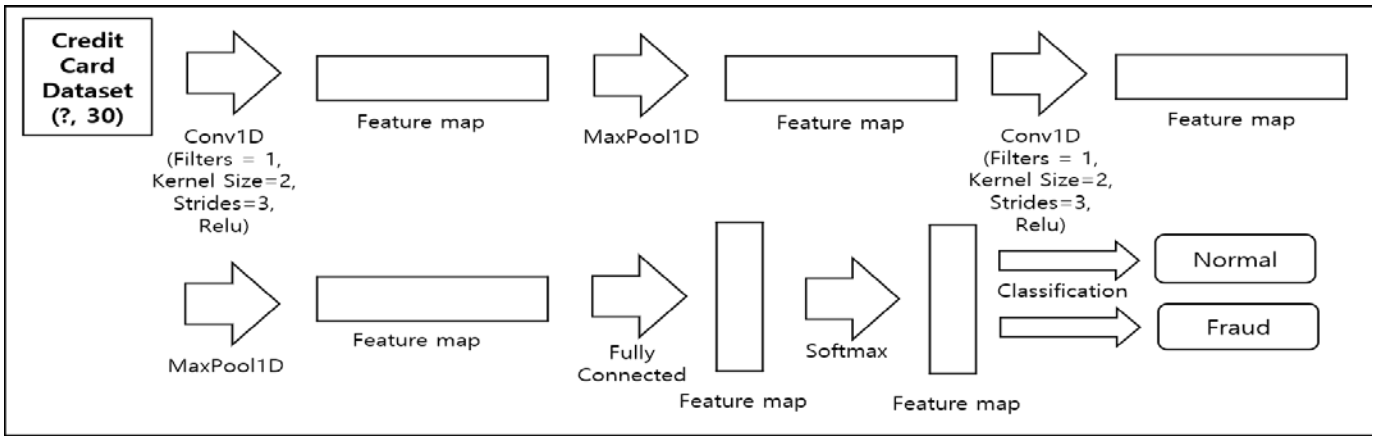


Fig. 4. Configuration of Source Encoder with Classifier.

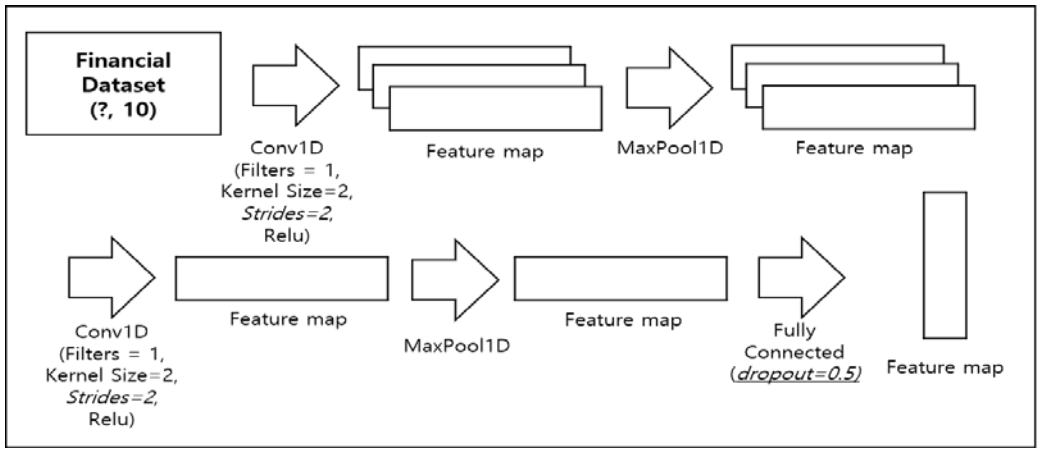


Fig. 5. Configuration of Target Encoder.

The entire learning process optimizes the loss functions described above, operating in a stepwise fashion. Based on the credit card fraud dataset (including the class information), the source encoder and classifier learn first, followed by the target encoder and discriminator. The source encoder proceeds in a fixed state whilst the target encoder and discriminator are being trained; thus, the target encoder's and discriminator's learning can proceed smoothly, without checking the state of the source encoder and classifier. Fig. 6 illustrates the overall structure of the domain-adaptation model introduced in this study; the components denoted with solid lines indicate a state in which learning is completed, and components formed of dotted lines indicate that learning takes place. Thus, the entire test process is as follows. First, the source encoder and classifier are trained on the source domain, and the discriminator and target encoder are trained from the source encoder and target domain. Finally, the proposed domain-adaptation model terminates the process when the target and source encoder can completely derive the classification results of the target and source domains, respectively.

E. Evaluation

The test results were evaluated using the area-under-curve (AUC) score, which is a classification-model performance evaluation index. The receiver operating characteristic (ROC) curve is a performance measure commonly used in binary classification and medical applications. Table I shows the confusion matrix; here, True (T)/False (F) indicates that the predicted value is the same/differs from the actual value, and Positive (P)/Negative (N) indicates how the predicted value was obtained. The ratio between the true-positive rate (TPR) and false-positive rate (FPR) is expressed as a graph of the ROC curve, and the AUC score is the area underneath this curve. The AUC score of a model with 100% incorrect prediction is expressed as 0.0, and the AUC score of a model with 100% correct predictions is expressed as 1.0; the performances of the models used in this study were evaluated accordingly.

TABLE. I. PARAMETERS IN ROC CURVE BY CONFUSION MATRIX

	Normal Prediction	Fraudulent Prediction
Normal Transaction	TN	FP
Fraudulent Transaction	FN	TP
$TPR = \frac{TP}{TP + FN}$		
$FPR = \frac{FP}{FP + TN}$		

IV. EXPERIMENTS

To evaluate the classification performance of the proposed domain-adaptation model, an SVM and CNN were employed as comparison machine and deep learning methods, respectively. Among machine learning methods, SVM has received particular attention for their excellent performance. It is a supervised learning model mainly used for pattern recognition and data analysis (in particular, classification and regression). Here, because both credit card and financial transaction datasets have class labels, SVM was used to detect abnormal transactions. The kernel of SVM uses a radial basis function. After testing values from 1 to 10,000, the hyperparameter C was set as 1000, which was found to deliver the optimal time and accuracy performances. The compositions of the source and target encoders in the proposed domain-adaptation model are as shown in Figs. 4 and 5. The source encoder sets the filter, kernel size, strides, and activation function, as shown in Fig. 4; the feature map (which undergoes max pooling after the CNN layer) passes through the fully connected layer. The output of the fully connected layer is passed to the classifier, to derive the classification result. The target encoder sets the number of strides to 2, to derive an output value with the same shape as the output value of the source encoder; other parameters (i.e., filter, kernel size, and activation function) are set identically to those of the source encoder. In addition, to prevent overfitting, a dropout was applied to the fully connected layer, with a ratio of 0.5.

The loss function of the classifier was calculated from the softmax cross-entropy, and the loss function of the discriminator was calculated using the sigmoid binary cross-entropy and optimized through the Adam optimizer (learning rate = 0.0001, beta 1 = 0.5, beta 2 = 0.99). The CNN model used the source encoder, target encoder, and classifier of the domain adaptation model. The credit card fraud data were used as the input data of the source encoder, and the financial transaction fraud dataset was used as the input data of the target encoder, to compare the classification results. The number of nodes of the hidden layer used in the GAN-based oversampling method was set to 128, the epoch was set to 20, and the Adam optimizer was set identically to the domain-adaptation model. The random noise was set as a random number extracted from a uniform distribution within the range [-1, 1]. The Ubuntu 18.04.4 LTS test environment consisted of an Intel(R) Xeon CPU E5-2620 v4 with a 2.10 GHz CPU, GTX 1080 GPU, and 64 GB RAM.

V. RESULTS AND ANALYSIS

Table II compares the classification performance results of the SVM, CNN, and proposed domain-adaption models. The experiment was conducted, and the results of the classification performance were averaged by summing only values above 0.8; this expresses the ratio between the majority and minority class when augmenting or oversampling a dataset. In other words, if the majority class is 1, a quantity of data equal to the ratio is generated to oversample the minority class.

Table III shows the time taken for each model to receive data, train it, and derive its classification results. Table IV shows the time taken to oversample each dataset with GAN and SMOTE, respectively. Fig. 7 compares the performances of the GAN- and SMOTE-based oversampling methods. The left-hand and right-hand graphs describe results for the credit card and financial transaction fraud datasets, respectively; the x-axis denotes the ratio mentioned in Table II. The AUC scores on the y-axis represent the averaged classification performances for all methods; the GAN-based oversampling method takes slightly longer than SMOTE to complete, but it exhibits a superior performance (as shown in Fig. 7). The left-hand graph in Fig. 8 shows the average classification performance for the dataset in which the GAN-based oversampling method was applied. The right-hand graph shows the time-averaged values of the GAN-based oversampling method in Table III. In Fig. 8, although the

classification performance of the domain-adaption model was inferior to those of the CNN and SVM, it was found to be suitable as an abnormal transaction detection model for both test domain datasets, because it reduced the required learning time when performing abnormal transaction detection on two datasets with similar domains. The SVM outperformed the CNN and domain-adaption models; however, it is not readily applicable to larger datasets, because its learning time increases sharply when the dataset increases. Compared to the domain-adaption model, the CNN model shows no significant difference in classification performance; however, because it requires more learning time, it is limited as a classification model for different domains.

TABLE. II. CLASSIFICATION RESULTS OF THE SVM, CNN, AND PROPOSED DOMAIN-ADAPTION MODELS IN AUC SCORE

Dataset	Method	Oversampling using GAN				Oversampling using SMOTE			
	Ratio	0.3	0.5	0.7	1	0.3	0.5	0.7	1
Credit Card	SVM	0.9984	0.9989	0.9994	0.9996	0.9639	0.9648	0.9662	0.9723
	CNN	0.9844	0.9895	0.9903	0.9862	0.9073	0.9261	0.915	0.899
	Domain Adaptation	0.9842	0.9889	0.9910	0.9888	0.9067	0.9257	0.921	0.9011
Financial	SVM	0.9986	0.9990	0.9989	0.9996	0.9701	0.9726	0.9754	0.9801
	CNN	0.88	0.8973	0.8988	0.9284	0.861	0.864	0.8721	0.9078
	Domain Adaptation	0.8821	0.8967	0.8927	0.9235	0.868	0.8701	0.8858	0.9125

TABLE. III. TIME IN SECONDS TAKEN FOR EACH MODEL TO RECEIVE DATA, TRAIN IT, AND DERIVE ITS CLASSIFICATION RESULTS

Dataset	Method	Oversampling using GAN				Oversampling using SMOTE			
	Ratio	0.3	0.5	0.7	1	0.3	0.5	0.7	1
Credit Card	SVM	198	215	257	261	185	229	265	311
	CNN	145	159	187	203	151	161	199	238
	Domain Adaptation	143	161	184	208	150	158	201	231
Financial	SVM	23940	24146	26756	28869	25442	26541	29287	30218
	CNN	1345	1973	2329	2975	1421	1898	2423	3033
	Domain Adaptation	456	657	823	1206	445	558	901	1158

TABLE. IV. DATA AUGMENTATION PROCESSING TIME IN SECONDS WITH GAN AND SMOTE

	Oversampling using GAN				Oversampling using SMOTE			
Ratio	0.3	0.5	0.7	1	0.3	0.5	0.7	1
Credit Card	16	28	43	63	0.61	0.72	0.82	1.01
Financial	303	492	714	1204	5.67	6.10	6.91	10.2

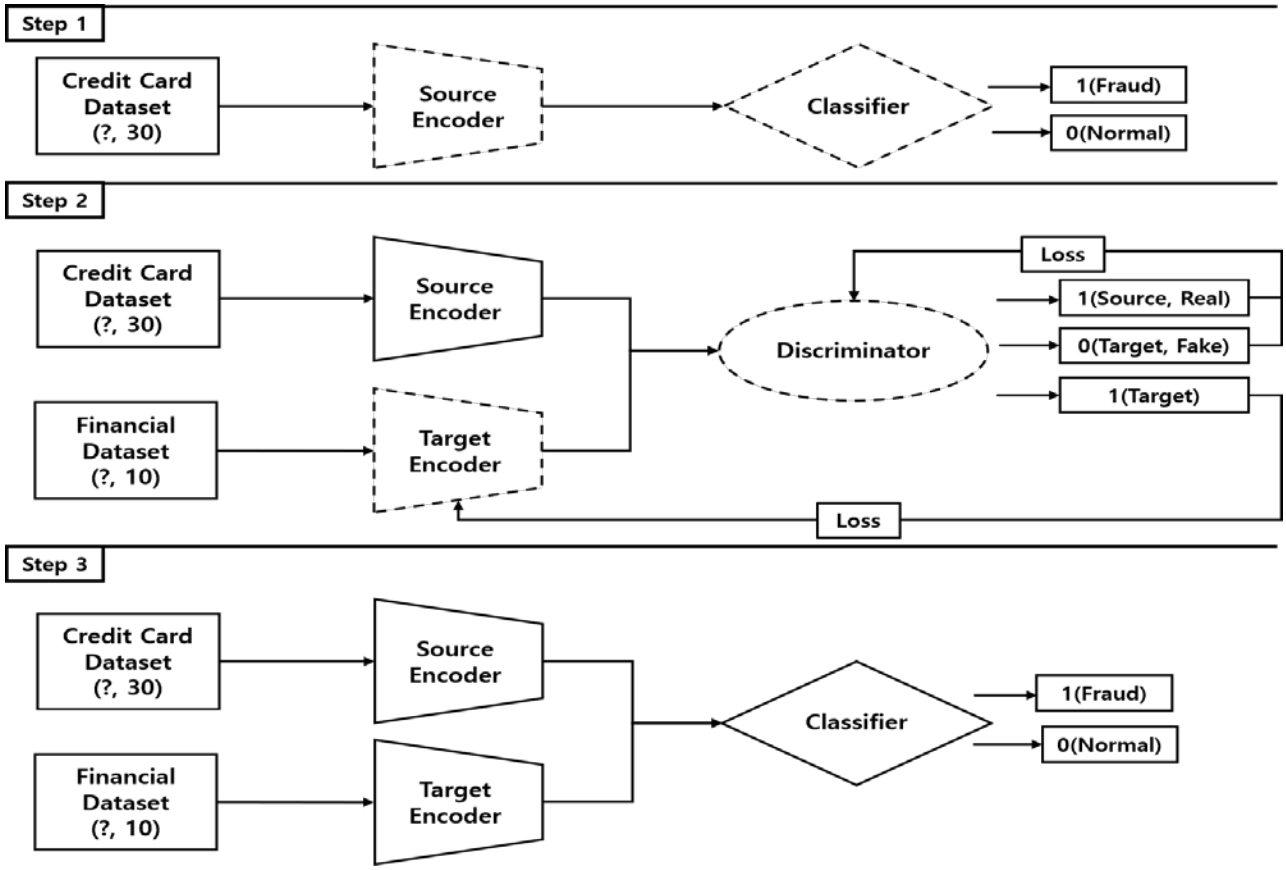


Fig. 6. Architecture of Proposed Domain-Adaptation Method for Fraud Detection.

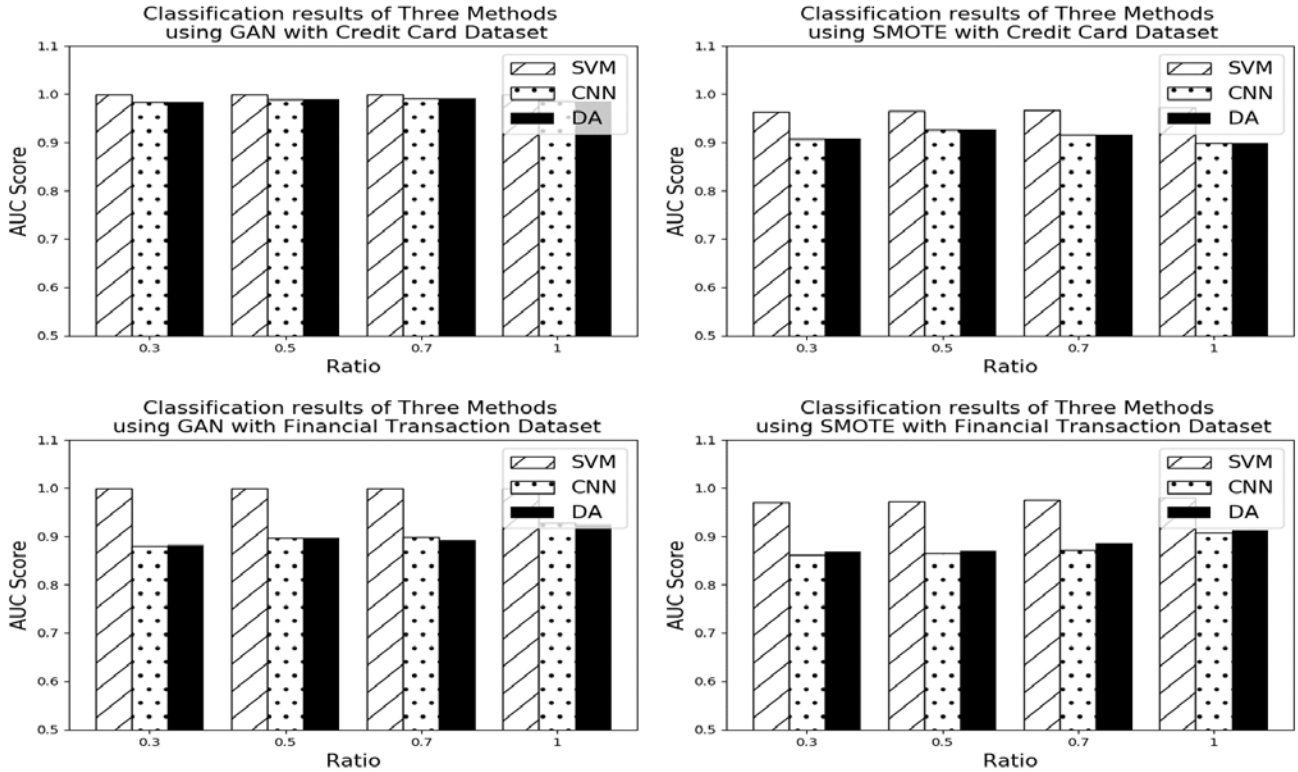


Fig. 7. AUC Scores for Datasets Augmented by GAN and SMOTE.

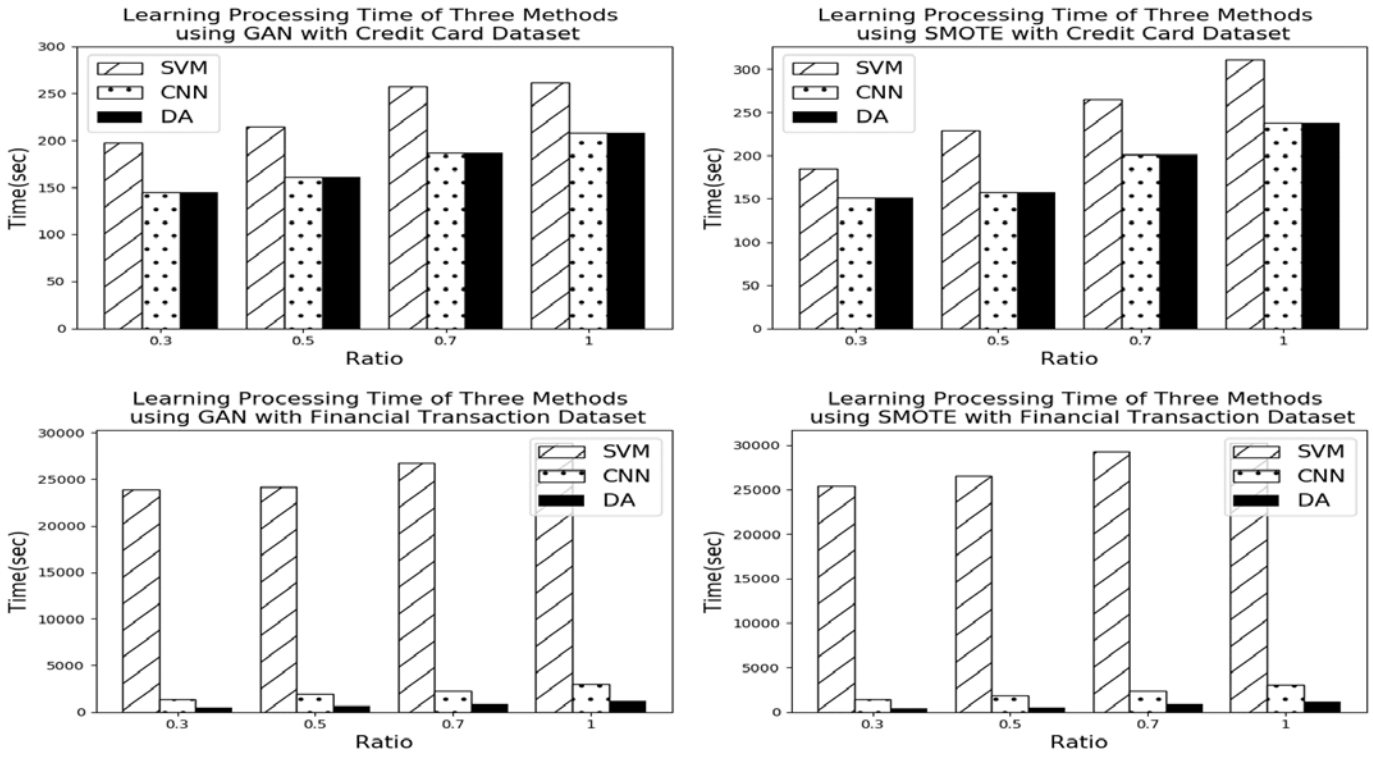


Fig. 8. Learning Processing Time for Datasets Augmented by GAN and SMOTE.

VI. Conclusions

In this study, a domain-adaptation method, applicable to data in similar domains, was proposed. The model to which the proposed domain-adaptation method was applied has the advantage of minimizing domain shifts when the domains are similar, even if the dataset has changed. In the experiments, credit card and financial transaction fraud datasets were used to evaluate the model's performance. Both datasets had a class imbalance problem; thus, oversampling was conducted using GAN and SMOTE; then, these data were used as input data of the model. Moreover, a classification performance comparison was made against SVM and CNN, to evaluate the model's performance. As a result, though the proposed domain adaption model did not achieve a better classification performance than the SVM or CNN, its performance was comparable thereto, while requiring a shorter learning time. Moreover, the GAN-based oversampling method, which was used to solve the class imbalance problem, outperformed SMOTE. Although the CNN showed a similar classification performance to the domain-adaption model, it required a longer learning time. The SVM had a high classification performance; however, it required a comparatively longer learning time than the CNN when the dataset size was increased. As a result, the proposed domain-adaptation model was shown to be capable of simultaneously classifying two datasets with similar domains and shortening the learning time compared to the SVM and CNN. However, there are several limitations to this study, which should be addressed in the future: both datasets were constructed using CNN models, to smoothly reuse the feature maps; the classification performance was insufficient compared to that of the SVM; and various domain data and results were absent. Therefore, in future research, structural changes will be made to the oversampling method proposed in this study, to make use of the various abnormal transaction data (including time-series data) and judge the performance of the model more objectively.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT: Ministry of Science and ICT) (No. NRF-2019R1F1A1059036).

References

[1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A Theory of Learning from Different Domains", Machine Learning, vol. 79, pp. 151-175, 2010, https://doi.org/10.1007/s10994-009-5152-4.

[2] L. Y. Pratt, "Discriminability-Based Transfer between Neural Networks", Advances in Neural Information Processing Systems, vol. 5, pp. 204-211, 1992, https://doi.org/10.5555/645753.668046.

[3] Machine Learning Group ULB. Credit Card Fraud Detection, 2017, https://www.kaggle.com/mlg-ulb/creditcardfraud.

[4] D. A. Lopez-Rojas, A. Elmir, and S. Axelsson, "PaySim: A Financial Mobile Money Simulator for Fraud Detection", 28th European Modeling and Simulation Symposium (EMSS 2016), vol. 28, pp. 249-255, 2016, https://doi.org/1016/j.ecolmodel.2006.04.

[5] G. Ian, P-A. Jean, M. Mehdi, X. Bing, W-F. David, O. Sherjil, C. Aaron, and B. Yoshua, "Generative Adversarial Nets", NIPS: Advances in Neural Information Processing Systems, vol. 2, pp. 2672-2680, 2014, doi/10.5555/2969033.2969125.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002, https://doi.org/10.1613/jair.953.

[7] C. Cortes and V. Vapnik, “Support-Vector Networks”, Machine Learning, vol.20,pp.273-297, 1995, https://doi.org/10.1007/BF00994018.

[8] Y. LeCun, B. Bose, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L D. Jackel, “Backpropagation Applied to Handwritten Zip Code Recognition”, Neural Computation, vol. 1, no. 4, pp. 541-551, 1989, https://doi.org/10.1162/neco.1989.1.4.541.

[9] M. Tatsuya and N. Ayahiko, “Proposal of Credit Card Fraudulent Use Detection by Online Type Decision Tree Construction and Verification of Generality”, International Journal for Information Security Research, vol.1, pp. 229–235, 2013, https://doi.org/10.20533/ijisr.2042.4639.2013.0028.

[10] M. Singh, S. Kumar, and T. Garg, “Credit Card Fraud Detection Using Hidden Markov Model”, International Journal of Engineering and Computer Science, vol. 8, pp. 24878-24882, 2019, https://doi.org/10.18535/ijecs/v8i11.4386.

[11] N. Malini and M.Pushpa, “Analysis on Credit Card Fraud Identification Techniques Based on KNN and Outlier Detection”, International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017, https://doi.org/10.1109/AEEICB.2017.7972424.

[12] N. K. Gyamfi and J-D. Abdulai, “Bank Fraud Detection Using Support Vector Machine”, IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2018, https://doi.org/10.1109/IEMCON.2018.8614994.

[13] M. A. Al-Shabi, “Credit Card Fraud Detection Using Autoencoder Model in Unbalanced Datasets”, Journal of Advances in Mathematics and Computer Science, vol. 33, pp. 1-16, 2019, https://doi.org/10.9734/jamcs/2019/v33i530192.

[14] U. Fiore, A. D. Santis, F. Perla, P. Zanetti, and F. Palmieri, “Using Generative Adversarial Networks for Improving Classification Effectiveness in Credit Card Fraud Detection”, Information Sciences, vol. 479, pp. 448-455, 2019, https://doi.org/10/1016/j.ins.2017.12.030.

[15] H. He and E. A. Garcia, “Learning from Imbalanced Data”, IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, 2009, https://doi.org/10.1109/TKDE.2008.239.

[16] S. Barua, M. M. Islam, X. Yao, and K. Murase, “MWMOTE: Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning”, IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 2, pp. 405-425, 2012, https://doi.org/10.1109/TKDE.2012.232.

[17] G. Menardi and N. Torelli, “Rose: Random Over-sampling Examples”, Data Mining and Knowledge Discovery, vol. 28, no. 1, pp. 92-122, 2014, https://doi.org/10.1080/24699322.2019.1649074.

[18] B. Baesens, S. Höppner, I. Ortner, and T. Verdonck, “robROSE : A Robust Approach for Dealing with Imbalanced Data in Fraud Detection”, 2020, ariv:2003.11915v1.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPPS), vol. 1, pp. 1097-1105, 2012, https://doi.org/10.1145/3065386.

[20] T. Karras, T. Alia, S. Laine, and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation”, Computing Research Repository (CoRR), vol. abs/1710.10196, 2017, http://arxiv.org/abs/1710.10196.

[21] A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”, Proceedings of the 25th International Conference Learning Representations (ICLR), pp. 1-16, 2016, https://arxiv.org/abs/1511.06434.

[22] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN”, ICML'17: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 214-223, 2017, https://arxiv.org/abs/1701.07875.

[23] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “Synthetic Data Augmentation using GAN for Improved Liver Lesion Classiffication”, IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), vol. 15, 2018, https://doi.org/10.1109/ISBI.2018.8363576.

[24] J. T. Springenberg, “Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks”, 2015, https://arxiv.org/abs/1511.06390.

[25] F. H. K. d. S. Tanaka, and C. Aranha, “Data Augmentation Using GANs”, Computing Research Repository (CoRR), 2019, http://arxiv.org/abs/1904.09135.

[26] R. Caruana, “Multitask Learning,” Machine Learning, vol. 28, pp. 41-75, 1997, https://doi.org/10.1023/A:1007379606734.

[27] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, “Self-Taught Learning: Transfer Learning from Unlabeled Data”, Proceedings of the 24th International Conference on Machine Learning(ICML), pp. 759–766, 2007, https://doi.org/10.1145/1273496.1273592.

[28] H. Daume III, and D. Marcu, “Domain Adaptation for Statistical Classifiers”, Journal of Artificial Intelligence Research, vol. 26, pp. 101-126, 2016, https://doi.org/10.1613/jair.1872.

[29] B. Sun, and K. Saenko, “Deep CORAL: Correlation Alignment for Deep Domain Adaptation”, European Conference on Computer Vision, pp. 443-450, 2016, http://arxiv.org/abs/1607.01719.

[30] Y. Ganin, and V. Lempitsky, “Unsupervised Domain Adaptation by Backpropagation”, Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pp. 1180-1189, 2015, https://doi.org/10.5555/3045118.3045244.

[31] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, “Simultaneous Deep Transfer Across Domains and Tasks”, IEEE International Conference on Computer Vision (ICCV), 2015, https://doi.org/10.1109/ICCV.2015.463.

[32] A. Gretton, A. J. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf, “Covariate Shift and Local Learning by Distribution Matching”, Dataset Shift in Machine Learning (in Book), pp. 131-160, Dec. 2008, https://doi.org/10.7551/mitpress/9780262170055.003.0008.

[33] X. Xu, X. Zhou, R. Venkatesan, G. Swaminathan, and O. Majumder, “d-SNE: Domain Adaptation Using Stochastic Neighborhood Embedding”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2497-2505, 2019, https://doi.org/10.1109/CVPR.2019.00260.

[34] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial Discriminative Domain Adaptation”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2017, https://doi.org/10.1109/CVPR.2017.316.

[35] G. Lemaitre, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning”, Journal of Machine Learning Research, vol. 18, pp. 1-5, 2017, https://arxiv.org/abs/1609.06570.

Evaluation of Student Core Drives on e-Learning during the Covid-19 with Octalysis Gamification Framework

Fitri Marisa[1]
Department of Informatic Engineering, Faculty of Information and Communication Technology
Widyagama University of Malang
Universiti Teknikal Malaysia Melaka
Malang, Indonesia, Melaka, Malaysia

Sharifah Sakinah Syed Ahmad[2]
Zeratul Izzah Mohd Yusoh[3]
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

Anastasia L Maukar[4]
Industrial Engineering Department
President University, Jakarta, Indonesia

Ronald David Marcus[5]
Faculty of Information Technology
University of Merdeka Malang, Malang, Indonesia

Anang Aris Widodo[6]
Faculty of Information Technology
Merdeka University of Pasuruan, Pasuruan, Malang

***Abstract*—Learning activities during the Covid-19 pandemic were carried out with an online system even though in reality many institutions had not prepared their systems and infrastructure properly. Some e-learning media that are generally used based on survey results include 53.81% google classrooms combined with other applications that are not integrated with the institution's Learning Management System. This condition provides research opportunities to evaluate the effectiveness of online learning, especially how students are motivated to learn the method, where the results can be used as a reference in developing and refining the method. Based on many studies, that the gamification model can increase individual motivation in carrying out activities, this study uses a gamification octalysis framework to analyze the extent of the role of gamification in the learning process and measure the amount of student motivation in online learning activities. The evaluation results show that the conclusion of the Likert scale results in a "High" level, while the highest score is "Very High". As for the octalysis test scale, the average score of 6.5 on a scale of 1 to 10. The conclusion from the results of this evaluation is that the motivation to learn e-learning during the Covid-19 period is quite high and has the potential to be developed. While the results of the Octalysis framework with 8 core drives are still average, for that we need innovation in E-learning which aims to increase student motivation based on Octalysis's 8 core drives. The results of this study recommend that gamification is needed to increase student learning motivation in order to improve learning outcomes.**

***Keywords*—Gamificaton; education; Covid-19 pandemic; octalysis framework**

I. Introduction

During the Covid-19 pandemic, the learning process in all countries was carried out using an online system [1],[2],[3]. At least, some fundamental problems affect the success of online learning, especially in the Covid-19 pandemic, one of them is the problem of limited infrastructure [2],[1]. However, with this emergency, many institutions have not prepared their systems and infrastructure properly, therefore, it is important to pay attention to this problem. [2],[1].

In conditions of limited infrastructure, the learning process is still carried out by utilizing available online learning technologies [4] on the internet such as Google classroom, moodle, zoom meeting, and others which are mostly not integrated with each other. From the results of a survey of 419 student respondents at several universities in Indonesia in August 2020, the data presented in Table I and Fig. 1. The highest number of categories is Google classroom mixed with other applications (53.81%), for other media, was 25.71%. Meanwhile, only 20.48% use e-learning applications, and this data represents the percentage of the availability of an integrated e-learning system provided by the institution (Learning Management System) [2],[1]. It can be inferred that the infrastructure which supports e-learning still needs improving.

The second problem is the decline in students' mental health during the lockdown period which has an impact on the low motivation of students towards the online learning system during Covid-19 pandemic. [5], [6], [7], [8]. However, this pandemic condition cannot be avoided, therefore efforts to increase user acceptance and motivation of online learning are also important [3]. Several studies [9] that observed the acceptance and motivation of teachers and students towards online learning have been carried out, such as [3] has observed student acceptance of online learning during a pandemic with observed parameters including attitude, influence, motivation, behavior control, and cognitive engagement [3]. The results of the analysis stated that the students preferred face-to-face learning [3].

TABLE I. PERCENTAGE OF E-LEARNING APPLICATION USED (BASED ON SURVEI ON AUGUST, 2020)

Media	%
Google Classroom and others	53.81
e-Learning Application	20.48
Others	25.71

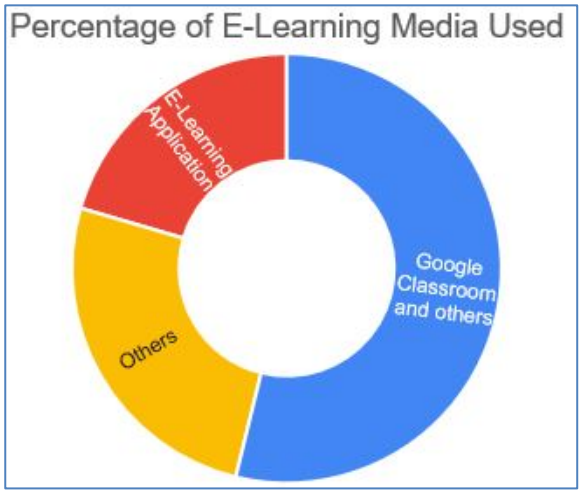


Fig. 1. Graph of Percentage of e-Learning media used (Based of Survei on August, 2020).

However, so far there has been no study that has a deeper evaluation of how students accept and perceive online learning without having to compare it to traditional learning. This is important to do considering that online learning is currently the only option in organizing learning during the Covid-19 pandemic. For this reason, efforts are needed to support the online learning system to get recommendations that can improve their performance. Although the development of online learning features in the future, it is necessary to develop knowledge pattern extraction and extraction [10] and a recommendation system [11],[12] from the resulting data, the extent to which users accept online learning itself is more fundamental. Exploring and providing recommendations about student perceptions is useful for re-evaluating online learning systems in the hope of knowing which parts should be improved or which should be maintained. This condition provides research opportunities to evaluate the effectiveness of online learning during this covid-19 pandemic. One of them is by proposing the concept of how to increase student learning motivation towards the method used, therefore that the results can be used as a reference for developing and improving existing methods. Based on many studies, gamification models can support increased motivation, and also measure it.[13],[14].

Gamification is one of the fields of science that comes from computer science became to multidiscipline science that adopts a game mechanism to be applied to a non-game system that aims to increase student motivation [15],[16]. In the field of education, gamification is also developed in a learning system that aims to increase student participation and motivation [13],[14]. Several gamification frameworks have been developed including MDA (Mechanics Dynamics Aesthetics), MDE (Emotional Mechanics Dynamics), and one of them is Octalysis [17].

The Octalysis framework was developed with the idea that each individual action is motivated by core drives that influence its action [18], [19]. Octalysis divides core drives into 8 types which are grouped into two groups, namely, the right brain (creativity) and the left brain (analytical) and White hat (top) and Black hat (bottom) [18], [19], [17]. With this potential, the Octalysis framework can measure and evaluate students' core drives against the online learning system that has been undertaken. The results of the further evaluation can provide recommendations to institutions in developing online learning to make it more optimal. This research contributes to the design method, analysis, and evaluation of the actual condition of the core drive experienced by users on a running system, especially in e-learning.

II. LITERATURE REVIEW

A. e-Learning during Covid-19 Pandemic

In this Covid-19 pandemic, online learning is the only solution for the learning process in all institution around the world [20], [3], [7], [1],[2],[3],[21]. The most emphasized thing on the characteristics of online learning is towards student independence and control is with students [20], for that effort related to it is important to do. However, common problems experienced in online learning today are limited infrastructure [1],[2], limited human resource capabilities [20], and low motivation [5], [6], [7], [8].

B. Gamification

Gamification is currently one of the technological trends that originated in the computer science and developed into a multidisciplinary science, by adopting game mechanisms to solve non-game system. [17], [16],[15]. Meanwhile, the focus that gamification wants to solve is increasing user acceptance, motivation and participation for the system being implemented [17], [16],[15]. Gamification has several frameworks where MDA (Mechanics, Dynamics, Aesthetics) is the basis of all existing frameworks in gamification. [22].

Fig. 2 described MDA framework, where Mechanics (M) are a set of rules / algorithms that cause a player to perform an action in a gamified system, and generally Mechanics are in the form of components such as levels, ratings, etc. [23], [22], [24]. Dynamics (D) is the movement of players resulting from the interaction process between players and mechanics. As an illustration, it can be said that if the points are mechanics, then the results of the points that the player gets after carrying out the rules set by the system are dynamics [23], [22], [24]. Aesthetics (A) is a component that deals with psychological aspects such as how players respond to game dynamics. Each player has a different response and Aesthetics is abstract. The MDE Aesthetics framework is also termed Emotional [23], [22], [24].

C. Octalysis Gamification Framework

The Gamification Octalysis framework was first invented by Yu-kai Chou, who based 8 core drives as a background or individual cause for action. [17],[18], [19]. Octalysis is the process of applying the core behavior drives that motivate a

Sponsored by: Ministry of Research and Technology of the Republic of Indonesia RISTEK- BRIN.

user to complete a task efficiently through an interactive experience. The Octalysis approach believes that no action will occur if there is no drive before [18], [19]. 8 core drives Octalysis is divided into 2 types of division, the first is right brain (creativity, expression) and left brain (analytical thinking) [17],[18],[19]. And the second group is grouped into 2 groups up (white hat) and bottom (black hat). White hat plays a role in positive motivation and brings a sense of meaning and a greater sense of control, while black hat acts as a negative impulse, but can inspire balance to achieve maximum results (Fig. 3) [17],[18], [19].

The description of the 8 core octalysis drives (Fig. 3) is as follows [18],[25]:

1) Epic meaning: Drive that plays a role in motivating to do work on the basis of the interests of the crowd above personal interests. A person is motivated to take action on the basis of being motivated that his actions will participate in completing a mission that is bigger than his own interests, and usually involves the interests of many people.

2) Development and accomplishment: Drive to take action on the basis of wanting to achieve an achievement. In general, someone takes an action because he hopes for a reward or something that he will have later. This drive is the easiest to expand compared to other 7 drives. Some of the mechanics that represent this drive include points, badges, and levels.

3) Empowerment of creativity and feedback: Drive for the creative process or they are looking for new things. Someone is motivated to take action for reasons of liking the freedom to create the object at hand and is involved in creative rposes. One of the mechanics that are often used in this drive is Evergreen Combos.

4) Ownership and possesion: This drive works by motivating ownership of something and trying to improve it. This drive is based on the idea that the more someone feels they own an object, the more he wants to maintain and improve it. Mechanics that are usually used in this drive include Avatar and virtual good.

5) Social influence and relatedness: This drive is motivated from social interaction so as to influence individual actions. This drive is based on the idea that a person takes action because of several things such as wanting to compete with others, wanting to cooperate with others, being inspired by other people, or because an object is connected to past memories. Mechanics involved in this drive include Group Quest, and mentorship.

6) Scarcity and impatience: This drive is driven by the motivation to take action simply because it is extremely rare, exclusive, or not immediately achievable. Someone is attracted to an object because it is exclusive and hard to find. Mechanics that are usually involved in these drives include countdown timers and price pacing.

7) Unpredictability and curiosity: A motivated drive to take action because something is difficult to predict. In other words, this drive works by provoking one's curiosity about the object. The more curious someone is, the more he will try to keep up. Mechanics who are usually involved in this drive include Rolling Reward, Sudden Reward.

8) Loss and avoinance: Drive that is motivated to take action for fear of missing an opportunity. This drive is in line with the "Unpredictability and Curosity" drive but the difference is that if this drive focuses on anticipating bad things and failures then someone tries to defend the object. The mechanics involved include Progress Loss, Status Quo Slots.

D. Related Works

In recent years gamification has emerged as a framework that also plays a role in the development of learning systems. [26], [27]. Gamification is applied to develop various learning materials, including: language, computers, statistics, and others [28],[29],[30]. The application of gamification is generally applied in learning design as a strategy to make learning methods more attractive [30], therefore increase students knowledge retention [31]. In another study, gamification was also shown to increase achievement motivation[32] and increase material absorption [29]. In its development, gamification in the learning method is collaborated with Artificial Intelligence to build a user personalization system so that the idea is expected to better accommodate user needs and increase motivation. [26].

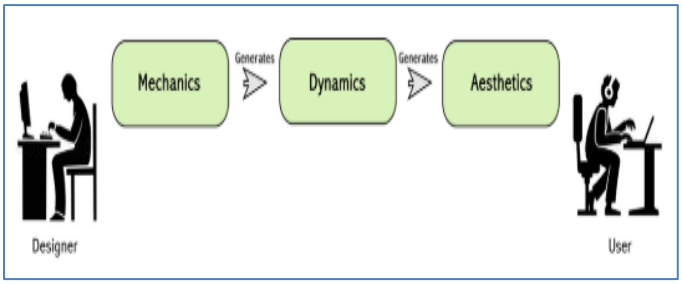


Fig. 2. MDA Framework [17].

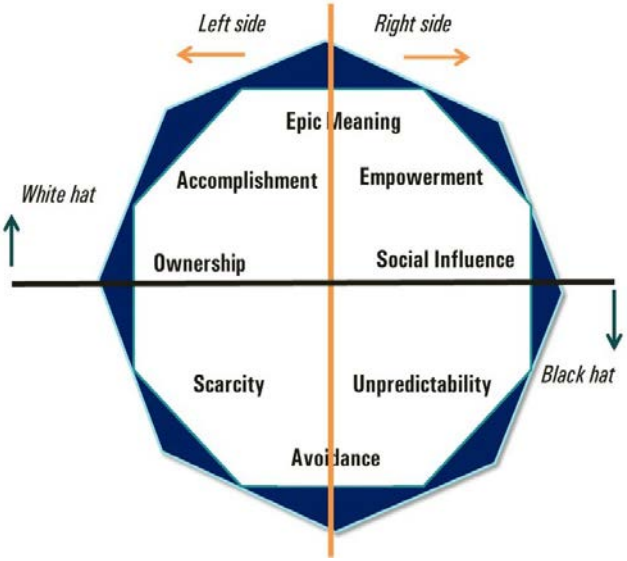


Fig. 3. Octalysis Framework [18].

III. Research Methodology

The research methodology described in Fig. 4.

1) Literature review: Literature studies are focused on collecting literature related to online learning, gamification theory in particular, the octalysis framework. The data were grouped using a questionnaire given to 419 student respondents in the provinces of East Java and West Java, Indonesia. The literature study also discusses the Likert scale and sample determination.

2) Determining of questionnaire: In this stage, the weight of the questionnaire is determined using a Likert scale [33] with four (4) criterias and making eight (8) questions, each of which represents 8 core drives.

3) Analysing each core drive: The stage of calculating each core drive with a Likert scale calculation and the results are categorized into four (4) levels of the range scale that have been determined based on the difference from the multiplication of the criteria weights and the number of respondents.

4) Analysing result measurement of likert scale test: In this stage, all the results of the Likert scale categorization of each core drive are filled in the table. This aims to find out and compare which core drives have a maximum or minimum value.

5) Converting and analysing result measurement of octalysis scale test: In this stage, the conversion process of the Likert scale calculation results of each core drive is converted into the octalysis test scale and the determination of the octagon pattern of the eight core drives that were tested. In this stage, an evaluation of the octagon pattern formed is also carried out compared to the octagon pattern which is the standard of achievement.

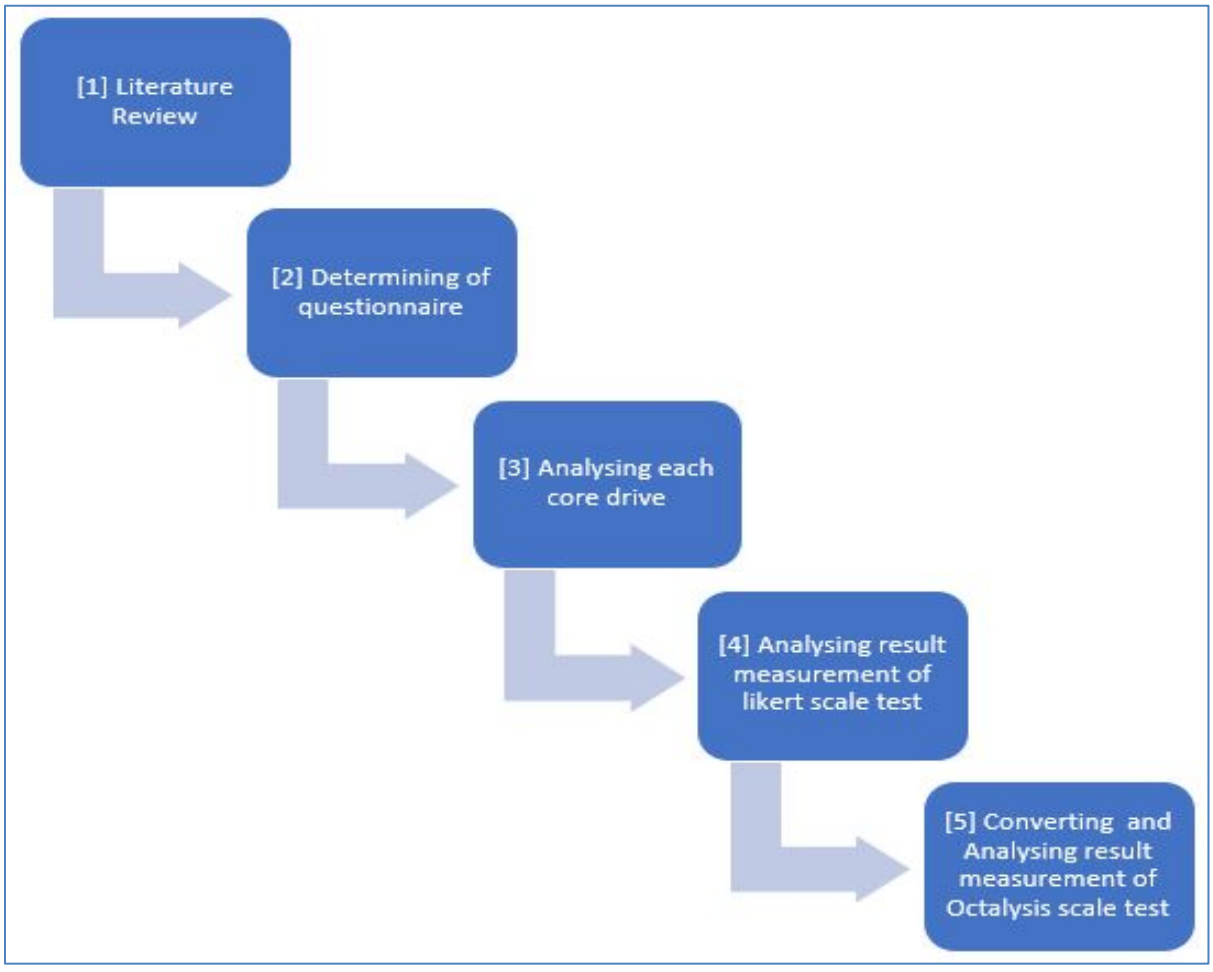


Fig. 4. Research Methodology.

V. Results and Discussion

A. Determining of Questionnaire

The analysis process begins by determining a questionnaire with eight questions representing each of the octalysis core drives. The questionnaire was filled with 419 respondents consisting of students spread across several campuses in the provinces of East Java and West Java, Indonesia. Questionnaires were distributed for two weeks in August 2020. The scale of measurement used a four-level Likert scale with the highest score of 4 and the lowest of 1 as in Table II.

Each core drives are grouped with a score multiplied by a weighted score. The results of all scores are totaled then grouped and given a predicate according to the range of the predicate score categories. The predicate for the category of score is determined based on the following calculations:

- Determine the lowest point of range, which is obtained from the multiplication of the number of respondents (N) with the lowest weight score (value = 1), where N x 1 (419 x 1 = 419).
- Determine the highest point of range, which is obtained from the multiplication of the number of respondents (N) with the highest weight score (value = 4), where N x 4 (419 x 4 = 1676).
- Calculate the difference in range with the formula for the highest range minus the lowest range (1676 - 419 = 1257).
- Calculating the range value by dividing the range difference with the highest weight value (1257: 4 = 314.25), rounded up to 314. This value is used as a benchmark for the distance between levels in the measurement scale.
- Determine the level of the measuring scale consisting of 4 levels using the range of distances between levels that have been previously calculated at 314. The resulting measuring scale levels are as follows:

  Very Low is 419 until 733,

  Low is 734 until 1047,

  High is 1048 until 1362,

  Very High is 1363 until 1677.

B. Analyzing Each Core Drive

In this stage, the calculation and analysis of each core drives is carried out using Likert scale testing. The following are the measurement results of each core drives:

- Core Drive-1: Epic Meaning and Calling.

The question:

"The existence of the e-learning system used gives students the opportunity to participate in maintaining the stability of the learning process during the Covid-19 period"

Table III described the total score on the core drive Epic Meaning and Calling is 1279, which is in the "high" category. This data has shown that students have a high enough awareness because they believe that taking good online learning during a pandemic is taking part in maintaining the stability of the learning process. This condition needs to be maintained, for example by providing a clear introduction to the objectives and benefits of following the material to be discussed.

- Core Drive-2: Development and Accomplishment.

The Question:

"The mechanism of the e-learning system that is used encourages students to improve their achievement"

Table IV described the total score for the Core Drive Development and Accomplishment is 1125, which is in the "high" category. This data has shown that students have a high enough awareness to achieve the targeted achievements. This condition needs to be maintained, for example by maintaining and increasing the reward of each achievement in material mastery.

- Core Drive-3: Empowerment of Creativity and Feedback.

The question:

"The existing e-learning system mechanism gives students the opportunity to freely choose the media to collect assignments or consult"

TABLE II. Measurement of Questionnaire

Value	Declaration
1	Very Weak
2	Weak
3	Strong
4	Very Strong

TABLE III. Result of Data Analyzing Core Drive-1 EPIC meaning and Calling

Answer	Number of Respondents	Value	Total
Very Strong	80	4	320
Strong	289	3	867
Weak	42	2	84
Very Weak	8	1	8
∑	419		1279

TABLE IV. The Result of Data Analyzing Core Drives-2 Development and Accomplishment

Answer	Number of Respondents	Value	Total
Very Strong	37	4	148
Strong	230	3	690
Weak	135	2	270
Very Weak	17	1	17
∑	419		1125

Table V described the total score on the core drive-3 Empowerment of Creativity and feedback is 1212, which is in the "high" category. This data has shown that students feel comfortable with online learning systems that provide flexibility in managing learning personalization. This condition needs to be maintained, for example by maintaining user personalization facilities on the online learning system.

- Core Drive-4: Ownership and Possession

The Question:

"The e-learning system mechanism used has user personalization facilities that motivate students to be responsible for the existence of their respective accounts".

Table VI described the total score on the core drive-4 ownership and possession is 1245, which is in the "high" category. This data has shown that students feel ownership and try to maintain ownership in response to user personalization imposed by online learning. This condition needs to be maintained, for example by improving user personalization facilities on online learning systems so that the desire to maintain and improve personalization increases.

- Core Drive-5: Social Influence and Relatedness.

The question:

"The mechanism of the e-learning system that is used provides encouragement to be more competitive or collaborate between friends"

Table VII described the total score on the Core Drive-5 social influence and relatedness is 1154, which is in the "high" category. This data has shown that students feel the benefits of socializing with friends to support their progress in achieving success in learning. Students feel a connection with friends in increasing the motivation to compete and cooperate so that it automatically increases learning motivation. This condition needs to be maintained, for example by improving facilities or group assignments or increasing the intensity of tasks that are competitive among friends.

- Core Drive-6: Scarcity and Impatience

The question:

"The mechanism of the e-learning system that is used sometimes provides challenges that are not mandatory, for example in the form of time-limited quizzes that cause curiosity to try."

Table VIII described the total score on the core drive-6 scarcity and impatience is 1154, which is in the high category. This data shows that students have a high interest in one of the assignment models imposed by the online learning system which is intermittent and unscheduled. This model encourages students to be motivated to participate and not want to miss opportunities. This condition needs to be maintained, for example by maintaining the assignment model.

- Core Drive-7: Unpredictability and Curiosity.

The question:

"The e-learning system mechanism used includes a formative quiz facility which is announced unscheduled, which has encouraged students to frequently check their respective accounts for these opportunities."

Table IX shows the Result of Data Analyzing Core Drive-7 unpredictability and Curiosity.

TABLE V. RESULT OF DATA ANALYZING CORE DRIVE-3 EMPOWERMENT OF CREATIVITY AND FEEDBACK

Answer	Number of Respondents	Value	Total
Very Strong	54	4	216
Strong	274	3	822
Weak	83	2	166
Very Weak	8	1	8
∑	419		1212

TABLE VI. RESULT OF DATA ANALYZING CORE DRIVE-4 OWNERSHIP AND POSSESION

Answer	Number of Respondents	Value	Total
Very Strong	52	4	208
Strong	309	3	927
Weak	52	2	104
Very Weak	6	1	6
∑	419		1245

TABLE VII. RESULT OF DATA ANALYZING CORE DRIVE-5 SOCIAL INFLUENCE AND RELATEDNESS

Answer	Number of Respondents	Value	Total
Very Strong	43	4	172
Strong	252	3	756
Weak	102	2	204
Very Weak	22	1	22
∑	419		1154

TABLE VIII. RESULT OF DATA ANALYZING CORE DRIVE-6 SCARCITY AND IMPATIENCE

Answer	Number of Respondents	Value	Total
Very Strong	38	4	152
Strong	253	3	759
Weak	115	2	230
Very Weak	13	1	13
∑	419		1154

TABLE IX. RESULT OF DATA ANALYZING CORE DRIVE-7 UNPREDICTABILITY AND CURIOSITY

Answer	Number of Respondents	Value	Total
Very Strong	45	4	180
Strong	274	3	822
Weak	84	2	168
Very Weak	16	1	16
∑	419		1186

The total score on the core drive-7 unpredictability and curiosity is 1186, which is in the "high" category. This data has shown that students have a high level of anticipation for the assignments given. This condition supports students' readiness for learning materials and motivates them to always optimize their mastery of the material. This condition needs to be maintained, for example by maintaining the assignment model.

- Core Drive-8: Loss and Avoidance

The question:

"The mechanism of the e-learning system used includes achievement board facilities that can be known by all students, it has encouraged students to always maintain / increase levels so that they are not judged poorly in the learning environment."

Table X described the total score on the core drive-8 loss and avoidance is 1167, which is in the "high" category. This data has shown that students are trying to maintain a good social status so that this stimulus will increase motivation in maintaining good status, and will automatically have a positive impact on learning achievement. This condition needs to be maintained, for example by maintaining the assignment model and creating innovations in the form of maintaining status in the form of social facilities.

A. Analyzing Result Measurement of Likert Scale

The following is a comparison of all the results for the categories of 8 core drives that have been calculated (Table XI):

Therefore, from measuring the Likert scale of a questionnaire involving 8 core drives, it has been found that the motivational drive of 8 core drives is "high". Even though the research [3] has stated that students prefer traditional learning, the results of this study indicate that there is considerable potential to optimize online learning.

Meanwhile, the role of 8 core drives can explore more specifically about the diversity of students' motivations in taking online learning. The results of extracting data obtained are useful for evaluating and determining models for developing online learning systems to make them more targeted. From 8 core drives it is also possible to generate different or opposite category values, therefore, it is recommended that trials of online learning using the Octalysis framework are still needed.

TABLE X. RESULT OF DATA ANALYZING CORE DRIVE-8 LOSS AND AVOIDANCE

Answer	Number of Respondents	Value	Total
Very Strong	51	4	204
Strong	250	3	750
Weak	95	2	190
Very Weak	23	1	23
∑	419		1167

B. Converting and Analyzing Result Measurement of Octalysis Scale Test

After knowing the results of the questionnaire analysis with the Likert scale, the optimal gamification results need to be tested with the octalysis scale. For testing with the octalysis scale, the value obtained from the questionnaire results must be converted into the octalysis scale rule. Each core drive has a value of 1 to 10. The conversion steps to the octalysis scale are as follows:

- Determining the lowest point of range is obtained from the multiplication of the number of respondents (N) with the lowest weight score (value = 1), where N x 1 (419 x 1 = 419).
- Determine the highest point of range that has been obtained from the multiplication of the number of respondents (N) with the highest weight score (value = 4), where N x 4 (419 x 4 = 1676).
- Calculate the difference in range with the formula for the highest range minus the lowest range (1676 - 419 = 1257).
- Calculating the range value by dividing the range difference with the highest octalysis score (1257: 10 = 125.7), rounded up to 125. This value is used as a benchmark for the distance between levels in the measurement scale.
- Determine the level of a measuring scale consisting of 10 levels using the range of distances between levels that have been previously calculated at 125. The resulting measuring scale levels are as follows:

  Scale-1 is 419 until 544,

  Scale-2 is 545 until 670,

  Scale-3 is 671 until 797,

  Scale-4 is 798 until 923,

  Scale-5 is 924 until 1049,

  Scale-6 is 1050 until 1175,

  Scale-7 is 1176 until 1301,

  Scale-8 is 1302 until 1427,

  Scale-9 is 1428 until 1553,

  Scale-10 is 1554 until 1679.

Furthermore, the octalysis score of 8 core drives can be determined based on the previously calculated values and adjusted for the level of the measuring scale (Table XI).

From the calculation of the octalysis scale, the octagon pattern is obtained as shown in Fig. 5 and Fig. 6.

In Fig. 5, each octalysis score generated in Table XII is measured and matched to the octalysis scale provided on the official octalysis website (https://yukaichou.com/octalysis-tool/) [34]. The measurement results can be seen in Fig. 5.

TABLE XI. CONCLUSION OF LIKERT SCALE TEST FOR 8 CORE DRIVES

Octalysis Core drive		Category
1	Epic Meaning and Calling	High
2	Development and Accomplishment	High
3	Empowerment and Creativity	High
4	Ownership and Possession	High
5	Social Influence and Relatedness	High
6	Scarcity and Impatience	High
7	Unpredictability and Curiosity	High
8	Loss and Avoidance	High

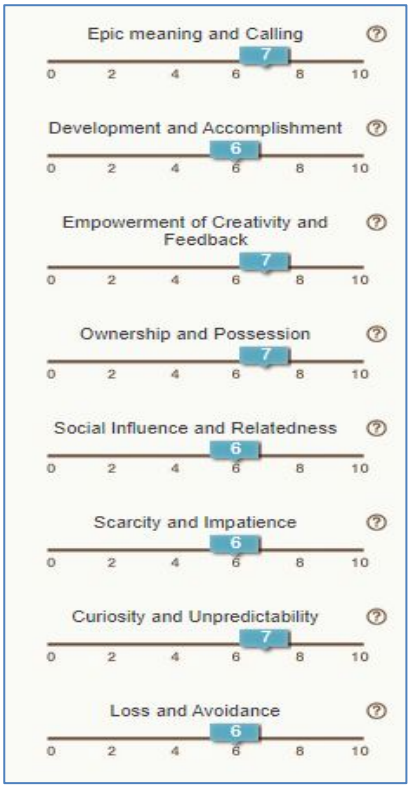


Fig. 5. Octalysis Scale.

In the octagon graph, you can see a visualization of the balance of core drives based on the questionnaire that has been generated. The octalysis value generated on the octalysis scale and octagon graph and the average manual count in Table XII is 340 which means that there is a good balance between Black Hat and White hat, and there is a balance between the left brain and the right brain. The balance of extrinsic and intrinsic motivation is also considered good.

C. Gap Analysis of e-Learning during Covid-19

In Table XIII, Table XIV, and Fig. 7, GAP data has been presented between the actual conditions and the desired conditions. For this reason, it is necessary to evaluate the causes of GAP according to the Octalysis framework.

1) Epic Meaning and Calling.

Target of statement : Very High

Target of Octalysis scale: 10

The existence of the e-learning system that is used provides the opportunity for students to participate in maintaining the stability of the learning process during the Covid-19 period optimally and consistently so that it can help strengthen the continuity and progress of e-learning.

Actual statement: High

Actual Octalysis scale value: 7

Students were motivated enough in their enthusiasm to play a role in the success of e-learning during the covid-19 period. However, students still need to improve and maintain the stability of their motivation because the actual target score has not reached the status of "Very High" and the value of the Octalysis scale still reaches 7 from the highest score of 10.

2) Development and Accomplishment.

Target statement: Very High

Target Octalysis scale value: 10

The existence of the e-learning system that is used to motivate students to improve their achievement optimally, through lecturer feedback and an assessment system that can encourage students.

Actual Statement: High

Actual Octalysis scale value: 6

If you look at the value of the octalysis scale obtained, it shows that the feedback system and the assessment system applied do not really motivate students to improve achievement. Efforts should be made to provide feedback and a more attractive scoring system because the actual target value has not yet reached the “Very High” status and the Octalysis scale value still reaches 6 from the highest score of 10.

TABLE XII. CONCLUSION OF OCTALYSIS

Octalysis Core drive		Questionnaire Score	Octalysis Scale	Octalysis Score
1	Epic Meaning and Calling	1279	7	49
2	Development and Accomplishment	1125	6	36
3	Empowerment and Creativity	1212	7	49
4	Ownership and Possession	1245	7	49
5	Social Influence and Relatedness	1154	6	36
6	Scarcity and Impatience	1154	6	36
7	Unpredictability and Curiosity	1186	7	49
8	Loss and Avoidance	1167	6	36
Total score				340

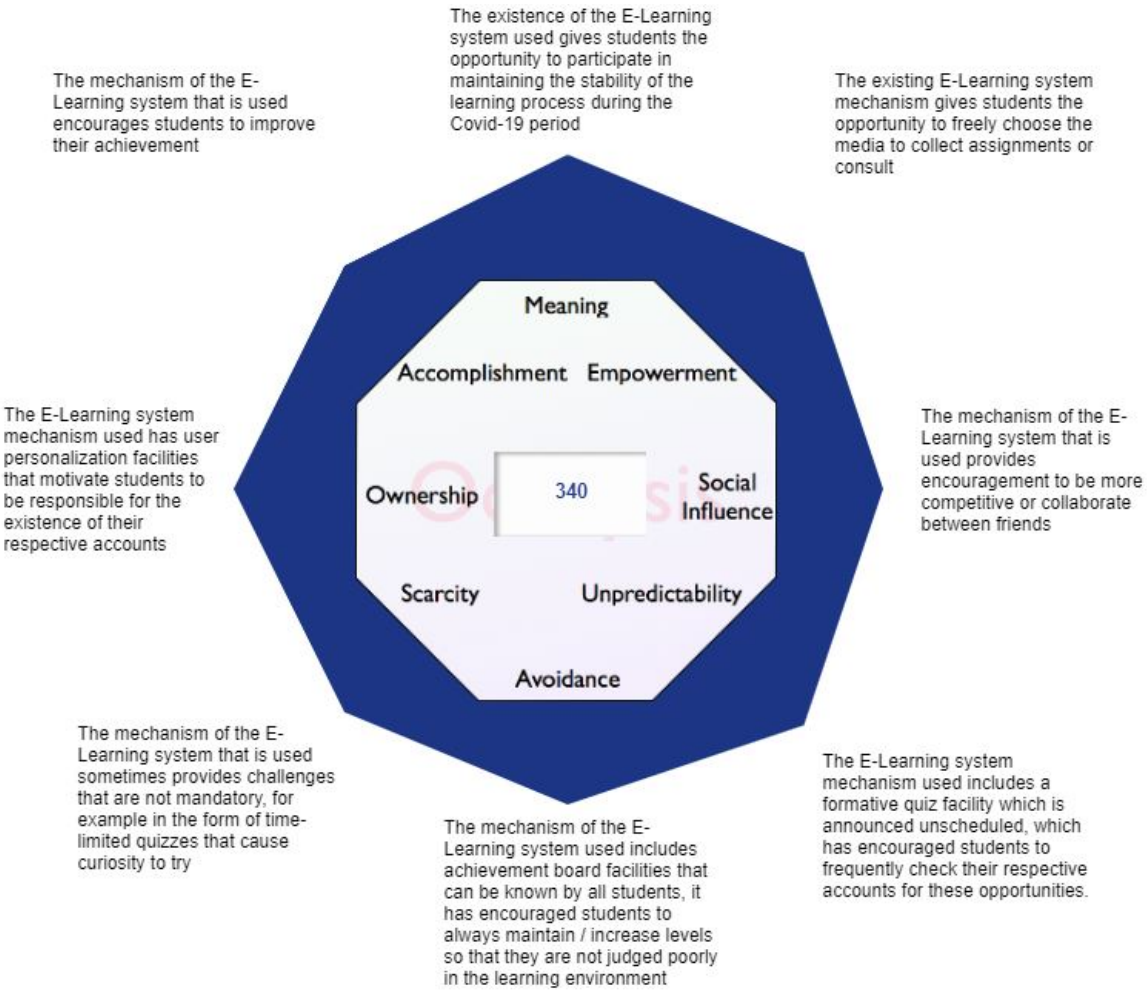


Fig. 6. Octagon Graph.

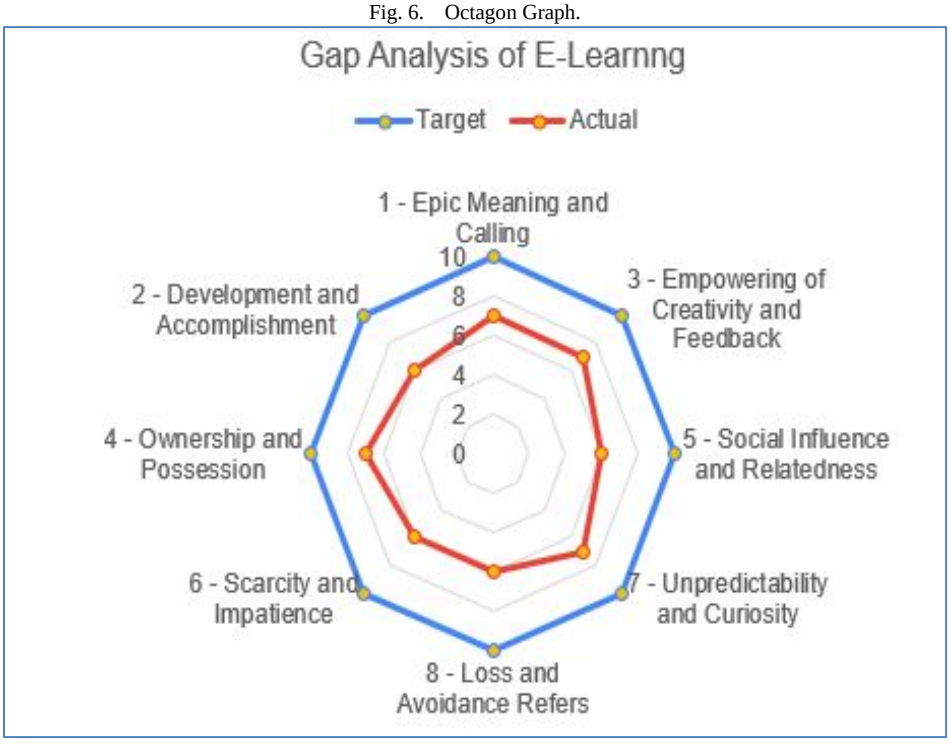


Fig. 7. The Graph of Gap Analysis.

TABLE XIII. GAP ANALISIS OF LEARNING WITH QUESTIONNAIRE SCALE

Octalysis Core drive		Actual	Target
1	Epic Meaning and Calling	High	Very High
2	Development and Accomplishment	High	Very High
3	Empowerment and Creativity	High	Very High
4	Ownership and Possession	High	Very High
5	Social Influence and Relatedness	High	Very High
6	Scarcity and Impatience	High	Very High
7	Unpredictability and Curiosity	High	Very High
8	Loss and Avoidance	High	Very High

TABLE XIV. GAP ANALISIS OF LEARNING WITH OCTALYSIS SCALE

Octalysis Core drive		Actual	Target
1	Epic Meaning and Calling	7	10
2	Development and Accomplishment	6	10
3	Empowerment and Creativity	7	10
4	Ownership and Possession	7	10
5	Social Influence and Relatedness	6	10
6	Scarcity and Impatience	6	10
7	Unpredictability and Curiosity	7	10
8	Loss and Avoidance	6	10
Average		6.5	

3) Empowerment of Creativity and Feedback.

Target statement: Very High

Target Octalysis scale value: 10

The flexibility of the e-learning system facilities, such as a media option for collecting assignments, provides comfort for students to use them optimally, thereby increasing optimal motivation for their use.

Actual Statement: High

Actual Octalysis scale value: 7

Based on the value of the octalysis scale obtained, it shows that e-learning facilities such as the freedom to choose media in collecting assignments quite motivating students to use them. However, this condition still needs to be improved because the actual target value has not yet reached the “Very High” status and the Octalysis scale value still reaches 7 from the highest value of 10.

4) Ownership and Possession.

Target statement: Very High

Target Octalysis scale value: 10

User personalization in the e-learning system motivates students to maintain ownership and increase it so as to increase achievement in learning.

Actual Statement: High

Actual Octalysis scale value: 7

Based on the value of the octalysis scale obtained, it has shown that the user personalization facility on e-learning is sufficient to motivate students to care for and maintain account ownership. However, this condition still needs to be improved, for example by adding facilities to user personalization because the actual target value has not reached the status of "Very High" and the Octalysis scale value is still on a scale of 7 from the highest value of 10.

5) Social Influence and Relatedness.

Target statement: Very High

Target Octalysis scale value: 10

The existence of the e-learning system motivates students to compete and collaborate with other friends in increasing achievement.

Actual Statement: High

Actual Octalysis scale value: 6

Based on the value of the octalysis scale obtained, it has shown that e-learning facilities still do not motivate students to compete and cooperate optimally. So, this condition needs to be improved because the actual target value has not yet reached the “Very High” status and the Octalysis scale value still reaches 6 from the highest value of 10.

6) Scarcity and Impatience.

Target statement: Very High

Target Octalysis scale value: 10

The implementation of random assignment facilities in the e-learning system arouses students' curiosity. This condition motivates them not to be absent from doing their assignments and to improve their performance.

Actual Statement: High

Actual Octalysis scale value: 6

Based on the value of the octalysis scale obtained, it shows that the random assignment facility in the existing e-learning system still does not motivate students to follow it. So, this condition needs to be improved because the actual target value has not yet reached the “Very High” status and the Octalysis scale value still reaches 6 from the highest value of 10.

7) Unpredictability and Curiosity.

Target statement: Very High

Target Octalysis scale value: 10

The implementation of the random assignment facility on the e-learning system generates an anticipatory attitude towards the given assignment. This condition motivates them not to be absent and to always check every notification of assignments so they don't miss their work.

Actual Statement: High

Actual Octalysis scale value: 7

Based on the value of the octalysis scale obtained, it shows that the random assignment facility in the applicable e-learning system is quite motivating students to anticipate the random assignment given. However, this condition needs to be improved because the actual target value has not yet reached the “Very High” status and the Octalysis scale value still reaches 7 from the highest value of 10.

8) Loss and Avoidance.

Target statement: Very High

Target Octalysis scale value: 10

The implementation of the achievement board sharing facility provides a stimulus to students to always show their best performance so that they are not judged poorly in their learning environment. This condition motivates them to always maintain and improve their performance.

Actual Statement: High

Actual Octalysis scale value: 6

Based on the value of the octalysis scale obtained, it has shown that the achievement board facility imposed by the e-learning system is not sufficient to motivate students to encourage students to maintain and improve their achievement. So, this condition needs to be improved because the actual target value has not yet reached the “Very High” status and the Octalysis scale value still reaches 6 from the highest value of 10.

VI. Recommendations

Based on the results of the GAP evaluation of online learning conducted by questionnaire and octalysis scale analysis, it is suggested that recommendations based on 8 core drives are as follows:

1) As an emergency system, the potential for "Epic Meaning and Calling" is quite good in the applicable e-learning. This potential needs to be maintained and increased because it involves a large amount of intrinsic motivation. Students are motivated to involve themselves in the e-learning system during the Covid-19 period because their involvement can take part in the mission of successful learning in an emergency. With the value of the "High" scale and the octalysis scale of 7, this condition still needs to be improved, such as providing an overview and explanation of the objectives of each material to be studied, so that students understand the purpose of studying.

2) e-Learning efforts to provide a stimulus that can motivate achievement have been implemented, but this has not been able to motivate students to improve achievement optimally. The feedback and assessment system model that is applied needs to be evaluated and developed to make it more attractive. Therefore, adding variations in the assessment model is needed so that students have a choice in choosing a model. The achievement presentation model can adopt the types of gamification mechanics such as points, levels, badges, and leaderboards to package achievements in the hope that it can further improve students to increase achievement. With the value of the "High" scale and the octalysis scale of 6, this condition still needs to be explored.

3) The flexibility to use e-learning media plays a role in increasing drives "Empowerment and Creativity". The current e-learning has implemented facilities such as the freedom to collect assignments, and this is enough to attract students as evidenced by the value of the "High" scale. However, this method needs to be developed considering the octalysis scale still reaches a value of 7. "Empowerment and Creativity" is based on the idea that people not only need a way to express creativity, but they need a way to see the results of that creativity, receive feedback, and respond in turn. Several mechanics need to be tried out to improve this drive, such as blank fills combos to give students the opportunity to write free notes, instant feedback to give students the opportunity to give feedback on the material being discussed.

4) The application of user personalization facilities in the applicable e-learning system has created sufficient student comfort, as evidenced by the acquisition of a "High" questionnaire scale value and an octalysis scale value of 7. However, this condition still needs to be improved again considering the achievement of the octalysis value and the level of the questionnaire scale has not reached the highest point. “Ownership and Possession is based on encouragement in which users are motivated because they feel they have something. When a player has a sense of belonging, he automatically wants to make what he has better and even more. For this reason, exploration of mechanics related to this drive needs to be explored, such as applying avatars as student identities according to achievement and character, or holding "virtual good" to increase ownership.

5) The current existence of e-learning is not enough to generate motivation for cooperation or competition between students. This is evidenced by the octalysis scale obtained by 6 while the highest score is 10. Drive "Social Influence and Relatedness" is based on the encouragement of someone to do something because it is inspired by others, both in terms of competition and cooperation. It is necessary to explore possible methods to generate this thrust. Some mechanics that need to be tried include group quests, friending, and social treasures, each of which are mechanics.

6) e-Learning that has been running has implemented a facility that generates a "Scarcity and Impatience" drive, for example the facility to provide random assignments that are unscheduled in nature. In fact, this is not enough to encourage students to improve this drive, which is evidenced by the value of the Octalysis scale which is still in the range of 6. The “Scarcity and Impatience” drive is based on the idea that something that is rare and difficult to obtain motivates individuals to want to have. Several mechanics that can be adopted in e-learning include "countdown timer" and "Appointment Dynamics". These two mechanics are used in e-learning for assignment systems with the hope of motivating students to always follow them.

7) The "Unpredictable and Curiosity" drive in e-learning has been implemented in the form of random assignments. Its existence is sufficient to motivate students, as evidenced by the octalysis scale value is in the range of 7, but this still needs to be improved to reach 10. This drive uses the same facilities as the 6th drive "Scarcity and Impatience" but the difference is from the point of view of the drive that is raised. This drive is based on the individual always thinking about what will happen next, thus encouraging him to always anticipate something that will happen that will come in his favor. The task systems that generate these drives need to be explored more deeply. This mechanic drive model that can be applied in e-learning includes "random reward", "rolling reward" both of which can be applied in the assessment and assignment systems.

8) Drive "Loss and Avoidance" in the current e-learning system has been applied in the form of sharing achievement boards with the hope that it can motivate students to anticipate by keeping their achievements so they are not judged negatively in their environment. However, this drive is not optimally absorbed, as evidenced by the Octalysis scale value of 6 from the highest scale of 10. This drive is based on individual efforts to avoid something negative, so someone takes anticipatory action against the conditions at hand. In the e-learning system, the developer can adopt the mechanics that are usually used in this drive such as "Progress Loss" and "Status Quo Slot" in the scoring system with the hope that students can be motivated to take anticipatory action against the actual conditions that occur.

VII. Conclusion and Future Works

In general, the implementation of e-learning during the Covid-19 period, especially in Indonesia, was quite acceptable to students, as evidenced by the results of the GAP analysis on the questionnaire analysis on a "high" scale while the highest achievement was "very high". However, if motivation is analyzed more deeply using the Octalysis scale, the average results are in the range of 6.5 from the scale range 1 to 10. Therefore, on the one hand, this actual condition has the potential to be developed, but more innovative development is needed to generate student core drives so that it will increase motivation to learn.

Gamification can optimally support the e-learning system, if it is taken seriously. Because in reality society really needs refreshment, something new that can make an object interesting and that continues to develop [35]. The Octalysis framework provides a choice of mechanics to explore learning models that can generate and enhance core drives more optimally. Each core drive has provided recommendations and guides available on the official website https://yukaichou.com/octalysis-tool/[34] which can be accessed freely as a standard. Several mechanics in accordance with the needs of the system are adopted and explored, therefore, that it is expected to support the optimization of students' core drives for e-learning.

Acknowledgment

All Authors would like to thank all who have gave contributions to this research starting from DRPM KEMENRISTEK BRIN of Ministry of Research and Technology & Higher Education, Indonesia; Universiti Teknikal Malaysia Melaka (UTeM); Malaysia, Widyagama University of Malang, Indonesia; President University, Indonesia; Indonesia and all colleagues who have contributed for valuable directions and suggestions to strength this quality paper.

References

[1] L. Mishra, D. T. Gupta, and D. A. Shree, "Online Teaching-Learning in Higher Education during Lockdown Period of COVID-19 Pandemic," Int. J. Educ. Res. Open, p. 100012, 2020, doi: 10.1016/j.ijedro.2020.100012.

[2] A. Sharma, S. B. Borah, and A. C. Moses, "Responses to COVID-19: The Role of Governance, Healthcare Infrastructure, and Learning from Past Pandemics," J. Bus. Res., 2020, doi: 10.1016/j.jbusres.2020.09.011.

[3] A. Patricia, "College Students' Use and Acceptance of Emergency Online Learning Due to COVID-19," Int. J. Educ. Res. Open, p. 100011, 2020, doi: 10.1016/j.ijedro.2020.100011.

[4] T. Y. Chang et al., "Innovation of dental education during COVID-19 pandemic," J. Dent. Sci., no. 155, 2020, doi: 10.1016/j.jds.2020.07.011.

[5] E. M. Aucejo, J. French, M. P. U. Araya, and B. Zafar, "The Impact of COVID-19 on Student Experiences and Expectations: Evidence from a Survey.," J. Public Econ., vol. 191, p. 104271, 2020, doi: 10.1016/j.jpubeco.2020.104271.

[6] D. Marques de Miranda, B. da Silva Athanasio, A. Cecília de Sena Oliveira, and A. C. Simoes Silva, "How is COVID-19 pandemic impacting mental health of children and adolescents?," Int. J. Disaster Risk Reduct., vol. 51, no. September, p. 101845, 2020, doi: 10.1016/j.ijdrr.2020.101845.

[7] N. Hasan and Y. Bao, "Impact of 'e-Learning crack-up' perception on psychological distress among college students during COVID-19 pandemic: A mediating role of 'fear of academic year loss,'" Child. Youth Serv. Rev., vol. 118, no. July, p. 105355, 2020, doi: 10.1016/j.childyouth.2020.105355.

[8] B. Oosterhoff, C. A. Palmer, J. Wilson, and N. Shook, "Adolescents' Motivations to Engage in Social Distancing During the COVID-19 Pandemic: Associations With Mental and Social Health," J. Adolesc. Heal., vol. 67, no. 2, pp. 179–185, 2020, doi: 10.1016/j.jadohealth.2020.05.004.

[9] M. Saefi et al., "Survey data of COVID-19-related knowledge, attitude, and practices among indonesian undergraduate students," Data Br., vol. 31, p. 105855, 2020, doi: 10.1016/j.dib.2020.105855.

[10] F. Marisa, S. S. S. Ahmad, Z. I. M. Yusof, Fachrudin, and T. M. A. Aziz, "Segmentation model of customer lifetime value in Small and Medium Enterprise (SMEs) using K-Means Clustering and LRFM model," Int. J. Integr. Eng., vol. 11, no. 3, pp. 169–180, 2019, doi: 10.30880/ijie.2019.11.03.018.

[11] F. Marisa, S. Sakinah, S. Ahmad, Z. Izzah, M. Yusoh, and T. M. Akhriza, "Performance Comparison of Collaborative-Filtering Approach with Implicit and Explicit Data," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 10, pp. 110–116, 2019, [Online]. Available: https://thesai.org/Downloads/Volume10No10/Paper_16-Performance_Comparison_of_Collaborative_Filtering_Approach.pdf.

[12] I. Istiadi, F. Rofii, A. Qustoniah, F. Marisa, and G. D. Putra, "Online Expert System for Consultation Services using a Mobile," J. Teknol., vol. 78, no. 6–3, pp. 41–46, 2016, [Online]. Available: https://journals.utm.my/index.php/jurnalteknologi/article/view/8926.

[13] M. C. Carnero, "Fuzzy multicriteria models for decision making in gamification," Mathematics, vol. 8, no. 5, 2020, doi: 10.3390/MATH8050682.

[14] S. A. A. Freitas, A. R. T. Lacerda, P. M. R. O. Calado, T. S. Lima, and E. D. Canedo, "Gamification in education: A methodology to identify

student's profile," Proc. - Front. Educ. Conf. FIE, vol. 2017-Octob, no. December, pp. 1–8, 2017, doi: 10.1109/FIE.2017.8190499.

[15] J. Kasurinen and A. Knutas, "Publication trends in gamification: A systematic mapping study," Comput. Sci. Rev., vol. 27, pp. 33–44, 2018, doi: 10.1016/j.cosrev.2017.10.003.

[16] A. M. Toda, R. M. C. do Carmo, A. P. da Silva, I. I. Bittencourt, and S. Isotani, "An approach for planning and deploying gamification concepts with social networks within educational contexts," Int. J. Inf. Manage., vol. 46, no. October 2018, pp. 294–303, 2019, doi: 10.1016/j.ijinfomgt.2018.10.001.

[17] J. Landsell and E. Hägglund, Towards a Gamification Framework: Limitations and opportunities when gamifying business processes. 2016.

[18] Y. Chou, Actionable Gamification, Beyond Points, Badges, and Leaderboards - Google Play. 2019.

[19] Y.-K. Chou, Actionable gamification: Beyond points, badges, and leaderboards. 2016.

[20] R. Radha, K. Mahalakshmi, V. S. Kumar, and A. R. Saravanakumar, "E-Learning during Lockdown of Covid-19 Pandemic: A Global Perspective," Int. J. Control Autom., vol. 13, no. 4, pp. 1088–1099, 2020.

[21] Mailizar, A. Almanthari, S. Maulina, and S. Bruce, "Secondary school mathematics teachers' views on e-learning implementation barriers during the COVID-19 pandemic: The case of Indonesia," Eurasia J. Math. Sci. Technol. Educ., vol. 16, no. 7, 2020, doi: 10.29333/EJMSTE/8240.

[22] K. Robson, K. Plangger, J. H. Kietzmann, I. McCarthy, and L. Pitt, "Is it all a game? Understanding the principles of gamification," Bus. Horiz., vol. 58, no. 4, pp. 411–420, 2015, doi: 10.1016/j.bushor.2015.03.006.

[23] Risal, "Pembangunan Gamification (Game Mechanics) Framework," 2013.

[24] P. H. Bachtiar, W. S. Wardhono, and T. Afirianto, "Pendekatan MDA Framework Pada Pengembangan Permainan Baby Care Augmented Reality dengan Outfit Marker," J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya, vol. 2, no. 12, 2018.

[25] F. Marisa, S. Sakinah, S. Ahmad, and Z. I. Mohd, "Analysis Of Relationship CLV with 8 Core Drives Using Clustering K-Means and Octalysis Gamification Framework," J. Theor. Appl. Inf. Technol., vol. 98, no. 20, pp. 3151–3164, 2020, [Online]. Available: http://www.jatit.org/volumes/Vol98No20/6Vol98No20.pdf.

[26] M. Urh, G. Vukovic, E. Jereb, and R. Pintar, "The Model for Introduction of Gamification into E-learning in Higher Education," in Procedia - Social and Behavioral Sciences, 2015, vol. 197, no. February, pp. 388–397, doi: 10.1016/j.sbspro.2015.07.154.

[27] L. Aguiar-Castillo, L. Hernández-López, P. De Saá-Pérez, and R. Pérez-Jiménez, "Gamification as a motivation strategy for higher education students in tourism face-to-face learning," J. Hosp. Leis. Sport Tour. Educ., vol. 27, no. September, 2020, doi: 10.1016/j.jhlste.2020.100267.

[28] N. L. Mingoc, E. Louwe, and R. Sala, "Design and Development of Learn Your Way Out: A Gamified Design and Development of Learn Your Way Out: A Gamified Content for Basic Java Computer Programming," in Procedia Computer Science, 2019, vol. 161, pp. 1011–1018, doi: 10.1016/j.procs.2019.11.211.

[29] D. Kayimbaşioğlu, B. Oktekin, and H. Haci, "Integration of Gamification Technology in Education," in Procedia Computer Science, 2016, vol. 102, no. August, pp. 668–676, doi: 10.1016/j.procs.2016.09.460.

[30] N. Z. Legaki, N. Xi, J. Hamari, K. Karpouzis, and V. Assimakopoulos, "The effect of challenge-based gamification on learning: An experiment in the context of statistics education," Int. J. Hum. Comput. Stud., vol. 144, no. June, 2020, doi: 10.1016/j.ijhcs.2020.102496.

[31] L. M. Putz, F. Hofbauer, and H. Treiblmaier, "Can gamification help to improve education? Findings from a longitudinal study," Comput. Human Behav., vol. 110, no. November 2019, p. 106392, 2020, doi: 10.1016/j.chb.2020.106392.

[32] G. P. Kusuma, E. K. Wigati, Y. Utomo, and L. K. Putera Suryapranata, "Analysis of Gamification Models in Education Using MDA Framework," in Procedia Computer Science, 2018, vol. 135, pp. 385–392, doi: 10.1016/j.procs.2018.08.187.

[33] V. H. Pranatawijaya, W. Widiatry, R. Priskila, and P. B. A. A. Putra, "Penerapan Skala Likert dan Skala Dikotomi Pada Kuesioner Online," J. Sains dan Inform., vol. 5, no. 2, p. 128, 2019, doi: 10.34128/jsi.v5i2.185.

[34] Y.-K. Chou, "Gamification Framework," Gamification Framework, 2020. https://yukaichou.com/gamification-examples/octalysis-complete-gamification-framework/.

[35] G. Wang and A. Y. Ariyanto, "Gamification: Strengthening The Relationship," J. Theor. Appl. Inf. Technol., vol. 97, no. 17, pp. 4490–4507, 2019.

Covid-19 Ontology Engineering-Knowledge Modeling of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)

Vinu Sherimon[1]*, Sherimon P.C.[2], Renchi Mathew[3], Sandeep M. Kumar[4]
Rahul V. Nair[5],Khalid Shaikh[6], Hilal Khalid Al Ghafri[7], Huda Salim Al Shuaily[8]
Department of IT, University of Technology & Applied Sciences, Muscat, Sultanate of Oman[1]
Faculty of Computer Studies, Arab Open University, Muscat, Sultanate of Oman[2]
Senior Consultant, Internal Medicine, Royal Oman Police Hospital, Muscat, Sultanate of Oman[3]
Specialist, Family Medicine & Tropical Medicine, Royal Oman Police Hospital, Muscat, Sultanate of Oman[4]
Medical officer (A), Emergency department, Royal Oman Police Hospital, Muscat, Sultanate of Oman[5]
Specialist, Internal Medicine and Diabetes, Royal Oman Police Hospital, Muscat, Sultanate of Oman[6]
Specialist, Internal Medicine, Royal Oman Police Hospital, Muscat, Sultanate of Oman[7]
Department of IT, University of Technology & Applied Sciences, Muscat, Sultanate of Oman[8]

***Abstract*—COVID-19 pandemic has rapidly spread across the world since its arrival in December 2019 from Wuhan, China. This pandemic has disrupted the health of the citizens in such a way that the impact is enormous in terms of economy and social aspects. Education, employment, income, well-being of the humankind is affected very crucially by this corona virus. Nations world-wide are struggling to battle this emergency. Intensive studies are being carried out to control this pandemic by researchers all over the world. Medical science has advanced a lot with the application of computer assisted solutions in health care. Ontology based clinical decision support systems (CDSS) assist medical practitioners in the diagnosis and treatment of diseases. They are well known in data sharing, interoperability, knowledge reuse, and decision support. This research article presents the development of ontology for SARS-CoV-2 (COVID-19) to be used in a CDSS, which is proposed in the satellite clinics of Royal Oman Police (ROP), Sultanate of Oman. The key concepts and the concept relationships of COVID-19 is represented using an ontology. Semantic Web Rule Language (SWRL) is used to model the rules related to the initial diagnosis of the patient and Semantic Query Enhanced Web Rule Language (SQWRL) is used to retrieve the data stored in the ontology. The developed ontology successfully classified the patients into one of the different categories as non-suspected, suspected, probable, and confirmed. The reasoning time and the query execution time is found to be optimal.**

***Keywords*—COVID-19; ontology; SARS-CoV-2; ontology reasoning; SWRL; SQWRL**

I. INTRODUCTION

Around the world, the crisis of coronavirus disease is growing unprecedently. Reported from Wuhan, the capital city of the Hubei Province in China in December 2019, World Health Organization (WHO) declared this as a pandemic on 11th March 2020 [1]. The virus which cause this disease is termed as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). As of November 17th, 2020, the number of active cases in the world are 15,526,034 and the number of deaths is 1,335,263 [2]. Due to the coronavirus, countries have issued travel bans. Many countries have implemented 14-21 days of complete lockdown. The sad reality is that people are locked down in houses and are filled with anxiety, stress and depression. People are afraid to go out. Relevant authorities of almost all countries provide necessary support and guidance to all the cases, but still the fear conquers the world.

Physicians and researchers from all over the world are conducting intensive studies to know about this novel virus and its pathogenesis, to find epidemiological factors, to explore different candidate vaccines, to investigate the effect of therapies, to perform randomized trials etc. to come up with effective drugs to handle this global pandemic. Countries are competing among themselves to become the first to develop effective drug to control this situation. Every day, news agencies are reporting about the success stories of testing vaccines to control this virus. Till this date, no approved therapies are known for this virus.

Ministry of Health (MOH), Sultanate of Oman reported the first cases of corona virus in February 24, 2020 [16]. Now with nine months completed, 7922 active cases exist, and 1350 deaths occurred as of November 17th, 2020 in Oman [3]. Country witnessed several lockdowns as part of the effective control measures to prevent the spread of the virus. MOH being the main agency to oversee the health sector in Oman, Royal Oman Police (ROP) hospital is also at the front-line to serve the citizens of the country. ROP operates several satellite clinics in the interior regions of Oman. These centres may not have expert doctors to diagnose such diseases. Also, in some of these centres, there are not enough facilities to diagnose coronavirus diseases. Because of the fear of community spread, people are unable to travel to hospitals in the capital city to seek expert care and support. So, to support the health care workers of satellite clinics of ROP in diagnosing COVID-19, we have proposed a knowledge based clinical decision support system (CDSS).

The objective of this research paper is to outline the design of COVID-19 ontology which will be used to represent the

*Corresponding Author

knowledge base of the proposed decision support system. The ontology covers the symptoms, risk factors, epidemiological factors, initial diagnosis, lab tests, clinical diagnosis, recommendations and treatment in the context of Sultanate of Oman. National Clinical Management Protocol for Hospitalized Patients with COVID-19, ICU protocol for management of COVID-19 published by Ministry of Health, Sultanate of Oman is used as the clinical guidelines for this research [13]. We created the knowledge base as per the above guidelines currently followed in Oman. The development details of CDSS is outside the scope of this article.

The rest of the paper is outlined as follows: - Background is given in Section 2 followed by the Methodology. The methodology section includes the modelling of the ontology and the construction of rules and queries. Results are given in Section 4 followed by Discussion in Section 5. The performance of the developed ontology is shown in Section 6 and the Conclusion & Future is presented in Section 7.

II. Background

Semantic web applications interpret not just the content and structure of any presentation, rather it can understand the meaning of the content. Machines conduct automatic reasoning to accomplish the task of interpreting the textual meaning. Concepts must be structured and accompanied by a set of inference rules to perform logical reasoning [4]. Technologies in semantic web includes Extensible Markup Language (XML), Resource Description Framework (RDF), Ontology and many more. These technologies ensure to connect the data, rather than the connection between documents through links. Instead of URLs between documents, semantic web implements URLs between facts. RDF express any sentence in the form of triplets (subject, predicate and object) [4]. Moreover, a Universal Resource Identifier (URI) is used to identify every subject and object [4]. XML doesn't support semantics [4]. Instead, it is used to represent the RDF triplets [4]. RDF use RDF Schema (RDFS) to describe the terms used in a sentence [4].

In 1993, Tom Gruber proposed the definition of ontology [5] as "An ontology is a formal, explicit specification of a conceptualization". Ontology defines an abstract model about a particular domain, explains precisely the terms in a domain and expresses the relationships between the terms thus making it explicit, it is machine understandable, so it is formal, and it is accepted by a group and shared. The primary aim of ontology is to ease knowledge sharing and reuse [4]. The integration of RDF and the underlying ontology makes the meaning understandable to the machines [4]. Web Ontology Language (OWL) extends RDFS and add features such as union, intersection, cardinalities, reasoning and inferring capabilities [4]. Description Logics (DL) are used to represent the semantics of OWL [6].

In the case of epidemic outbreaks, action research by integrating different disciplines such as medicine, computer science, sociology, psychology etc. is mandatory. Such research requires the integration of data from varied sources [7]. Ontologies are well known among biomedical researchers for data sharing, integration and reuse. Ontologies are used to represent any domain knowledge in a formal way. Biomedical researchers use ontology to include the concepts (entities) in the biomedical domain and to express their connections (relationships). The relationships are semantic in nature, and accurately represent the domain terms and its relationships. Ontology plays an important role in knowledge sharing and reuses and it supports automatic reasoning. They are used in clinical decision support systems to support health care workers.

Since its inception as the foundation of Artificial Intelligence (AI), ontologies are used in different disciplines for knowledge sharing. Among the successful ontologies in life sciences include Gene Ontology [8], Disease Ontology [9], SNOMED-CT [10], ICD-10 [11] etc.

III. Methodology

A. Conceptual Modelling

Fig. 1 illustrates the entity-relationship diagram of the proposed system. Patient entity has a relationship with Symptom, Risk Factor, Epidemiology, Diagnosis, Test, Clinical Diagnosis, Recommendation and Treatment entities. Based on the Symptoms, Risk Factors, and Epidemiology, the patient will be first categorized into one of the cases – non-suspected, suspected, probable, and confirmed. Then based on the Clinical Diagnosis and Test reports, suitable Recommendation and Treatment will be provided to the Patient. The relationship between Patient and Diagnosis and Patient and Clinical Diagnosis is 1:1 as the Patient will be categorized into one of the states as mentioned above. Relationship of Patient entity with all other entities are M:N.

B. Modelling of Domain Knowledge

The process of modeling the domain using ontologies starts with defining the top concept of the domain. It is usually represented by *owl:Thing*. Then the concept classes in the domain is defined as $C = \{C_1, C_2, \ldots\ldots, C_n\}$ [17]. For each of the concept class C_n, sub-classes are also defined as $C_n = \{s_1, s_2, \ldots\ldots, s_m\}$ to form the concept hierarchy [12].

The key knowledge required to build our knowledge base is to understand about the symptoms, diagnosis, available treatments etc. of SARS-CoV-2. The important concepts are represented as classes in the ontology. Patient, Diagnosis, Symptom, Background_history, LabTest, RiskFactor, ClinicalDiagnosis, Recommendation, and Treatment are the main concepts.

Fig. 2 represents the relevant concepts (such as demographic, epidemiological, clinical based features etc) in SARS-CoV-2 domain.

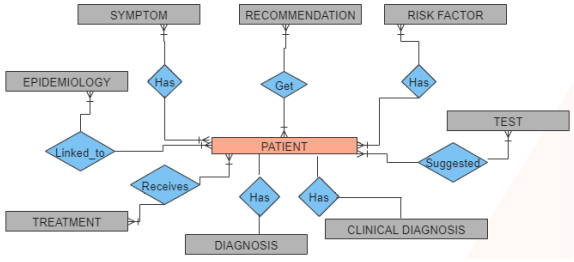


Fig. 1. Conceptual Schema Modelling.

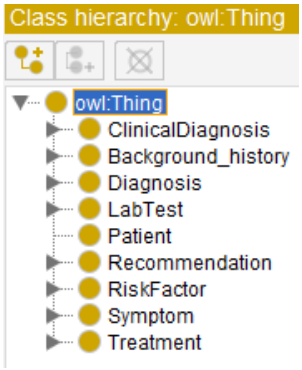


Fig. 2. Taxonomy of Main Domain Concepts Represented as Classes in SARS-CoV-2 Ontology.

All the patient cases will be represented as instances of Patient class. To represent the different symptoms of SARS-CoV-2, Symptom class is created. All the reported symptoms as per the Oman and WHO guidelines [13,14] are represented as subclasses of this class. Epidemiological link of the case is represented using the Background_history class through various sub-classes. RiskFactor class includes subclasses to represent the important risk factors of the case (comorbidities, age greater than 60, consumption of immunosuppressive drugs, etc.).

The Diagnosis class is used to categorize a case into one of the different categories non-suspected, suspected, probable, and confirmed. All these categories form the subclasses of Diagnosis class. Classification of a case into one of the above categories is done after the reasoning of the ontology. ClinicalDiagnosis class is used to represent the cases into different categories such as mild, moderate, severe and critical based upon the clinical conditions. LabTest class include different investigations (mandatory tests and additional tests) to be performed, depending on the case. Recommendation class is used to provide recommendations to the users. The treatment suggestions as per the situation of each case is represented in Treatment class.

Object properties and data properties are shown in Fig. 3 and Fig. 4.

C. Semantic Web Rule Language and Semantic Query Enhanced Web Rule Language

We have used here Semantic Web Rule Language (SWRL) to construct the rules. It is an OWL based language which is expressive in nature and adds power to Description Logic (DL). SWRL is a combination of DL and First Order Logic (FOL). The antecedent-consequent way of expressing a rule allows SWRL to link with relational databases. The rules written in SWRL provide more reasoning ability than OWL.

The body part of the rule is referred as antecedent and the head part is referred as the consequent. It takes the form,

$atom1 \wedge atom2 \ldots\ldots \rightarrow atom \wedge atom \ldots\ldots$

The rule is basically a combination of conjunction of atoms where each atom is expressed in the form of

$predicate(arg1, arg2, \ldots\ldots, argn)$

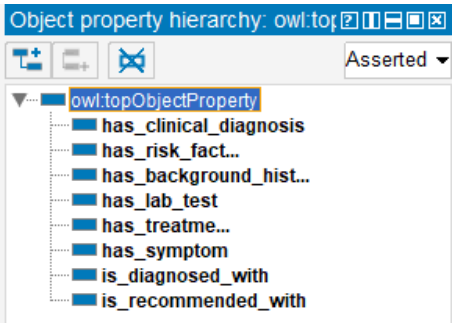


Fig. 3. Object Properties.

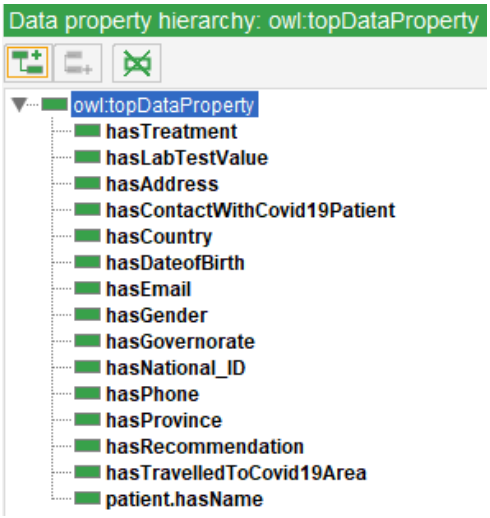


Fig. 4. Data Properties.

The concepts in OWL such as classes, properties, data types, ranges and built-in functions are used as SWRL predicates. Arguments include the values of the properties or OWL instances.

Semantic Query Enhanced Web Rule Language (SQWRL) is used to extract information from the ontology [15]. This query language supports SQL like operators.

It is used along with SWRL to extract the information retrieved by SWRL rules [15]. We have also used SQWRL collections to construct queries which involves disjunction, which otherwise is not possible using SWRL. The left side of SQWRL takes the form like SWRL antecedent and retrieve the data in the consequent part in the right side of the query. sqwrl:select is the primary operator used in SQWRL. Here, the SWRL inference rules are executed initially followed by the execution of queries.

D. Construction of Rules and Queries

The proposed CDSS will have a separate user interface for patient (online) and several other interfaces for health workers. So, accordingly we divide the working of this ontology into two parts. The first part will be used by Module 1 (patient interface) of our proposed CDSS. OWL classes such as Background_history, RiskFactor, Symptoms, Diagnosis and Recommendation are used in this part. Module 1 provides the necessary recommendations to the patient from the home itself regarding his/her present health condition related to COVID-

19, based on the initial diagnosis. Module 2 of the proposed CDSS will use the concepts related to Clinical Diagnosis, Lab Test, and Treatment. It provides the required suggestions and recommendations to the medical practioners based on the results of clinical diagnosis and lab test values. In this article, we focus on the initial diagnosis part of the ontology.

1) Rules: Rules are constructed to initially diagnose different patient cases into one of the categories of Diagnosis class: Nonsuspected, Suspected-Asymptomatic, Suspected-Symptomatic, Probable and Confirmed (Fig. 5). SWRL rules related to the initial diagnosis is presented in Table I.

Case#1 represents cases that doesn't have any symptoms and background history yet. The corresponding rule is used to diagnose such non-suspected cases. Antecedent part of Rule 1 consists of four atoms. Each atom is expressed in the form of predicate with arguments. The first predicate Patient (?p) is used to retrieve all the patients from the ontology and store the value in variable 'p'. Next predicate has_symptom (?p,No_symptom) checks whether the patient has no symptoms. No_symptom is an individual of the same class.

hasContactWithCovid19Patient (?p, false) checks the background history of the patient whether the patient has contact with any infected cases and hasTravelledToCovid19Area (?p, false) checks about the travel history of the patient. There are 3 atoms in the consequent part of this rule. If the conditions given in the predicates of the body are satisfied, then the predicate is_diagnosed_with is used to display the diagnosis result as Nonsuspected_case.

The predicat is_recommended_with(?p,Nonsuspected_case_recommendation) fetches the recommendation suggested by the system to such non-suspected patients. Also, such patients are automatically categorized as instances (individuals) of Nonsuspected_case class. This is the functionality of the predicate Nonsuspected_case(?p). When the patient is categorized as a non-suspected case, a concerned SWRL rule will provide the recommendation through a data property hasRecommendation.

The cases with no symptoms, but suspected history of travel to infected areas/contact with COVID-19 cases are classified as a Suspected_Asymptomatic case (Case#2). If the predicates has_symptom(?p,No_symptom), hasContactWith Covid19Patient (?p,true), and hasTravelledToCovid19Area (?p,false) returns true, then such patients are automatically categorized as instances (individuals) of Asymptomatic class. Also, the corresponding diagnosis and the recommendation is also provided.

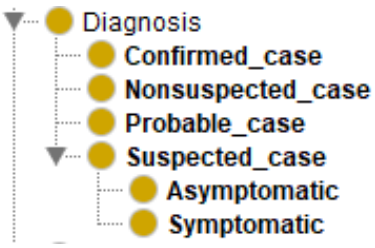


Fig. 5. Subclasses of Diagnosis Class.

TABLE I. SWRL SAMPLE RULES TO DIAGNOSE DIFFERENT PATIENT CASES BASED ON THE INPUT PROVIDED

No	Patient case	SWRL rule
Case# 1	Nonsuspected case	Patient (?p) ^ has_symptom(?p,No_symptom) ^ hasContactWithCovid19Patient (?p,false) ^ hasTravelledToCovid19Area (?p,false) ^ →Nonsuspected_case(?p) ^ is_diagnosed_with(?p,Nonsuspected_case) ^ is_recommended_with (?p,Nonsuspected_case_recommendation)
Case# 2	Suspected Asymptomatic case	Patient (?p) ^ has_symptom(?p,No_symptom) ^ hasContactWithCovid19Patient (?p,true) ^ hasTravelledToCovid19Area (?p,false) ^ →Asymptomatic(?p) ^ is_diagnosed_with(?p,Suspected_asymptomatic_case) ^ is_recommended_with (?p,Suspected_asymptomatic_case_recommendation)
Case# 3	Suspected Symptomatic case	Patient (?p) ^ has_symptom(?p,Chills) →Symptomatic(?p) ^ is_diagnosed_with(?p,Suspected_symptomatic_case) ^ is_recommended_with (?p,Suspected_symptomatic_case_recommendation)
Case# 4	Probable case	Patient (?p) ^ has_symptom(?p,Breating_Difficulty) →Probable_case(?p) ^ is_diagnosed_with(?p,Probable_case) ^ is_recommended_with (?p,Probable_case_recommendation)
Case# 5	Confirmed case	Patient (?p) ^ has_lab_test(?p,RTPCR) ^ hasLabTestRTPCRValue(?p,true) →Confirmed_case(?p) ^ is_diagnosed_with(?p,Confirmed_case) ^ is_recommended_with (?p,Confirmed_case_recommendation)

Patients with upper respiratory tract viral infection reports non-specific symptoms such as chills, fatigue, cough, muscle pain etc. which are uncomplicated. As per the guidelines of WHO and Oman, it is suggested not to have immediate hospital referrals for such cases. System advice such patients to have home quarantine and in case if the symptom worsens, it suggests taking an appointment in the hospital to do PCR swab test. For example, in Table I, Case#3 represents a suspected symptomatic case and the given SWRL rule checks the patients with the mild symptoms reported. As SWRL doesn't support disjunction of atoms, we have written separate rules to check the occurrence of every symptom. We didn't use the OWL class union operator which supports disjunction. All the mild symptoms are included in the ontology to check all possible symptomatic cases.

As per the MOH guidelines, a person suffering from high fever, shortness of breath, and chest discomfort should be immediately reported to the hospital for admission and for further treatments. These cases are considered as probable cases of COVID-19. An SWRL rule regarding the probable

case is given in Table I (Case#4). The probable cases of COVID-19 are immediately suggested to do RTPCR test. Case#5 shows the SWRL rule regarding the confirmed case.

Table II represents the recommendations suggested by the system. For each case, the corresponding SWRL rules are added to the ontology.

2) Querying the ontology: After constructing SWRL rules to diagnose different patient cases, SQWRL is used to construct queries to retrieve the relevant data. For example, the execution of the SWRL rule related to Case#5 in Table I will classify all instances of confirmed cases under $Confirmed\_case$ class of the ontology. The data related to all such patients can be retrieved using the following SQWRL query in Table III. $hasPatientCivilID$, $hasAddress$, $hasAge$ and $hasGender$ are the data properties and the *select* operator is used to retrieve this information of patients confirmed with COVID-19. Similarly, we constructed several queries to retrieve information of patients with one or more risk factors, symptoms, etc.

TABLE II. INITIAL RECOMMENDATIONS

Patient case	Recommendation
Non-suspected case	Your inputs suggest that at present you are safe. In case any new symptoms develop, revisit our Symptom Checker to get recommendations.
Suspected Asymptomatic case	Recommended home quarantine for 14 days. In case any new symptoms develop, revisit our Symptom Checker to get recommendations
Suspected Symptomatic case	Recommended home quarantine for 14 days. In between, if symptoms develop, take an appointment in the nearest hospital and do PCR swab test.
Probable case	Immediately proceed to the nearest hospital and do PCR swab test.
Confirmed case	Strict home quarantine is advised, if symptoms develop, immediately proceed to the emergency.

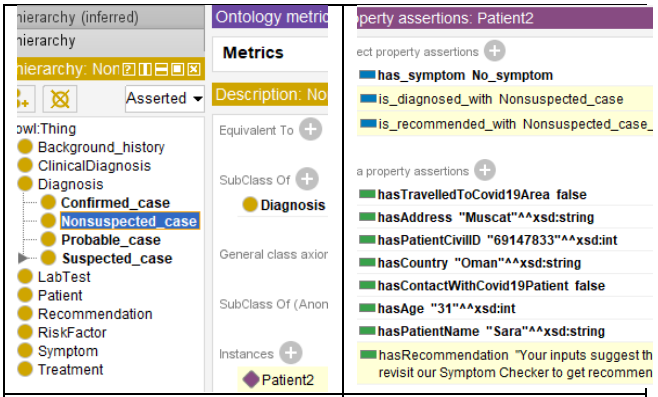
TABLE III. A SAMPLE SQWRL QUERY

SQWRL
$Confirmed\_case(?p, patient)$ ^ $hasPatientCivilID(?patient, ?CivilID)$ $hasAddress(?patient, ?Address)$^ $hasAge(?patient, ?Age)$ $hasGender(?patient, ?Gender)$ $-> sqwrl:select(?CivilID, ?Gender, ?Age, ?Address)$

IV. IMPLEMENTATION AND RESULTS

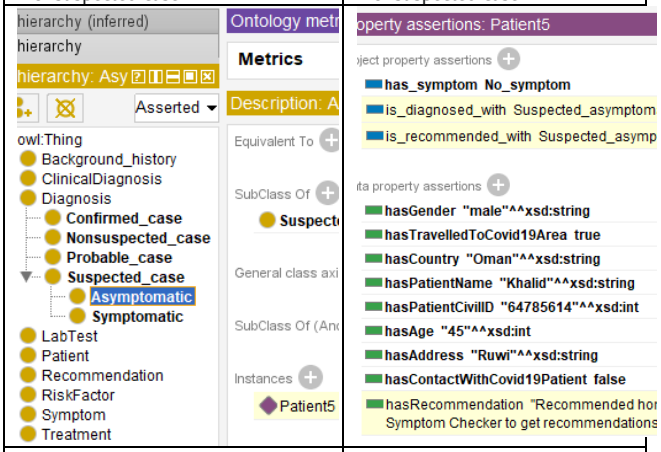
This section describes the implementation and results of the semantics of the knowledge base. Protégé 5.5.0 is used to implement the ontology [16]. Reasoners play a critical role in interpreting the semantics of ontologies and instances. The explicit facts are directly asserted in the knowledge base through properties. The semantics of these facts are interpreted by the inference mechanism of the knowledge bae. Reasoners perform the inference and extracts additional information from the knowledge base. Here rule-based reasoning is used in which the reasoner interprets the logical rules along with the asserted facts in the knowledge base to extract new information.

Forward chaining inference method is used here to add all the implied facts to the knowledge base. Reasoner do the reasoning in a forward-fashion, by considering the facts and the rules in the knowledge base and infer the new facts they imply. Whenever a new fact is inferred, it may lead to the inference of other facts. Pellet reasoner is used to reason the knowledge base. Fig. 6 shows the inference results based on SWRL rules given in Table I and Fig. 7 display the results of the query given in Table III.



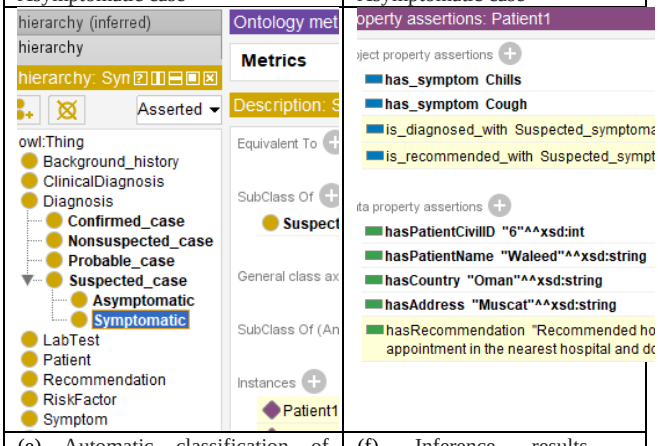
(a) Automatic classification of Nonsuspected case

(b) Inference results – Nonsuspected case



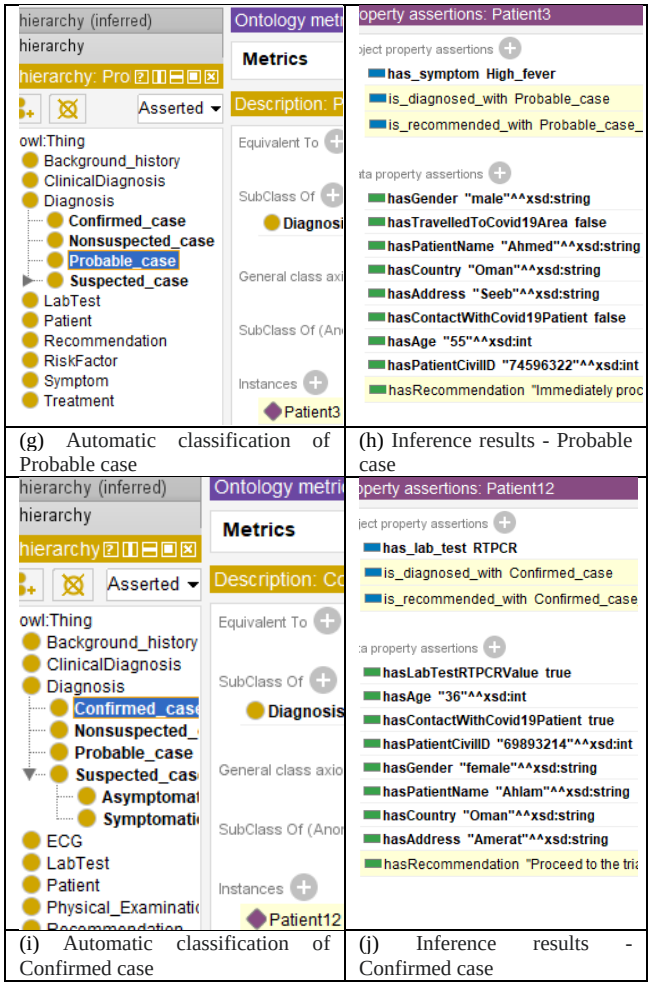
(c) Automatic classification of Asymptomatic case

(d) Inference results - Asymptomatic case



(e) Automatic classification of Symptomatic case

(f) Inference results - Symptomatic case



(g) Automatic classification of Probable case

(h) Inference results - Probable case

(i) Automatic classification of Confirmed case

(j) Inference results - Confirmed case

Fig. 6. Inference Results.

CivilID	Gender	Age	Address
72458963	male	55	Mattrah
70124589	male	35	Mattrah
69893214	female	36	Amerat
71258896	male	62	Baushar

Fig. 7. SQWRL Query Results of All Confirmed Cases.

V. Discussion

This section discusses the results of the reasoning process to infer the recommendations for different types of patient cases. The necessary conditions to infer the patient cases were written using SWRL as explained in the previous sub-sections. As shown in Fig. 8, Patient2 instance is created and the values of different properties (object and data) are asserted.

As per the SWRL rules given in Table I, the reasoner infers two facts initially for this patient instance. First fact is the diagnosis of this case as a non-suspected one. As a result, the instance is automatically inferred under the Non_suspected_case class using the object property is_diagnosed_with [Fig. 6a]. Next one is the corresponding recommendation for these types of cases. The property is_recommended_with is used to assign the recommendation [Fig. 6b]. The addition of the above inferred fact to the knowledge base leads to the automatic inferring of another fact, the final recommendation given as the value of has Recommendation property [Fig. 6b].

Similarly, the inference results of suspected asymptomatic cases are shown in Fig. 6(c) and (d). The instance Patient5 is automatically inferred under the Asymptomatic subclass. The inference results of suspected symptomatic cases are shown in Fig. 6(e) and (f). The instance Patient1 is automatically inferred under the Symptomatic subclass. Fig. 6(g) and (h) shows the inference results of probable cases. The instance Patient3 is automatically inferred under the Probable class. The confirmed case inference is shown in Fig. 6(i) and (j). The instance Patient12 is automatically inferred under the Confirmed class. The final recommendations are also shown in each of the above cases.

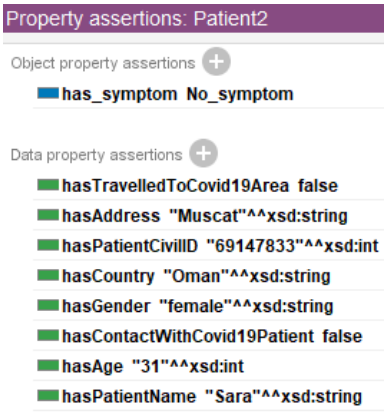


Fig. 8. Property Assertions of a Non-Suspected Case (Patient2) – Object Properties are Given in the Top Panel and Data Property Assertions are Given in the Bottom.

VI. Performance Metrics

In this section, we explain the performance of the ontology in terms of reasoning time and the execution time of queries. The reasoning task of classifying the ontology was done using Pellet reasoner. Six SQWRL queries corresponding to different category of patients were executed on a machine of configuration 8GB RAM and i7 processor. Fig. 9 shows the execution time of different queries in Protégé.

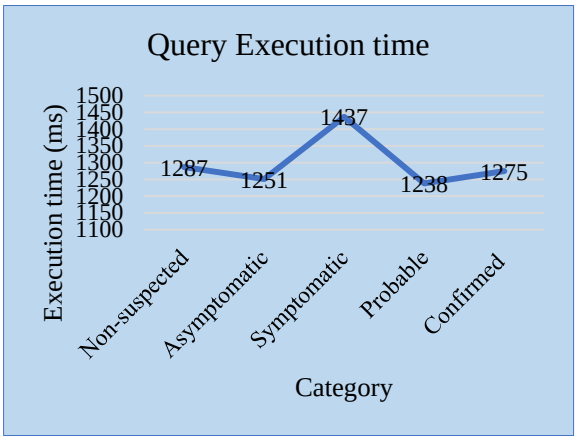


Fig. 9. Execution Time.

The developed ontology was loaded with some input data. OWL 2 DL reasoner, Pellet was then used to reason the ontology. The ontology was processed by Pellet in 77 ms. Then from SQWRLTab, each of the six queries were selected and run. The total execution time of the six queries was 6488 ms and the average execution time was calculated as 1297.6 ms. This shows that the reasoning time and the query execution time is optimal.

VII. Conclusion and Future

Bio-medical ontologies play a significant role in the design and development of CDSS. In this paper, we have presented the development of an ontology (classes, properties, rules etc.) to represent the domain concepts of SARS-CoV-2 (COVID-19). We have presented the initial part of the ontology in this article, which categorize a person into any one of the categories -unsuspected, symptomatic, asymptomatic, probable and confirmed. SWRL rules are constructed to check the necessary condition for the classification and categorization of patients into the above categories. We constructed several queries using SQWRL to retrieve the information stored in the ontology. The intended ontology will be used in a CDSS, which is currently under development to support the medical practioners of satellite clinics of ROP, in diagnosing COVID-19 in Sultanate of Oman. The future work will be of two-fold: (a) to present the full ontological concepts related to Clinical Diagnosis, Lab Test, and Treatment and (b) the design and development of the above mentioned CDSS.

Acknowledgment

The research leading to these results has received Research Project Funding from The Research Council (TRC) of the Sultanate of Oman, under Commissioned Research Program, Contract No. TRC/CRP/AOU/COVID-19/20/13.

References

[1] World Health Organization. (2020). WHO Director-General's opening remarks at the media briefing on COVID-19-11 March 2020 [Online]. Available at https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

[2] WHO Coronavirus Disease (COVID-19) Dashboard. https://covid19.who.int/ Accessed on August 27, 2020 [Online]. Available at https://covid19.who.int/?gclid=CjwKCAjw5p_8BRBUEiwAPpJO6_XYkPgVKKebOYNIjjGeZ24nlBQqM_cMN5ZHFFW_mYUhOcogZe6o0RoCCrQQAvD_BwE.

[3] MOH Registers First Two Novel Coronavirus (COVID-2019) in Oman [Online]. Available at https://www.moh.gov.om/en/-/--1226 Accessed July 24, 2020.

[4] Akerkar, Rajendra. Foundations of the semantic Web: XML, RDF & ontology. Alpha Science International, Ltd, 2009.

[5] Gruber, T. "Toward Principles for the Design of Ontologies Used for Knowledge Sharing". International Journal of Human-Computer Studies. 43 (5–6): 907–928. doi:10.1006/ijhc.1995.1081. 1993.

[6] Alexander, C. Y. Methods in biomedical ontology. Journal of biomedical informatics, 39(3), 252-266 2006.

[7] Babcock, S., Cowell, L. G., Beverley, J., & Smith, B. The Infectious Disease Ontology in the Age of COVID-19, 2020 (preprint).

[8] Ashburner, M., Ball, C., Blake, J. et al. Gene Ontology: tool for the unification of biology. Nat Genet **25,** 25–29. https://doi.org/10.1038/75556, 2000.

[9] Schriml, Lynn Marie, et al. "Disease Ontology: a backbone for disease semantic integration." Nucleic acids research 40.D1: D940-D946, 2012.

[10] Donnelly, Kevin. "SNOMED-CT: The advanced terminology and coding system for eHealth." Studies in health technology and informatics 121:279 2006.

[11] Nitsuwat, Supot, and Wansa Paoin. "Development of ICD-10-TM ontology for a semi-automated morbidity coding system in Thailand." Methods of information in medicine 51.06: 519-528 2012.

[12] Jin, Z. Environment modeling-based requirements engineering for software intensive systems. Morgan Kaufmann. 2018.

[13] National Clinical Management Protocol for Hospitalized Patients with Covid-19. Available at http://ghc.sa/ar-sa/Documents/Oman.pdf. Accessed July 24, 2020.

[14] World Health Organization. Clinical management of severe acute respiratory infection (SARI) when COVID-19 disease is suspected: interim guidance, 13 March 2020 (No. WHO/2019-nCoV/clinical/2020.4). World Health Organization.2020.

[15] O'Connor, M. J., & Das, A. K. (2009, October). SQWRL: a query language for OWL. In OWLED, Vol. 529, No. 2009.

[16] Musen, M.A. The Protégé project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), DOI: 10.1145/2557001.25757003 2015.

A Conceptual Data Modelling Framework for Context-Aware Text Classification

Nazia Tazeen[1]*
PhD Scholar, Department of Computer Science and Engineering, SPMVV Tirupati, India

K. Sandhya Rani[2]
Professor, Department of Computer Science SPMVV Tirupati, India

***Abstract*—Data analytics has an interesting variant that aims to understand an entity's behavior. It is termed as diagnostic analytics, which answers "why type questions". "Why type questions" find their applications in emotion classification, brand analysis, drug review modeling, customer complaints classification etc. Labeled data form the core of any analytics' problem, leave alone diagnostic analytics; however, labeled data is not always available. In some cases, it is required to assign labels to unknown entities and understand its behavior. For such scenarios, the proposed model unites topic modeling and text classification techniques. This combined data model will help to solve diagnostic issues and obtain meaningful insights from data by treating the procedure as a classification problem. The proposed model uses Improved Latent Drichlet Allocation for topic modeling and sentiment analysis to understand an entity's behavior and represent it as an Improved Multinomial Naïve Bayesian data model to achieve automated classification. The model is tested using drug review dataset obtained from UCI repository. The health conditions with their associated drug names were extracted from the reviews and sentiment scores were assigned. The sentiment scores reflected the behavior of various drugs for a particular health condition and classified them according to their quality. The proposed model performance is compared with existing baseline models and it is proved that our model exhibited better than other models.**

Keywords*—Text classification; topic modeling; natural language processing; sentiment analysis; drug dataset; context-aware model; diagnostic analytics; feature extraction

I. Introduction

Data analytics is a branch of data mining, that deals with extracting useful information from the data. There are four types of analytics, Predicitve, Prescriptive, Diagnostic and Descriptive analytics. The concern of this work is centered on the concepts related to Diagnostic analytics. It provides answers to why something happened?. Why a person likes a particular product?, why a particular drug is harmful?, why a particular event occurred? are some issues that can be solved using this analytics.The application of such analytics operations in Natural Language datasets pose a greater challenge. There are three reasons for that, availability of labeled data regarding the entity of interest, extracting topics or coherent terms from documents and size of varying topics (saliency). To solve this, a system needs to include contextual information of texts [11]. The context-aware systems are more concerned about, topic relevance, term relevance, and topic labels when compared to bag-of words approach. Such a system capable of revealing systematic use cases of the business problems through conceptual models help to arrive at solutions easily [14]. The proposed conceptual framework is designed in such a way that, the model not only supports Natural Language Understanding(NLU) but also enables Natural Language Processing(NLP) in unstructured data environments. To identify entities of interest, a topic modeler is used and to study its behaviour to predict future scenarios, a machine learning data modeler is used. The combined model is built as a classifier, since almost all analytical problems can be represented as a classification problem and also building a classifier model will make the system generic for all similar use-cases in future. For topic modeling and labeling uknown entities, the famous conventional Latent Drichlet Allocation technique is improved. It is enhanced to handle sparse features, extracts latent semantic relationships from the data, keywords for topics of all sizes and minimize polysemy issues.

Also, to answer the why type questions taken for consideration, the setimental value of text is also measured. Sentiments could represent the emtional reaction of a person, an operational failure of a component, a side effect of a drug based on the application area. Using the sentiment analysis, the entity's behaviour could be reasoned out. This additional information when used along with a Machine Learning(ML) data model, can further improve the accuracy of classification results. The proposed model improves the exisiting Multinomial Naïve Bayes using sentiment analysis scores and and latent topics and builds a classifier to automate analysis.

The result of the analysis could be used for efficient decision making in medical diagnosis, manufacturing error probes, automobile efficiency monitioring etc. The proposed system is tested using drug review dataset obtained from UCI repository. The health conditions with their associated drug names were extracted from the reviews and sentiment scores are assigned. The sentiment scores reflected the behavior of various drugs for a particular health condition and classified them according to their quality. Quality is decided based on its positive and negative side effects.

The aim is to extract disease names mentioned in drug reviews and find out the most suitable drug names for each ailment. Understanding the drugs based on their positive and negative feedbacks reflect the reason why a particular drug is suitable for a particular health condition and thus satisfies the aim of the study. The proposed model was compared with existing baseline models and found that our model exhibited better performance.

Corresponding Author

II. Literature Review

The advent of e-commerce not only replaced the physical shopping experience but also altered the word of mouth spread regarding goods of interest. People started commenting about their most favorite and not so favorite items. User reviews have almost become store house of user preferences. This work uses topic modeling with LSTM to understand review comments that improves the traditional topic clustering. However, the usage of smaller sliding windows failed to extract meanings of longer sentences leading to ambiguity [12]. Lexical Selection applications are concerned about context information of sentences and topics present in documents separately. This work combines them to achieve machine translation by exploiting term correlations [8]. It uses statistical probability based Gibbs sampling technique to achieve hidden variable prediction using source information. The context level features are split into local and global to learn topic distributions in data. But the model did not take into consideration, the phrase level information and word level attributes for classification accuracy improvements.

Another work uses Gaussian Mixture Neural Topic Model for extracting contextual information by using multi vector clusters [13]. The terms in each topics and sentences are jointly modeled. This order sensitive and context aware system can extract effective topics along with appropriate features. It does not consider polysemy of words and also manually detect topics which are time consuming. Topic identification in dialog systems is an interesting application of Topic Modeling [2]. It is applied in human and Chabot systems. It uses contextual key words related to each topic and the conversational features to achieve topic clusters with Deep Average and Contextual Attention Deep Average Networks. The coherent dialog process is easily achieved using this combined technique to annotate topics of interest. Unsupervised variant was not tested using this method.

Another interesting direction in modeling sentiment analysis is establishing relationship between ratings and sentiments of yelp reviews [6]. This helps to achieve a quantitative value for comments in better classifying favorite restaurants. The model was tested using Support Vector Machine algorithm in a binary classification setup. The model however, failed to model other rating values such as 3, 4, and 5 which is the norm in most rating systems. Also multiclass classification was not possible. To identify really helpful reviews from the available reviews is a challenging task for users. Most users rely on ratings to filter useful reviews [SA2]. The model uses TF-IDF for feature extraction which does not consider order of terms related to topics. Hence, the obtained topics will not be effective.

The Review Rating Prediction and Review Text Content analysis were combined to predict user preferences. The role of non-rated reviews is analyzed in this work [7]. It uses sentiment analysis of aspects to predict the ratings for all such non- rated reviews. This helps to identify most liked products in a quantitative manner. Usage of sentiments extracts contextual information also. To achieve this the model used Sentiment Base Conditional Random Fields to measure term co-occurrence. The model requires more training data for effective functioning which is not always possible. Another similar work analyses the multi aspect labeling of sentences and rating prediction for topic modeling [1]. It employs a semi-supervised approach for Perceptron Ranking technique. It is however, a weak prediction model not applicable for other stronger prediction models like support vector regression. In one more work related to rating analysis, physiological signals and reviews obtained from other data sources were used [10]. The global rating was assigned to global reviews using NLP techniques along with Electro encephalogram signals were recorded for each product. Sentiment analysis, Random forest with regression and Artificial Bee Colony algorithm is used to model global and local rating. But real emotions of patients were not considered. The work models the most subjective sentences in the document and measures its aspect sentiment orientation [17]. WordNet based clustering helps to analyze sentiment of clusters. Usage of hand coded rules, bigger training data are drawbacks of the model. Aspect is termed as topics.

An information retrieval based text classification is tried in this work to predict the rating value using the user comments [5]. By involving overall satisfaction of the customers through ratings and a clear description of why they have reacted in such a way to a product a justifiable decision making process is obtained. The classifier is modeled as supervised with vector space model and sentiment analysis to predict ratings on the scale of one to five. No usage of advanced NLP techniques was found. One major drawback in analyzing the sentiments of reviews, lies in the technique of modeling long term dependencies which are so complicatedly found in the short text reviews [19]. A deep sequential model is built for sentiment analysis using Gated RNN for dependency between sentences and LSTM to vectorize the sentences. It works on the basis of principle of compositionality which states that summarized meaning of reviews reflect actual concepts mentioned in the reviews. The model was not tested for document level analysis and its applicability for longer reviews was not available.

The mining of electronic patient records for decision-making and diagnostics requirements is a prominent issue in text classification [18]. The discharge summaries of patients were morphologically analyzed for extracting features with the help of two-dimensional attribute mapping through correspondence analysis. The class labels were assigned after obtaining the keywords and rank them accordingly. The distance between the data points and class labels are measured and those with shortest distance were assigned the corresponding label. It was found that decision trees though capture the structure of data, performed poor.

Another similar work on the same kind of data used statistical techniques, rules of associations and Extreme learning machine encoder for extracting patient specific features [16]. It works on the principle of sentence level sentiment analysis. However, it requires the sentences to be subjective and availability for a lexicon specific dictionary is scarce. In vehicle fault diagnosis applications, the text classification system can be used along with machine learning and search-prompt techniques [20]. The diagnostic codes are integrated with term weight matrix to obtain similarity scores

between documents and labels. Latent Semantic Indexing is found not so suitable for this kind of analysis. The diagnostic codes have to manually obtained and integrated with the system. In the place of diagnostic codes, the feedback or review data could be used which will be more effective.

Problems to address in the proposed work:

- Extracting topics present in random texts suffers from, high-frequent keyword elimination, subjectivity of sentences and lack of quantitative value of texts.
- Sentiment Analysis without rating information leads to imprecise classification results.
- Use of conventional TF-IDF is not suitable for opinion mining due to long term dependency in sentences and shorter expressions.

The proposed classifier model attempts to solve the above problems using a combined approach of topic modeling and text classification based on sentiment value of documents. The detailed approach is explained in the following sections.

III. Proposed Methodology

The proposed model is built as a two stage classifier for multi class classification in opinion mining problems. The aim is to extract entities of interest from the 'review' or 'comments' data. Then based on the subjectivity, the sentiment values are assigned to sentences that in turn make up the topics. The knowledge about sentiments related to topics reiterates the originality of ratings provided by the people. The proposed framework consists of a topic modeler, that extracts latent topics from the reviews and its behavior is studied using the sentimental values and ratings provided by the user. In the next step, a ML classifier data model is built for effective classification requirements off the data. The proposed framework is given in Fig. 1.

According to Fig. 1, the drug reviews dataset was retrieved from UCI machine learning repository. The dataset is a collection of names of drugs, its related conditions, review regarding the side effects and ratings. The aim is to extract unique drug names and health conditions using topic modeling techniques and in the next phase assign sentiment tags to each drug. The name of drugs mentioned in the dataset is not related to one particular health condition or disease, but generic opinions of patients that took the drugs. If the system could filter each disease along with its side effects, name of drug and sentiment value, it would assist the medical practitioners is swift diagnosis of patients.

In the initial data pre-processing stage, the dataset is thoroughly studied. The exploratory data analysis was done to extract complete information of the drug reviews which is spread over 2,15,0633 instances containing attributes such as names of drugs, type of health conditions, review of patients, ratings, date of review and count of users that found review helpful.

A. Preprocessing

The records were checked for null entries and it was not found anywhere in the dataset. The identifiers given to each reviews are unique. Regular expression rules were manually constructed to remove 'non-English characters', 'symbols', 'stop words' and 'upper case' characters. The average length of reviews is found to be around 500. The reviews were cleaned and stored in a separate column; this preprocessing is enough for the sentiment tagging stage. For topic modeling, more steps were carried out in cleaning such as stemming, lemmatization and tokenization. The corrupted reviews that were of no use for others were removed and accounted for 1171 reviews. Averagely, the number of reviews for a particular drug is found to be 3658 and that for a particular health condition is around 836. The top ten most reviewed drug names are plotted as a bar graph and can be found in Fig. 2.

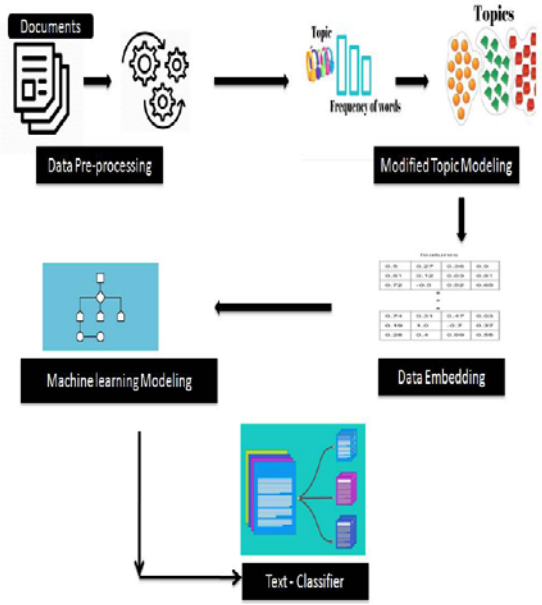


Fig. 1. Improved Multinomial Naive Bayesian Framework for Text Classification.

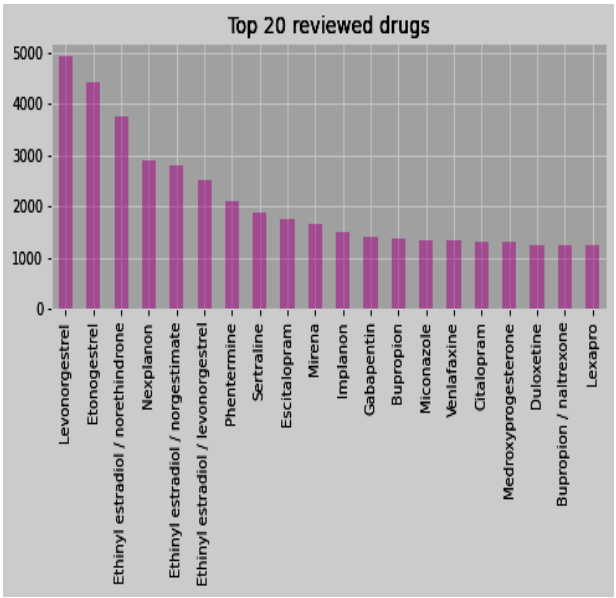


Fig. 2. Most Reviewed Drug Names.

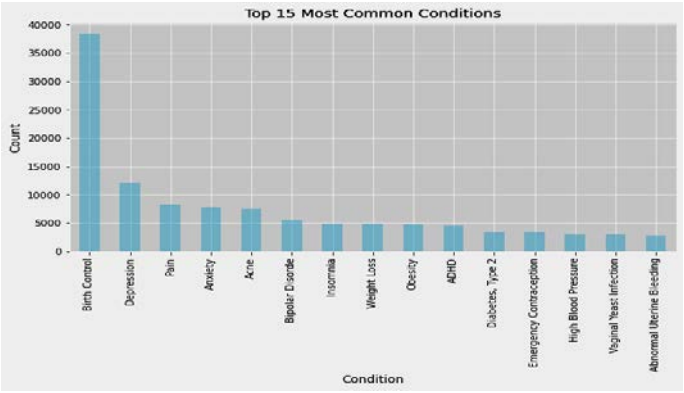


Fig. 3. Most Reviewed Health Conditions.

The statistics were extracted as part of the exploratory data analysis using count vectorizer function. The dataset was analyzed to obtain the most common health conditions of people. From the Fig. 3, we could see that birth control and depression are most talked about health conditions of people.

B. Modified Topic Modeling

The topic modeling is carried out using Latent Drichlet Allocation (LDA) algorithm as it does not require prior annotations. A group of words describes each topic or health condition. The application of LDA for text classification is a rare scenario especially in healthcare. The proposed LDA is an improvement over the existing model, wherein, sentiment analysis and rating information is included to model the topics apart from the general approach to topic modeling. The latent nature of topics could be substituted by using this technique to provide more accurate results. From the exploratory data analysis, it was identified that, on an average, there exist 3658 reviews for a particular drug and 836 reviews per health condition. The conditions for which only one drug name was given were eliminated. The unwanted 1171 corrupted reviews were also removed. The aim is to extract disease names mentioned in drug reviews and find out the most suitable drug names for each ailment.

All the disease names were selected as topic labels. Some of the topic labels are, 'Birth Control', 'Depression', 'Acne', 'Pain', 'Anxiety', 'Bipolar Disordered. For each of these diseases it was found that on an average 830 reviews existed. The topics were modeled as clusters and for each clusters 50 terms were set to maximum. The LDA generates probabilistic topic terms based on the concepts they share. The primary assumption is regarding the topics that are latent in the dataset, denoted by some constant say 'T'.

There exist 'N' number of documents for each latent topic which is nothing but a polynomial distribution over its constituent terms. It is assumed again that the 'T' latent topics generate the documents. It is important to extract the information that are hidden in the documents. This information is termed as, composition data of the unknown topic denoted by (α, x). The parameters of Richet can be mentioned as, β and λ. The distribution of random variables in a given document 'Doc' can be computed using the probability distribution,

$$P(\alpha, x, t \mid \beta, \lambda) = P(\alpha, \beta)\prod_{i=1}^{Q_e} P(x|\alpha)P(t|\alpha,\beta) \quad (1)$$

The probabilistic values are approximated using Gibbs Sampling. The next step is topic or word embedding, since words or terms constitute topics. Instead of using TF-IDF which removes repeated terms irrespective of its contribution to a particular topic, we have used Topic2vec[11].Topic2vec is used with a modification in this improved-LDA, where, the high frequent words in a particular topic are replaced with their single entry obtained from the initial iterations and assigned to topics with the same probability values. However, there is presence of sparse features due to unequal topic representation. It can be removed by appending a medical thesaurus with the topic dictionary through an index. Such a thesaurus is constructed using medical articles scraped from web and tokenizing words related to topics found in our dataset. This not only removes sparse feature issues but also minimizes polysemy. When using Topic2vec model, the keywords of all sizes a treated as a fixed length vector, thereby solving the lengthy keyword problem. Thus, using these techniques, the various disease names are obtained as topics along with their constituent terms.

Followed by this, the Sentiment analysis is carried out to identify positive and negative reviews for a particular drug. Using the sentiment polarity of each review, the corresponding drug name is chosen as the best and worst drug for a particular disease obtained in the previous stage. Based on the review comments' polarity and corresponding rating values, the most common health conditions of people are identified. Here, the sentimental value of reviews helps to identify both a particular health condition and its appropriate drug. The useful count of reviews was included to analyze the sentiments. The Harvard polarity dictionary was used to assign polarity to each review using useful count. The drugs for each health condition were discovered using a review's useful count and sentiment value obtained.

The rating information is appended to the drug and health condition names along with its sentiment to form a semantic knowledge base for the subsequent processing. After identifying the disease names, respective drugs and the rating information from the reviews are indexed in a topic dictionary. This dictionary helps to identify drugs for various diseases quantitatively. Using this knowledge, the topics and their most likely terms are represented as a dataset for further processing. From the topic dictionary, the top-10 drugs for the sample health conditions, 'Birth Control', 'Depression', 'Acne', 'Pain', 'Anxiety' and 'Bipolar Disorder' are given in Table I. The novelty is to extract health conditions from drug reviews and identifying the corresponding drug names through topic modeling. Understanding the drugs based on their positive and negative feedbacks reflect the reason why a particular drug is suitable for a particular health condition and thus satisfies one of the aims of the study.

TABLE I. TOP-10 DRUGS REVIEWED FOR MOST COMMON HEALTH CONDITIONS

BirthControl	Depression	Acne	Pain
Plan B	Niravam	Milk of Magnesia	Ketoprofen
Femcon Fe	Serzone	Benzaclin	Acetaminophen / phenyltoloxamine
Ortho-Novum 7 / 7 / 7	Alprazolam	Magnesium hydroxide	Dolophine
Kyleena	Parnate	Benzoyl peroxide	Fentora
Ortho Evra	Xanax	Phillips' Milk of Magnesia	Lodine
Levonorgestrel	Xanax XR	Benzoyl peroxide / erythromycin	Methadone Diskets
Lybrel	Nefazodone	Retin-A	Proctofoam

C. Improved Naïve Bayesian Classifier

The model is further enhanced as a generic text classifier capable of solving any type of opinion based text classification problem by introducing a machine learning data model. To achieve this, we use a variant of Bayesian algorithm, Multinomial Naïve Bayes (MNB). This algorithm suffers from, independency between the features of its classes. The context information of classes is disregarded in the process [15]. To avoid this, the Bayesian procedure has to include some form of dependency between its features since, text data is heavily depended on its neighborhood information. To attain this functionality, the information obtained from the I-LDA phase is used for training the MNB to finally derive Improved MNB. This technique is named as I-MNB throughout the article.

The probability distribution of topic terms is already available from the improved LDA output. For each topic, here we have the name of diseases as our topic. For each topic, the topic terms probability is taken from the previous phase, its again affixed with sentiment scores for each topic. These scores are adjusted relative to the intra topic terms. For example, the sentiment value for a particular drug that receives highest positive score will be the most suitable drug for that disease. The rest of the drugs will be scored using relative scoring technique to order them in descending order according to the relative scores they obtained. This relative score establishes a dependency among terms in the topics, thereby capturing the optimal Bayesian network. The classification result is improved through this representation, since; it gives a mathematical explanation to the classification results. In other words, it tells why a certain drug is prescribed over the others.

Probability of each term per class is give as:

$$P(\alpha, x, t \mid \beta, \lambda) = P(\alpha, \beta)\prod_{i=1}^{Q_e} P(x|\alpha)P(t|\alpha,\beta) \quad (2)$$

The sentiment weighted terms for a class is given as

$$P(\alpha, x, t \mid \beta, \lambda) = P(\alpha, \beta)\prod_{i=1}^{Q_e} P(x|\alpha)P(t|\alpha,\beta) + |\text{senti}(T_i)| \quad (3)$$

$|\text{senti}(T_i)|$ denotes the absolute value of sentiment score for a given class, it is calculated using subjectivity analysis. Here, we have not included the rating information since, ratings are not always available for all entities. The sentiment of the drug reflects the sentiment of the opinions of the patient or users that consumed them. Also, the aggregated sentiment value of a review is the sentiment of the drug, therefore the neighborhood information of drug reviews decides the overall sentiment of the drug. This captures the dependency among features related to a particular drug. The dependency between other drugs for the dame disease can be obtained through comparing the sentiment scores of each drug and ranking them relatively.

$$\text{senti}(T)=[\text{senti}(t_1),\text{senti}(t_2),\text{senti}(t_3),\ldots\text{senti}(t_n)] \quad (4)$$

Relative Scoring,

$$\max [\text{senti}(T)] = \text{senti}(t_1) > \text{senti}(t_2) > \text{senti}(t_3) > \ldots\ldots\ldots\ldots\ldots\ldots > \text{senti}(t_n) \quad (5)$$

The sentiment scores are normalized using laplace smoothing.

$$\text{senti(i|j)} = \frac{term_{ij}+\theta}{topic_j + |senti| + 1}, \theta = 0.001 \quad (6)$$

senti is the sentiment score of all terms in the topic. Now the combined probability distribution of a topic 'T' can be obtained using the distribution over its terms.

For a given topic j and term i, at the term frequency denoted by tf,

$$P(j) \propto \mu_j \prod_{i=1}^{|senti|} p(i|j)^{tf_i} \quad (7)$$

Taking log on the probability distribution will avoid term overflow for topics

$$P(j) \propto log\,(\,\mu_j \prod_{i=1}^{|senti|} p(i|j)^{tf_i}) \quad (8)$$

$$P(j) = log\mu_j + \sum_{i=1}^{|senti|} tf_i \log(P(i|j)) \quad (9)$$

Thus, the optimal I-MNB model can be obtained as,

$$P(j) = log\mu_j + \sum_{i=1}^{|senti|} \log(1 + tf_i)\log(P(i|j)) \quad (10)$$

The proposed model with the necessary improvements, have given impressive results in the experimental analysis, which will be discussed in the next section.

IV. PERFORMANCE EVALUATION

The model is tested using drug review dataset. The two-phased approach is built as a multi class text classifier to be generically able to deal with opinion mining problems in similar applications. This model is compared with other benchmark algorithms in text data classification. Some dataset-oriented changes were made appropriate to each of the benchmark algorithms, such as, discrete to continuous features, binary to multiclass, numerical categorical conversions, normalizations etc. The benchmark algorithms taken for the experimental study are, Linear Support Vector Classification [9], Logistic Regression [3] and Random Forest classifier models [4].

The performance is validated through the score of accuracy of the model. It is the measure of ratio between total predictions obtained correct and sum of predictions made in the entire dataset. Accuracy is not suitable for problems with

imbalanced class distributions. Hence, along with accuracy, it is required to measure other metrics such as 'Precision', 'Recall' and 'F1-score.

A. Precision

Precision verifies the number, percentage, or value of True Positive (TP) predictions for all classes to the True Positive and False Positive (FP) predictions of all classes in the dataset [15]. It is denoted by the formula,

$$\text{Precision} = \text{Sum}(TP(cs)) / \text{Sum}(TP(cs) + FP(cs)) \quad (11)$$

Where, cs denotes classes in the dataset

B. Recall

Recall verifies the number, percentage, or value of True Positive (TP) predictions for all classes to the True Positive and False Negative (FN) predictions of all classes in the dataset [15]. It is denoted by the formula,

$$\text{Recall} = \text{Sum}(TP(cs)) / \text{Sum}(TP(cs) + FN(cs)) \quad (12)$$

C. F1-Score

To improve precision and recall, the tweaking of one measure might increase or decrease another. To avoid this F1-score is used. It summarizes the overall performance of a system, by deriving the harmonic mean of both the precision and recall results [15].

$$\text{F1-score} = (2 * Pre * Rec) / (Pre + Rec) \quad (13)$$

The combined Precision, Recall and F1-scores obtained for the proposed model is compared with the benchmark algorithms and plotted as bar graph.

From the Fig. 4, we can see that the performance of our proposed I-MNB is higher than the other algorithms in all three metrics. This is due to better knowledge representation rendered by the I-LDA phase. The Random Forest Classifier though almost matches the proposed model's perfromance, it falls short in Recall values drastically. According to Eqns (10, 11 and 12) it is also proved that our model did not overfit the data by showing near perfect perfromance values.

D. Accuracy

Accuracy is the ratio of True Positive results and the entire number of results obtained for the given dataset [15]. It is denoted by,

$$\text{Acc} = (TP+TN)/(TP+FP+FN+TN) \quad (14)$$

From the Fig. 5, we can see that the performance of our proposed I-MNB is higher than the other models in accuracy. The accuracy achieved for our system is 91%. It can be noted that, the accuracy of Linear SVC falls way below other models, stating that this algorithm is not suitable for opinion mining problems.

E. AUC –ROC

Receiver Operating Curve and the relative Area Under Curve (ROC-AUC) state the separation between positive and negative classes [15]. The number of positive classes is known as True Positive Rate(TPR) and that of the negative classes are known as False Positive Rate(FPR) and range of separation is measured using various threshold values. It is a plot between TPR and FPRin other words sensitivity and 1-specificity denotes.

Sensitivty = Recall = Sum (TP (cs)) / Sum (TP (cs) + FN (cs))
1- Specificity = FP/(TN+FP)

From the Fig. 6 we can see that the pink lines represented by our proposed model I-MNB outperformed the others. Though the models LinearSVC and Random Forest Classifier converge with our model it is mainly with increasing thresholds which is not advisable since, it will drastically decrease the FP rate and increase FN rate simultaneously. The increase is not gradual. Also, the area occupied under the curve is still higher for our proposed model.

From the Fig. 7 we can see that AUC-ROC values obtained for some of our sample classes by the proposed I-MNB. The model has shown impressive results in all the six classes with the 'Pain' class topping the accuracy. It can also be seen that 80% of classes obtained better results and the remaining classes obtained satisfactory results. This is due to the people reviewing the particular disease and its constituent term ambiguity. Overall the model shows that the area occupied under the curve is still commendable for a multiclass text classifier.

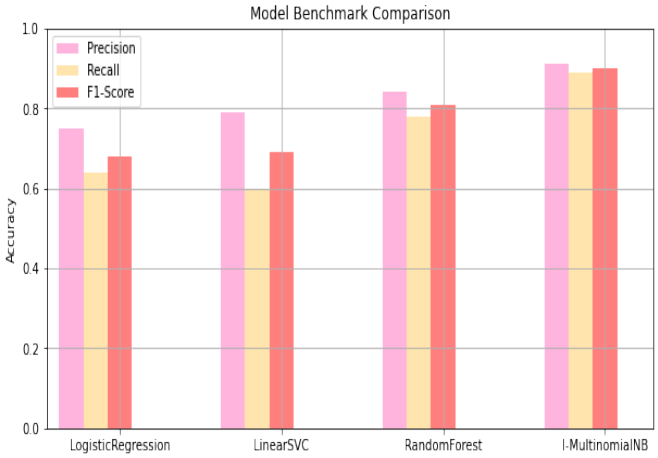


Fig. 4. Model Benchmark Comparison Results.

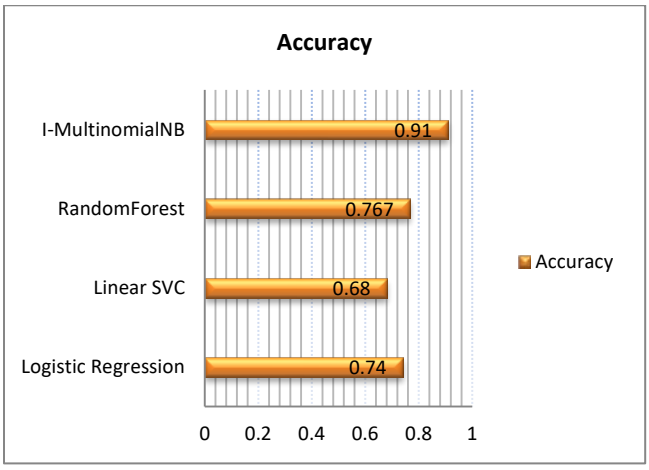


Fig. 5. Model Benchmark Comparison Results for Accuracy.

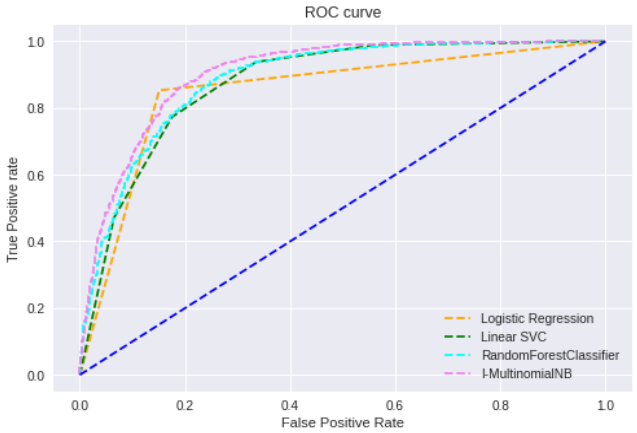


Fig. 6. AUC-ROC of Benchmark Models.

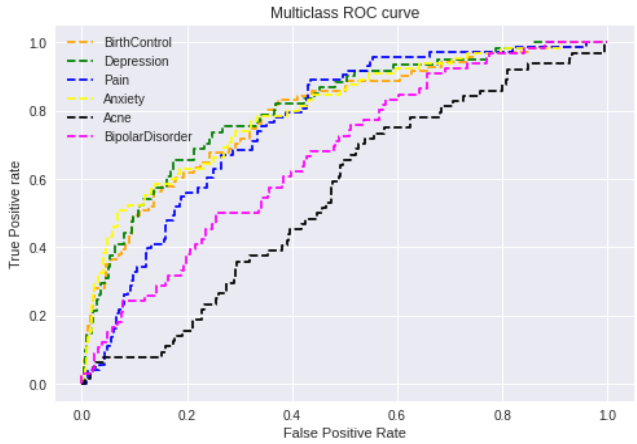


Fig. 7. AUC-ROC of Classes in I-MNB.

V. Discussion

The proposed model is built as a two phase classifier for multi class classification in opinion mining problems. The aim is to extract entities of interest from the 'review' or 'comments' data. Then based on the subjectivity, the sentiment values are assigned to sentences that in turn make up the topics. All the disease names were selected as topic labels. Some of the topic labels are, 'Birth Control', 'Depression', 'Acne', 'Pain', 'Anxiety', 'Bipolar Disordered. For each of these diseases it was found that on an average 830 reviews existed. Instead of using TF-IDF in LDA which removes repeated terms irresepctive of its contribution to a particular topic, we have used Topic2vec to improve LDA. The repplacement of high frequency words with their base words, sparse features minimization using medical thesaurus, polysemy representation and semantic modelling of knowledge base of classifier have shown remarkable perfromance compared to other baseline models. Theknowledge about sentiments related to topics reiterates the originality of ratings provided by the people in identifying the best drug for a health condition. The novelty in combining LDA and MNB and using sentiment scores to represent class features with neighborhood dependency improved the existing MNB to suit the needs of opinion mining problems. The accuracy obtained through various metrics is around 91%, which is remarkable, given the complex nature of test data and presence of multiple classes.

VI. Conclusion

The aim of the work is to extract entities of interest from the 'review' or 'comments' data. Based on the subjectivity, the sentiment values are assigned to sentences that in turn make up the topics. Some of the topic labels are, 'Birth Control', 'Depression', 'Acne', 'Pain', 'Anxiety', 'Bipolar Disordered. The replacement of high frequency words with their base words, sparse features minimization using medical thesaurus, polysemy representation and semantic modelling of knowledge base of classifier have shown remarkable perfromance compared to other baseline models. The knowledge about sentiments related to topics reiterates the originality of ratings provided by the people in identifying the best drug for a health condition. The novelty in combining LDA and MNB and using sentiment scores to represent class features with neighborhood dependency improved the existing MNB to suit the needs of opinion mining problems. The accuracy obtained through various metrics is around 91%, which is remarkable, given the complex nature of test data and presence of multiple classes. Also the AUC, ROC, Precision, Recall, F1-score values for the proposed system is obtained in the range above 90% signifying 10 to 12% improvement over the similar benchmark models. In future, other opinion or review mining problems can be considered with ensemble algorithms and different applications.

References

[1] Kaporo, H. (2019, April). Cross-collection Multi-aspect Sentiment Analysis. In Computer Science On-line Conference (pp. 107-118). Springer, Cham.

[2] Ahmadvand, A., Choi, J. I., & Agichtein, E. (2019, July). Contextual dialogue act classification for open-domain conversational agents. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1273-1276).

[3] Wu, T., Huang, Q., Liu, Z., Wang, Y., & Lin, D. (2020, August). Distribution-balanced loss for multi-label classification in long-tailed datasets. In European Conference on Computer Vision (pp. 162-178). Springer, Cham.

[4] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., & Androutsopoulos, I. (2019). Large-scale multi-label text classification on eu legislation. arXiv preprint arXiv:1906.02192.

[5] E. M. Alshari, A. Azman, and N. Malukssthaeprh, “Prediction of Rating from Comments based on Information Retrieval and Sentiment Analysis,” p. 5, 2016.

[6] Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: a review. Artificial Intelligence Review, 53(6), 4335-4385.

[7] Xu, H., Liu, B., Shu, L., & Yu, P. S. (2019). Bert post-training for review reading comprehension and aspect-based sentiment analysis. arXiv preprint arXiv:1904.02232.

[8] J. Su et al., “A Context-Aware Topic Model for Statistical Machine Translation,” in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, 2019, pp. 229–238, doi: 10.3115/v1/P15-1023.

[9] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.

[10] Kumar.S, Yadav.M & Roy. P.P (2018). Fusion of EEG response and Sentiment Analysis of Product Reviews to Predict Customer Satisfcation. Information Fusion.https://doiorg/10.1016/j.inffus. 2018.11.001.

[11] Li, W., Matsukawa, T., Saigo, H., & Suzuki, E. (2020, May). Context-Aware Latent Dirichlet Allocation for Topic Segmentation. In Pacific-

Asia Conference on Knowledge Discovery and Data Mining (pp. 475-486). Springer, Cham.

[12] M. Jin, X. Luo, H. Zhu, and H. H. Zhuo, "Combining Deep Learning and Topic Modeling for Review Understanding in Context-Aware Recommendation," p. 10.

[13] M. Yang, T. Cui, and W. Tu, "Ordering-Sensitive and Semantic-Aware Topic Modeling," p. 7.

[14] Nalchigar, S., & Yu, E. (2018). Business-driven data analytics: a conceptual modeling framework. Data & Knowledge Engineering, 117, 359-372.

[15] Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

[16] S. A. Waheeb, N. A. Khan, B. Chen, and X. Shang, "Machine Learning Based Sentiment Text Classification for Evaluating Treatment Quality of Discharge Summary," p. 16, 2020.

[17] S. Gojali and M. L. Khodra, (2016) "Aspect Based Sentiment Analysis for Review Rating Prediction," p. 6.

[18] S. Tsumoto, T. Kimura, H. Iwata, and S. Hirano, "Mining Text for Disease Diagnosis," Procedia Computer Science, vol. 122, pp. 1133–1140, 2017, doi: 10.1016/j.procs.2017.11.483.

[19] S. Verma, M. Saini, and A. Sharan, "Deep Sequential Model for Review Rating Prediction," p. 6, 2017.

[20] Y. L. Murphey, L. Huang, H. X. Wang, and Y. Huang, "Vehicle Fault Diagnostics Using Text Mining, Vehicle Engineering Structure and Machine Learning," International Journal of Intelligent Information Systems, p. 13.

Smart Start and HER for a Directed and Persistent Reinforcement Learning Exploration in Discrete Environment

Heba Alrakh[1], Muhammad Fahmi Miskon[2]
Center for Robotics and Industrial Automation
Fakulti Kejuruteraan Elektrik
Universiti Teknikal Malaysia Melaka
76100 Durian Tunggal, Melaka, Malaysia

Rozilawati Mohd Nor[3]
Center for Robotics and Industrial Automation
Fakulti Teknologi Kejuruteraan Elektrik dan Elektronik
Universiti Teknikal Malaysia Melaka
76100 Durian Tunggal, Melaka, Malaysia

***Abstract*—Reinforcement learning (RL) solves sequential decision making problems through trial and error, through experiences can be amassed to achieve goals and increase the accumulative rewards. Exploration-exploitation dilemma is a critical challenge in reinforcement learning, particularly environments with misleading or sparse rewards which have shown difficulties to construct a suitable exploration strategy. In this paper a framework for Smart Start (SS) and Hindsight experience replay (HER) is developed to improve the performance of SS and make the exploration more directed especially in the early episodes. The framework Smart Start and Hindsight experience replay (SS+HER) was studied in discrete maze environment with sparse rewards. The results reveal that the framework doubles the rewards at the early episodes and decreases the time of the agent to reach the goal.**

***Keywords*—Reinforcement learning; hindsight experience replay; smart start; limit search space; exploration-exploitation trade off**

I. Introduction

People learn through interacting with the environment around them from their childhood where children's walking or trying to play, which is considered the major source of learning. The same in reinforcement learning, the machine is trying to interact with environment to collect information then use it to discover the best possible performance [1].

Exploration means learning new knowledge by trying new actions that the agent did not select before which might lead to a better action selection in the future causing increase in the accumulative reward [2]. In contrast, exploitation is using the same actions that the agent tried in the past and was effective in producing rewards in order to maximize the immediate reward. Excessive exploration will be wasting of time and cause less immediate reward because the agent might spend most of the time doing irrelevant actions or less reward actions. On the other hand, more exploitation will cause suboptimal solution. So the balance between both of them is becoming a critical challenge and an essential matter in order to get better results [3]. This is called exploration and exploitation dilemma.

There are many exploration strategies to solve this problem [4] but most of them are depending on collecting more data. A relatively new strategy is focusing on reducing the search space such as Proximal Policy Optimization (PPO) [5] and Trust Region Policy Optimization (TRPO) [6]. Although there are number of new limiting search space techniques, there are no experiments combining two techniques together.

The objective of this paper is combining Smart Start and Hindsight Experience Replay (HER) in order to reach a more efficient exploration. Smart Start guides an agent to a state where it supposes to discover the newest information, which is named the Smart Start state, S_{ss}. Smart Start does not modify the functionality of RL algorithm in which it is utilized with, yet it adds more persistent and directed exploration to the algorithm. On top of Smart Start strategy, a conceptually simple framework (HER) is added which utilizes experience to improve exploration by splitting the main goal to sub goals to learn from the previous errors.

The rest of this paper is organized as follows: Section II gives the description of the related work. Section III contains the background of the research. In Section IV, Smart Start and HER is discussed in details. Section V presents the experiments. In Section VI, the results are displayed. Lastly, Section VII is conclusion.

II. Related Work

Balancing between exploration and exploitation is solved by [7] which depends on Stratonovich's value of information which consists of two steps. The first one generates the base line of agent performance by measuring the achievable return of a policy in where there is no information regarding the states, afterward offsets these costs with a term that evaluates the average penalties when the state-action information is bounded above by a prescribed amount. Though, the optimization of value of information shows a softmax random exploration. Obviously, it depends on factors where the value factors is decided by human. Also it does not cover the multistate case so the improvement of optimal average cost can be achieved per episode.

Another alternative solution, by applying Bayesian deep Q-networks (BDQN) is an efficient Thompson sampling based method in high dimensional RL problems. In [8]

Azizzadenesheli and Anandkumar studied the behaviour of BDQN and compared it to another method to solve exploration – exploitation trade off. Yet the problem is this method itself is difficult in implementing and time consuming and did not provide a sample efficiency guarantee.

On the other hand, Lin et al. [9] used demonstration data to guide the exploration of agents at the beginning of training, which can help agents learn faster. A demonstration data guided mechanism was proposed, which makes use of demonstration data to guide agents' actions in the training phase. But after running the experiment on Ant-v2 environment for random seed 3 and 4, the algorithm had experienced the unstable training problem at the beginning of training.

Also Colas et al. [10] tried to solve exploration – exploitation trade off in continuous environment especially in Continuous Mountain Car. By using "Goal Exploration Process - Policy Gradient" GEP-PG which contained of two stages: the first one was "Goal Exploration Processes" GEP which used a directed exploration of the continuous state action space for a specific environment. Then stored the results in the replay buffer of a deep RL algorithm, which processed them to perform sample efficient policy improvement.

Nair et al. [11] aimed to solve the exploration problem via imitation of a human expert. That combined demonstration-based imitation learning and reinforcement learning to solve exploration problems in robotic tasks. Also learn a policy from demonstrations and rewards, using demonstrations to make the RL problem easier. The main limitation of this work was small efficiency when solving tougher tasks.

III. BACKGROUND

A. Smart Start

Smart Start was developed for sparse or misleading rewards, in which the agent receives the rewards after achieving the goal which make the learning process harder [12]. Smart Start uses the previous information to find the region is expected to give the best information of the agent to solve the challenge and reach the goal. On the other hand, in normal learning the agent spends most of the time just re-exploring the states that have already visited.

Fig. 1 displays a normal RL contrasted with RL with Smart Start [13]. In normal RL, the agent starts in the first state s0 and continues its policy utilizing some exploration strategy, named πexplore, till the end of episode in a final state sterm. For Smart Start, in contrast, the agent firstly finds the Smart Start state Sss then obtains a policy πss leading the agent to Sss by utilizing past experiences. The policy πss is implemented till the agent is nearby Sss and consequently the agent implements the learned policy πexplore until finishing the episode in a terminal state sterm.

Here the functionality for discrete systems is considered in the environment. So it becomes essential to utilize the epsilon greedy or Upper Confidence Bound (UCB1) algorithm [14] for getting the Smart Start state and dynamic programming for guiding the agent to the Smart Start state.

1) Choosing smart start state: In choosing S_{ss}, select a reachable state. The state is called reachable when it has been visited no less than one time by the agent. Thus, the agent cannot determine the Smart Start state at the beginning of the learning process because it requires collecting more information about the surrounding environment at the beginning. The agent saves the visited states in a replay buffer (D) [15] which is used in many algorithms in Deep Reinforcement Learning, such as Deep Q-Network (DQN) [16], Deep Deterministic Policy Gradients (DDPG) and Hindsight Experience Replay (HER). The replay buffer has a specific capacity and uses many strategies for sampling the transition. Such as uniform sampling where each transition is sampled with equal probability or prioritized sampling where each transition is sampled with a high Temporal Difference error (TD) [17].

When selecting S_{ss}, the agent is searching for the optimal state in buffer *D* to begin exploring from. As a result, a state with a lower visitation density has a higher probability to be nearby unvisited states as a result has a high probability of leading to new information. In this project an easy approximation has been used and only taken into consideration the visitation density of discrete states. That can be verified simply by the visitation counts *C(s)* to every visited state. The following equation shows how to choose the smart start state [7]:

$$S_{ss} = \text{argmax}[\text{maxQ}(s,a) + c_{ss}\sqrt{\frac{\log \sum_{s \in S} C(s)}{C(s)}}] \forall s \in D \quad (1)$$

Where Q (s, a) is the action value function and D is the size of buffer. A constant css > 0 for varying the amount of exploitation and exploration and has to be selected suitably. Sometimes the agent only needs exploration especially at the start of the learning process to learn too much about the surrounding environment. A large value for css will produce more exploration. The css value may be reduced in the learning process to change from pure exploration to a suitable balance between exploration and exploitation.

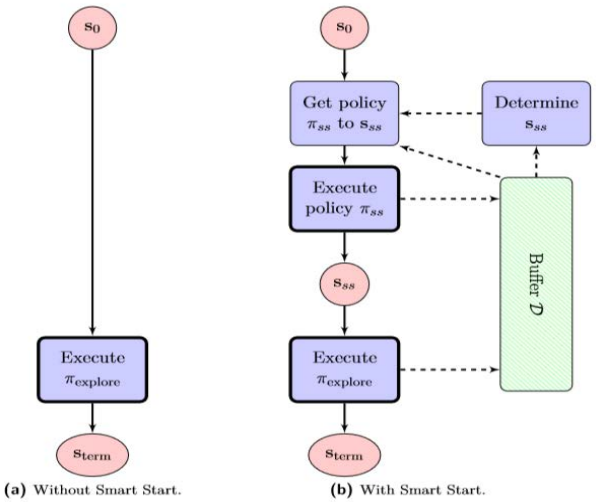


Fig. 1. Comparison between RL Episodes with and without Smart Start.

To explain the previous equation in details, suppose that there is a state with 10 visitation density then the value of $\sqrt{\frac{\log 10}{10}}$ will be 0.316. On the other hand, if a second state was with 20 visitation density then the value of the same part will be 0.255. Because of that the agent chooses the maximum value since it indicates to a less visitation density.

2) Leading the agent to smart start state: Leading the agent from the first state to S_{ss} is considered the second section of the Smart Start technique. As mentioned previously the Smart Start state has been visited before, consequently the trajectory to S_{ss} is known. The easiest method to reach the S_{ss} might be through replaying the trajectory. However this has clear consequences especially in stochastic environments because it has randomness behavior associated with it contrary to deterministic system [18]. Additionally, there is other issue which is the trajectory length as the trajectory might be totally a random path. After numerous iterations the trajectory to S_{ss} can involve numerous series random trajectories, leading to extremely long, complicated and time-consuming path to the Smart Start state. This is not a good option since it lessens the efficiency and the accumulative rewards.

The trajectory optimization method [19] can direct the agent to the Smart Start state. But it is needed to take into consideration the environment in the trajectory optimization with a view to prevent the agent going to regions with great penalties. This work will not consider the environment characteristics when leading the agent to the smart start state, because the main aim is finding a policy which causes to the shortest, most efficient and rapid track to S_{ss}.

This article focuses on discrete environments where a model based approach may be simply applied and almost has an optimum performance. The trajectory optimization may be done by dynamic programming which depends on Value Iteration to give the shortest and most reliable track to S_{ss}.

The agent keep counts the visitation states and learns a transition model as the Model Based Reinforcement Learning method where the transition model and reward function can be simply build. The agent is tracking the visitation counts of the total number of times an action *a* has been used in state *s* which symbolized through *C* (*s*, *a*) in equation 1, also the times number using action *a* in state *s* resulting a traversing to state *s0* which symbolized through *C* (*s*, *a*, *s0*). A rewards' sum for the reward function for each state-action pair is saved. Now the transition model and reward function can be built. Then transitions' reward is given to the S_{ss} in order to use this transitions, the whole transitions have a probability larger than zero for traversing to the Smart Start state receive a reward. That is because the agent is aiming to get into the area of the Smart Start but not exactly in the S_{ss} itself. An ideal policy to the S_{ss} may be obtained utilizing Value Iteration.

B. HER

Imagine that you want to cook any kind of food, and the first trial was bad because you did a mistake then the next time you will avoid that mistake to get a better result. This can simply explain the main idea behind HER which is letting the agent learn from all episodes even the episode was not successful for reaching the main goal *g*. Assuming that at the early episodes the agent cannot reach the final goal *g*, it is supposed that the state where the agent arrives is the virtual reward and the agent can get some reward instead of zero. Which can be seen in Fig. 2.

Suppose that there is an agent with a task *g*, every transition leads to the goal will give a reward 0 or 1. This is called a sparse reward environment where the agent gets the reward when reaching the goal only so that does not help in improving the actions next time. Because of that learning from the sparse reward is so difficult, but HER is solving this problem [20]. HER is taking into account the goal beside the state in the value function. So the transition will be saved in the replay buffer in the following format:

$(S_t \parallel g, a_t, r_t, S_{t+1} \parallel g)$

As mentioned above the agent must split the tasks so instead of storing the transitions regarding the goal *g* every time, the agent also stores the transitions regarding the new selected sub goals *g'* in order to decrease the sparsity. Because of that the agent can get more rewards which help in improving the next new actions which leading to lessen the time that the agent needs to reach the goal *g*. So the new transition will look like the next format:

$(S_t \parallel g', a_t, r', S_{t+1} \parallel g')$

It is not a good idea to tell the agent which sub goals to choose since it will cause a domain specific knowledge. There are four strategies to choose the sub goal. The first strategy uses final state in the episode. The second strategy uses random states that come from the same episode as the transition being replayed. The third strategy depends on choosing random states that come from the same episode. The last strategy chooses the sub goals in a random way. In this experiment, the second strategy is used mainly because it is considered the best one and has the maximum success rate [20].

Fig. 3 displays a diagram for normal reinforcement learning with (HER). In normal reinforcement learning, at every time step the agent receives a representative from the environment state $S_t \in S$ and a reward $R_t \in R$, on that basis selects an action a_t. After executing the action a_t, the agent receives again the modified s_{t+1} and r_{t+1}. The loop continues going on until the environment sends a terminal state, which finishes the episode. When HER is implemented, the same thing will happen except the experiences are stored in the replay buffer that is related to the goal. As a result, through each step the agent will get a batch of experiences, so the agent will store them in the replay buffer regarding the goal then it will choose a sub goal in order to use it.

The HER process can be described in steps as below:

1) Store tuple from the episodes using the goal *g*.
2) Select sub goals, *g'*, using one of the mentioned above strategy.
3) Store new tuples by replacing g to *g'*.

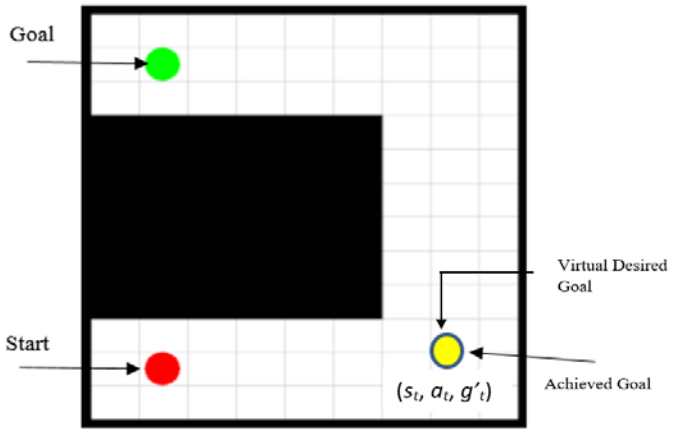


Fig. 2. Grid World Environment showing Virtual Goal in HER Technique.

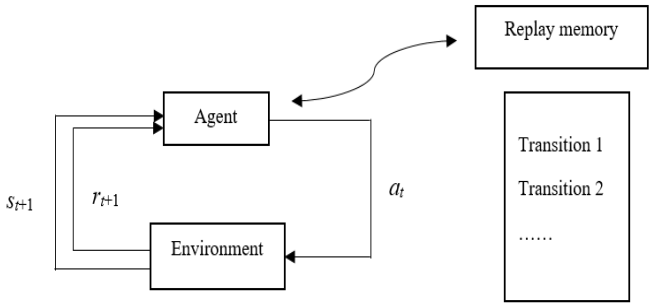


Fig. 3. The effect of Adding Hindsight Experience Replay.

IV. Smart Start and HER Algorithm

In this work both techniques are combined together in order to reach the more efficient exploration. Smart Start solves the challenges in a more efficient way. However, it is not a complete and independent exploration strategy. For this reason, Smart Start should be incorporated with other exploration strategies for increasing the accumulative rewards and decreasing the number of episodes. So a conceptually simple framework (HER) will be added which split the goal to improve exploration by saving the previous experiences.

Both Smart Start and HER was developed for sparse or misleading rewards. In Smart Start the agent receives the rewards after achieving the final goal which makes the learning process harder. So by adding HER the agent can receive some rewards before reaching the final goal that helping in decreasing the learning time.

The algorithm of how HER is implemented with Smart Start technique is provided in Algorithm 1. This algorithm can be utilized as a template for implementing HER with Smart Start framework.

Algorithm 1 Smart Start Framework with HER

1 : Initialize buffer D, and let AGENT starts from initial state $s0$
2 : **For** each episode **do**
3 : Sample a goal g and an initial state $s0$
4 : **For** t=0, T-1 **do**
5 : Sample an action a_t using the behavioral policy from A: $a_t \leftarrow \pi_b(s_t||g)$
6 : Execute the action a_t and observe a new state s_{t+1}
7 : **End for**
8 : **For** t=0, T-1 **do**
9 : $r_t := r(s_t, a_t, g)$
10 : Store the transition $(s_t||g, a_t, r_t, s_{t+1}||g)$ in D
11 : Sample a set of additional goals for replay G:=S (current episode)
12 : **For** $g' \in G$ **do**
13 : $r' := r(s_t, a_t, g')$
14 : Store the transition $(s_t||g, a_t, r_t, s_{t+1}||g)$ in D
15 : **End for**
16 : **End for**
// ---- Smart Start algorithm begins ---- //
17 : **if** $u \leq \eta$ and $|D| > 0$ **then** // Smart Start Algorithm
18 : // select S_{ss} utilizing upper confidence bound
19 : S_{ss} =argmaxa [max $Q(s,a) + c_{ss}\sqrt{\frac{\log|D|}{C(s)}}] \forall s \in D$
20 : // obtain policy using trajectory optimization
21 : π_{ss} = TRAJOPT (D, $s0$, S_{ss})
22 : // execute smart start policy
23 : **Repeat**
24 : Choose $a_t = \pi_{ss}(s_t)$
25 : Take action a_t and observe s_{t+1} and r_{t+1}
26 : Add (s_t , a_t , s_{t+1} , r_{t+1}) to D
27 : UPDATEAGENT (D)
28 : $t \leftarrow t+1$
29 : **Until** $d(s_t, S_{ss}) < \vartheta$, s_t is terminal or $t = T_{episode}$
// ---- Smart Start algorithm ends ---- //
30 : **End for**

From the algorithm, the buffer size D should be more than zero, to ensure there are states inside. Line 19 shows that the agent uses the buffer size instead of visitation count in equation 1 mainly because both of them are equal.

Normally, Smart Start is not used in each episode, it is utilized only in a specific ratio of the episodes. Now after implementing HER that leading to boost the accumulative rewards and enhance the performance of the agent in reaching the goal optimal solution as shown in Fig. 4.

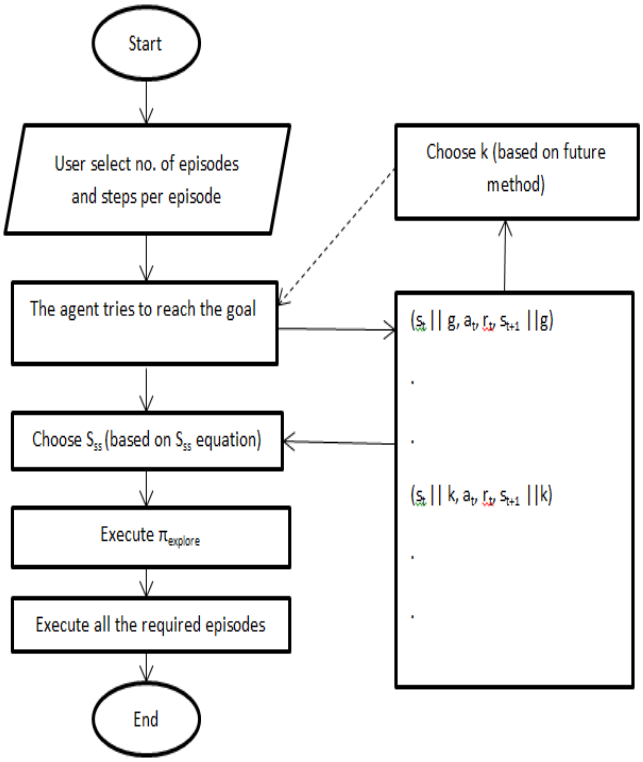


Fig. 4. Flowchart of Smart Start with HER.

V. Experiment

The experiment is done using python 19.2.3 on a grid world environment which is shown in Fig. 2. The initial state is represented by a red circle, and the goal state is represented by a green circle. The agent cannot pass through walls. Once the agent tries to pass a wall the state does not change. The episode is finished once the agent reaches either the state of goal or the limit of steps per episode which is 1000. The agent takes a reward only when the agent reaches the goal state as shown below:

R_{t+1} = {1 if $s_{t+1} = s_{goal}$

{0 if otherwise

The number of steps that the agent takes to reach the goal in this experiment is computed also the average reward per episode. The experiment is carried out on the Easy grid world environments [13]. Firstly, the experiment is carried out with Smart Start framework only then HER is added to the framework to compare between them.

The agent should utilize Smart Start each episode. So, in this experiment $\eta = 1$ will be used. Smart Start does not used in the first episode since there is no information has been stored yet. Through the experiment the value function will be zero. This makes the Smart Start parameter c_{ss} irrelevant in this experiment and can be set to an arbitrary positive value, a value of $c_{ss} = 0.1$ is utilized in this experiment.

VI. Results

Reinforcement learning agents learn from the reward which is given to the agent by the environment. There are certain suggestions in sparse rewards environment like ours. For an algorithm like Q-Learning this means the value function is zero until the goal has been reached for the first time. Also the number of steps that the agent takes to reach the goal for the first time is consequently a vital characteristic of the exploration strategy. But this problem was solved by adding HER which reduce the sparsity in the environment by choosing sub goals and give the agent rewards when reaching there.

To study the result of adding HER on the maze grid world environments, Fig. 5 and Fig. 6 display the average reward per episode and the number of steps required to achieve the goal using Smart Start alone and Smart Start with HER, respectively.

Fig. 5 shows the reward is doubled in SS+HER comparing to SS alone in the episodes which are less than 100. After that both SS alone and (SS+HER) give the same performance. So HER helps SS to be more directed and persistent in the beginning which is leading to more rewards and achieving the goal with a less time.

Fig. 6 shows the number of steps per episode for (SS+HER) is less than number of steps per episode for SS alone, the difference between them was 30% in the beginning then is reduced to reach 5% after the episode 100. When the number of steps is less in all the episodes as a result the time is needed in (SS+HER) is less than the time for SS alone. So (SS+HER) is faster than SS alone, as a result (SS+HER) reduces the learning time.

In both Fig. 5 and Fig. 6, the graph is becoming flat after the learning time. As in the early episodes, the agent is just collecting information. After that, the agent uses this information to reach the goal with the maximum reward and minimum number of steps.

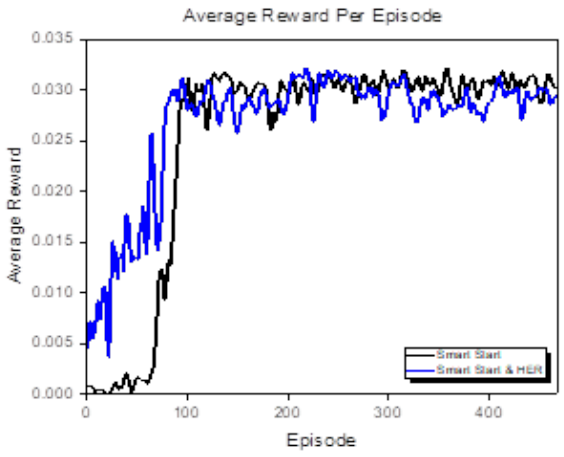


Fig. 5. The Average Reward of Smart Start alone and Smart Start with HER.

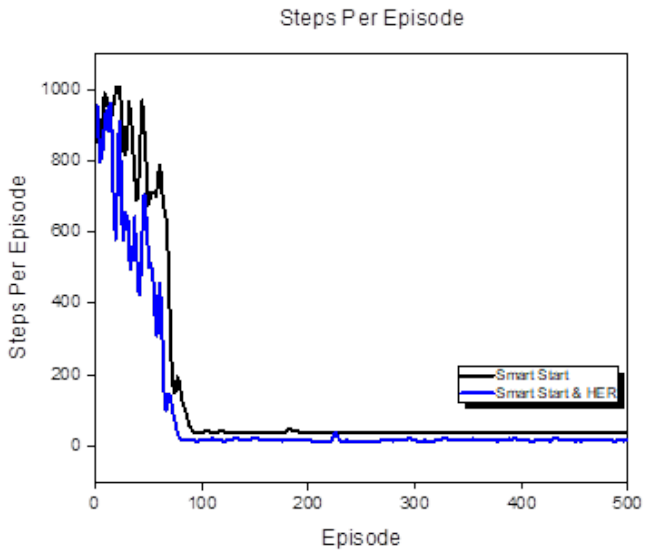


Fig. 6. Steps Per Episode of both Smart Start alone and Smart Start with HER.

VII. Conclusion

The Smart Start framework with HER was assessed by using the easy grid world environment in this experiment. The experiment considered the exploration performance of Smart Start in combination with a limit search space technique which is HER. The performance of exploration was determined as the average number of steps it took the agent to attain the goal state.

It has shown that Smart Start and HER together can enhance the exploration on discrete grid world environments. This clearly leads to efficient performance of the whole learning. The Smart Start and HER can simply be combined with several exploration strategies and reinforcement learning algorithms. That making Smart Start and HER a promising and attracting exploration basis of reinforcement learning challenges. The Smart Start technique was developed for environments with misleading or sparse rewards. This article assessed the performance of Smart Start and HER in discrete environments. For the future works, it is still an open area for other environments not only restricted to discrete environments with misleading or sparse rewards. This directly provides a rise to remarkable guidelines for future work in the same field in other environments such as continuous.

Acknowledgment

This research was supported by Ministry of Higher Education Malaysia Grant under project FRGS/2018/FTKEE-CERIA/F00384 and Center for Robotics and Industrial Automation (CeRIA), Faculty of Electrical Engineering (FKE), Universiti Teknikal Malaysia Melaka (UTeM).

References

[1] R. S. Sutton and A. G. Barto, Reinforcement Learning : An Introduction. London: MIT press, 2015.

[2] K. Shao, Z. Tang, Y. Zhu, N. Li, and D. Zhao, “A Survey of Deep Reinforcement Learning in Video Games,” no. 61573353, pp. 1–13, 2019, [Online]. Available: http://arxiv.org/abs/1912.10944.

[3] L. Shani, Y. Efroni, and S. Mannor, “Exploration conscious reinforcement learning revisited,” 36th Int. Conf. Mach. Learn. ICML 2019, vol. 2019-June, pp. 9986–10012, 2019.

[4] A. D. Tijsma, M. M. Drugan, and M. A. Wiering, “Comparing exploration strategies for Q-learning in random stochastic mazes,” 2016 IEEE Symp. Ser. Comput. Intell. SSCI 2016, 2017, doi: 10.1109/SSCI.2016.7849366.

[5] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal Policy Optimization Algorithms,” pp. 1–12, 2017, [Online]. Available: http://arxiv.org/abs/1707.06347.

[6] J. Schulman, S. Levine, P. Moritz, M. Jordan, and P. Abbeel, “Trust region policy optimization,” 32nd Int. Conf. Mach. Learn. ICML 2015, vol. 3, pp. 1889–1897, 2015.

[7] I. J. Sledge and J. C. Principe, “Balancing exploration and exploitation in reinforcement learning using a value of information criterion,” ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., pp. 2816–2820, 2017, doi: 10.1109/ICASSP.2017.7952670.

[8] K. Azizzadenesheli and A. Anandkumar, “Efficient exploration through Bayesian deep Q-networks,” 2018 Inf. Theory Appl. Work. ITA 2018, 2018, doi: 10.1109/ITA.2018.8503252.

[9] K. Lin et al., “Exploration-efficient Deep Reinforcement Learning with Demonstration Guidance for Robot Control,” 2020, [Online]. Available: http://arxiv.org/abs/2002.12089.

[10] C. Colas, O. Sigau, and P. Y. Oudeyer, “GEP-PG: Decoupling exploration and exploitation in deep reinforcement learning algorithms,” 35th Int. Conf. Mach. Learn. ICML 2018, vol. 3, pp. 1682–1691, 2018.

[11] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel, “Overcoming Exploration in Reinforcement Learning with Demonstrations,” Proc. - IEEE Int. Conf. Robot. Autom., pp. 6292–6299, 2018, doi: 10.1109/ICRA.2018.8463162.

[12] J. A. Arjona-Medina, M. Gillhofer, M. Widrich, T. Unterthiner, J. Brandstetter, and S. Hochreiter, “RUDDER: Return Decomposition for Delayed Rewards,” in 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), 2019, no. NeurIPS, pp. 1–12.

[13] Bart Keulen, “Smart Start A Directed and Persistent Exploration Framework for Reinforcement Learning,” 2018.

[14] E. Hartog and H. Moreines, “New techniques in automatic flight control system design,” SAE Tech. Pap., vol. 3, pp. 397–422, 1961, doi: 10.4271/610369.

[15] S. Zhang and R. S. Sutton, “A Deeper Look at Experience Replay,” 2017, [Online]. Available: http://arxiv.org/abs/1712.01275.

[16] J. Fan, Z. Wang, Y. Xie, and Z. Yang, “A Theoretical Analysis of Deep Q-Learning,” vol. 120, no. 1995, pp. 1–4, 2019, [Online]. Available: http://arxiv.org/abs/1901.00137.

[17] M. Tokic, P. Ertle, G. Palm, D. Söffker, and H. Voos, “Robust exploration/exploitation trade-offs in safety-critical applications,” IFAC Proc. Vol., vol. 8, no. PART 1, pp. 660–665, 2012, doi: 10.3182/20120829-3-MX-2028.00160.

[18] V. Anagnostopoulou, “Stochastic and deterministic absorption in neutron-interference experiments,” vol. 36, no. 9, pp. 1–17, 1987.

[19] J. T. Betts, “Survey of numerical methods for trajectory optimization,” J. Guid. Control. Dyn., vol. 21, no. 2, pp. 193–207, 1998, doi: 10.2514/2.4231.

[20] M. Andrychowicz et al., “Hindsight experience replay,” Adv. Neural Inf. Process. Syst., vol. 2017-Decem, no. Nips, pp. 5049–5059, 2017.

Implementation of Low Cost Remote Primary Healthcare Services through Telemedicine: Bangladesh Perspectives

Uzzal Kumar Prodhan [1], Tushar Kanti Saha [2]
Rubya Shaharin[3]
Department of Computer Science and Engineering
Jatiya Kabi Kazi Nazrul Islam University
Trishal, Mymensingh, Bangladesh

Toufik Ahmed Emon [4], Mohammad Zahidur Rahman [5]
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka 1212

***Abstract*—In this paper, we have implemented a low cost primary healthcare service for the remote rural people of Bangladesh. These services were delivered through our developed advanced telemedicine model. The main aim of this paper is to provide basic healthcare service through the developed low cost hardware. We have developed Arduino based low cost hardware's to be used for this telemedicine services. Remote patients of Bangladesh can get the expert doctors opinion without going to the urban areas. We have collected nine vital signs such as electrocardiogram (ECG), oxygen saturation (SPO2), blood pressure, temperature, body position, glucose level, airflow, height, and weight of patients to be used in our model. We have removed unwanted signals from the collected vital signs through several filtering algorithms. Our system was successfully tested with the patients of Marie Stopes Bangladesh Hospital. From our developed model, rural patients can get primary healthcare services from the pharmacy of any remote village of Bangladesh with the assistance of local doctor by using Raspberry PI. Finally, we can say that the deployment of the developed healthcare service will reduce the cost of the telemedicine services and advances the healthcare facilities for the remote people of Bangladesh.**

Keywords*—Raspberry PI; DGHS; Arduino; Portable; ECG; SPO2

I. Introduction

Healthcare services of Bangladesh are mainly divided into two areas: urban and rural. Most of the expert doctors are urban centric where healthcare services are available. But, most of the people of Bangladesh live in rural areas. There are limited healthcare facilities in rural areas. There is a huge shortage of expert doctors in rural areas. As a result, remote rural people will have come to the urban areas for better healthcare services. Patients will have to travel to a long distance. They have to reside in city for some days to get treatment. The fees of expert doctor and different service costs will have to bear of patients. As a result, total costs are very high for the patients. Sometimes, it is also very difficult for the remote patients to go to the urban areas for the treatment due to the physical conditions of the patients. Fig. 1 shows the real scenario of health workforce-population situation of Bangladesh. From Fig. 1, we see that this ratio is very low for Bangladesh. This ratio is also very low in the rural areas in comparison with the urban areas of Bangladesh.

Fig. 2 shows the comparative analysis of physician density among the western pacific regions. From Fig. 2, we can see that this ratio is very low for Bangladesh. Considering these scenario, telemedicine can be an alternate solution for the delivery of effective medical services to the large rural communities of Bangladesh.

Healthcare facilities are distributed in Bangladesh according to the Fig. 3. From Fig. 3, we can see that health services are distributed into five layers in Bangladesh. At present, telemedicine services are available up-to the selected hospitals in Upazilla level. So, there are huge scopes to introduce telemedicine services in Union and Ward level of Bangladesh.

S.N.	Description of the facilities	Ratio
1	No. of doctors per 10,000 population	1.43
2	No. of registered nurses per 10,000 population	2.90
3	No. of registered physicians per 10,000 population	4.90
4	No. of nurses per 10,000 population	1.05
5	No. of medical technologists per 10,000 population	0.37
6	No. of community and domiciliary health workers per 10,000 people	4.04
7	Hospital bed for 1528 population	1

Fig. 1. Health Workforce-Population Status of Bangladesh.

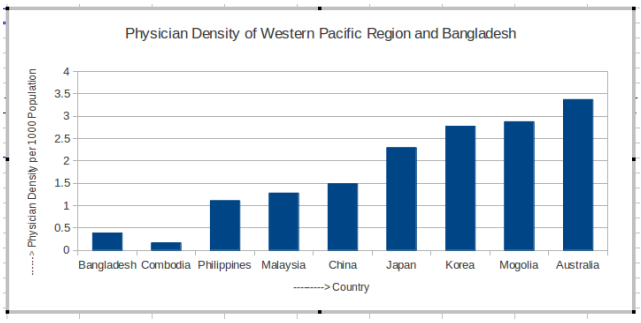


Fig. 2. Physician Density of the Western Pacific Countries.

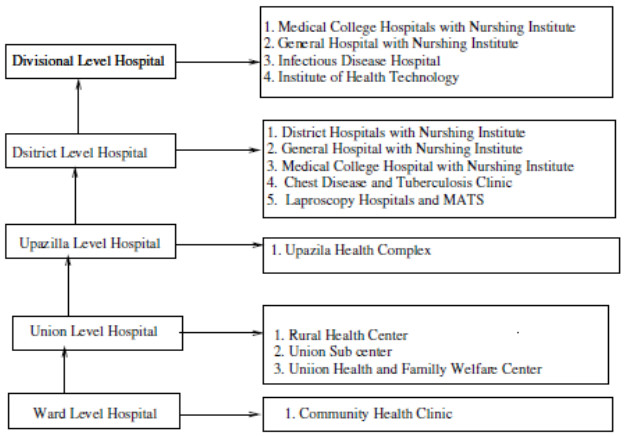


Fig. 3. Current Health Structure of Bangladesh.

Telemedicine can be an effective solution for addressing the current healthcare problems in Bangladesh. The scarcity of expert doctor in remote areas can be removed through telemedicine. Currently noninvasive technologies [1, 2-3] are combined with disease management process to provide more patient information, with a view to improve healthcare decision-making.

In this research, we have developed low cost remote primary healthcare services through telemedicine for addressing the financial problems of poor patients of Bangladesh. Remote people will go to the pharmacy of a remote village for their primary healthcare services. The pharmacy will be equipped with the developed necessary hardware's for telemedicine services. Local doctors will use our developed low cost portable telemedicine tool kit for the services. Patient's registration and telemedicine services will be delivered through the local pharmacy.

The rest of the paper is organized as follows: Section 2 is the review of literature section which presents the ongoing research activities done on the telemedicine and summarizes the scopes of further research, Section 3 is the materials and methods section which depicts the working methods of the proposed research and the corresponding materials, Section 4 is the results section which presents the outputs of this results, Section 5 is the discussion section which summarizes the research and Section 6 is the conclusion section which concludes the research.

II. Review of Literature

This section presents some of the recently published papers with a view to find the current status of the research. Literature review helps us to find out the current demand of proposed research and presents the requirement analysis of the research. Now, we are going to discuss some of the recently published research in the section below.

A health system was proposed and developed by Ashir Ahmed and other researchers in 2013 through Portable Health Clinic and GramHealth. They have tested their system in three villages of Bangladesh and found satisfactory result for pilot project. In that paper they have tried to managed the huge amount of patient data to be used for fruitful health solution in a cost effective way [4]. Jecinta Kamau et al. developed a model "Social Services" on Wheel with a view to enhance the services in 2014 for the unreached people. This model [5] was tested in rural community of Bangladesh. The model was operated by subsidized but their proposed policies could make the model cost effective.

Scalable and Internet-based architecture for ECG in telemedicine application is developed by Umme Sayma Busra and others in 2014 to include hospitals and medical specialists to focus the rural health centers. In this research, the authors [6] used an ECG kit to measure the signal of a patient and send data to a server. The data is then forwarded to the nearest health professionals for measurement and provide consultation to the patients.

Aditi Kaushik and Pooja Sabherwal developed an ECG monitoring system by using android phone in 2015. They have used Pam Tompkins algorithmic approach. In this research work, feature extraction was done from the ECG signal. After feature extraction, its implementation was done in an android environment. Effective results are derived from this proposed research [7]. The research conducted by Salman Ahmed and others in 2015 results the construction of a device that can be used for transferring the data of a patient's vital signs to a remote device wirelessly. This research used blue-tooth technology for the transmission of data to expert doctors for evaluation. Temperature and ECG signal are taken from the patients to be analyzed and processing for transmitting. In future, authors [8] want to use GSM technology for data transmission to a wide range.

The authors presented a wearable sensor based elderly home care system in a smart environment for the elderly people in 2015. They have used PC as monitoring system, sensors for sensing data, web-service to communicate between monitoring systems. They have used two modules wearable user and main module for this research. They got satisfactory results for this model [9]. According to Amanda Mohon Ghosh and others remote patient monitoring and guidance awareness in a authenticated manner are the vital in telemedicine application. The whole work conducted by the researchers was carried out by three modules called sensing, main module and interaction module. This research [10] was evaluated in the hospital with the real patients for its acceptance.

ECG is one of the most important vital signs of non-invasive technique in diagnostics of heart oriented diseases [11]. Noise of ECG signal can be removed through Least Mean Square algorithm and hardware based filtering techniques [12]. Blood pressure (BP) is identified as a vital sign for controlling hypertension and cardiovascular disease [13]. Widely used blood pressure measuring technique is the oscillometric method through automated non-invasive BP measurement devices [14-15]. A study was conducted by Nilanjan Dey et al. [16] on residential wireless sensor networks for ECG healthcare monitoring in 2017.

Real time activity recognition framework for health monitoring support in home environments is developed by Shaikh Farhad Hossain in 2017. This system collects and analyzes sensory data in real-time. It gives variety feedback to the users. In addition, it can generate alerts based on the

detected events and finally store the data to a server for further use [17]. Remote health monitoring system in a rural population: challenges and opportunities were conducted by Jacey-Lynn Minoi and Alvin W Yeo in 2014. In this study, blood pressure sensor is used for the pilot testing. Authors found that remote blood pressure monitoring is suitable for the management of hypertension in tele-homecare [18]. In 2016, Rafhanah Shazwani Binti Rosli et al. developed mobile heart rate detection system. This study develops a program for Arduino micro-controller that enables a pulse sensor module to detect alarming changes of a user's heart rate. They used a GM Shield to act as an SMS interface between the device and receivers. The developed is small sized and low cost which makes this device to be used more widely [19]. W. Yan [20] found that technology assists in meeting primary healthcare goals to understanding public healthcare difficulties, or to power individuals to engage in their own health. Authors [21] found the healthcare data breaches in this paper and they also found their primary causes with the role of PHI and HIPAA Act for dissuading data breaches. Authors [22] designed a healthcare system considering privacy and security through Blockchain technology for serving with deliberate medical care and benefits to the patients across. Authors [23] worked on the hiring of proper human resources and highly skilled professionals in the supply chain departments to improve the quality of supply chain healthcare services. Authors [24] proposed resource-aware service-oriented service model as a way to design systems that can efficiently provide quality healthcare services in this paper.

From the literature review section, we have found that real time vital information of remote patient is necessary for the expert doctor to deliver medical services. As we have limited doctors in rural areas, we are going to connect with the expert doctor in urban areas with the local doctors in a village with the low cost portable telemedicine tool kit. As a result, poor patient can enjoy the healthcare facilities at low cost from the remote pharmacy.

III. Materials and Methods

In this section, we have developed the primary healthcare model to be deployed in the rural areas of Bangladesh. The model is shown in Fig. 4. The model is expressed through Business Process Modeling Notation (BPMN 2). There are four processes in our developed model for Bangladesh. The processes are local administrator in pharmacy, remote local doctor, expert doctor and health system administrator. The activities of the process are shown in Fig. 4. At first, we have developed the components of telemedicine model at low cost. According to the model, pharmacy is the local hub for the telemedicine services. Patients will come to the pharmacy of a local village for receiving telemedicine services. Administrator in pharmacy is responsible for inputting the patient's record to the health system. Local doctors in village are equipped with our developed portable tool kit. He is responsible for doing the primary health check-up of the patients and interacts with the expert doctor. Expert doctors will log-in to the health system through our developed client module. We have used Raspberry PI 3 Model B to make our remote client log-in module. All the medication plan and services will be given to the patients in a printed form from the pharmacy and the fees for the service is also received by pharmacy administrator.

The arrangements of the components of the developed telemedicine model for Bangladesh are shown in Fig. 5. From Fig. 5, we can see that there are six main components of the model. The components are sensors, developed portable telemedicine tool kit, android application, staging server, HL7 based health system, Raspberry PI based client log-in module. The sensor components are used to collect different vital signs of patients.

Portable telemedicine kit is developed with ardunio uno, e-health sensor shield and Bluetooth. This kit is very small in size, low power consumption, easily portable and has easy interface with the sensors. This tool kit is powered from USB. As a result, we can easily use this kit in rural remote areas. Physical arrangement of the toolkit for Bangladesh is shown in Fig. 6. According to our design, this toolkit will be used by the local doctors for collecting vital signs of patients. In order to complete this task, we have used ardunio uno as the microcontroller.

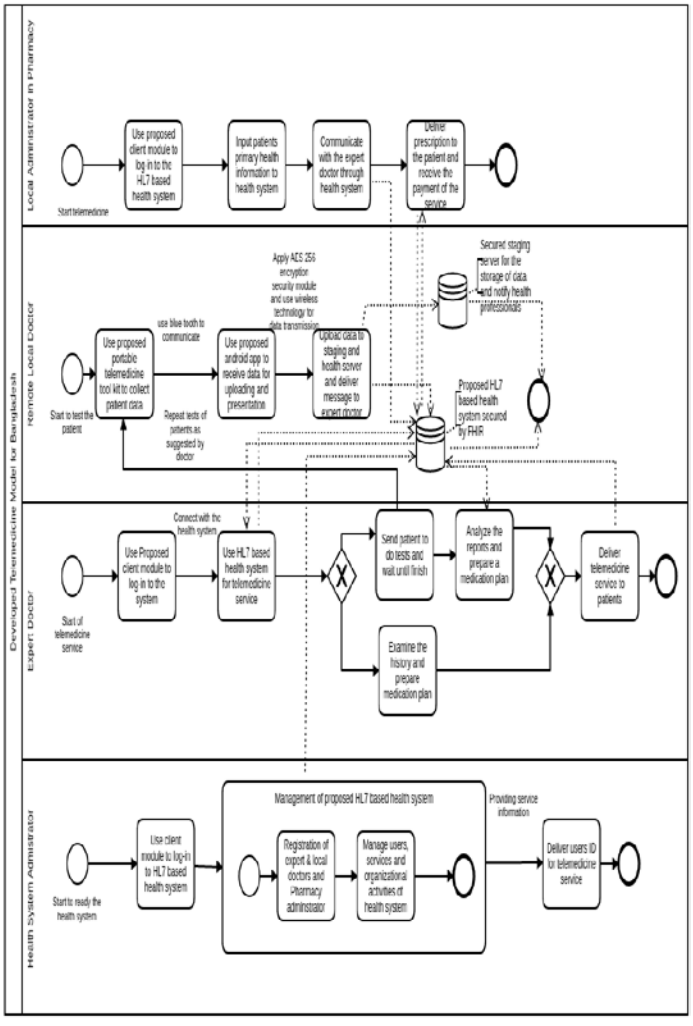


Fig. 4. Developed Primary Healthcare Model for Bangladesh.

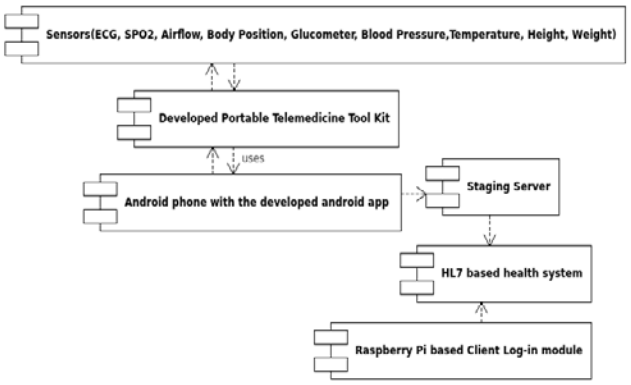


Fig. 5. Arrangements of the Components of Primary Healthcare Model.

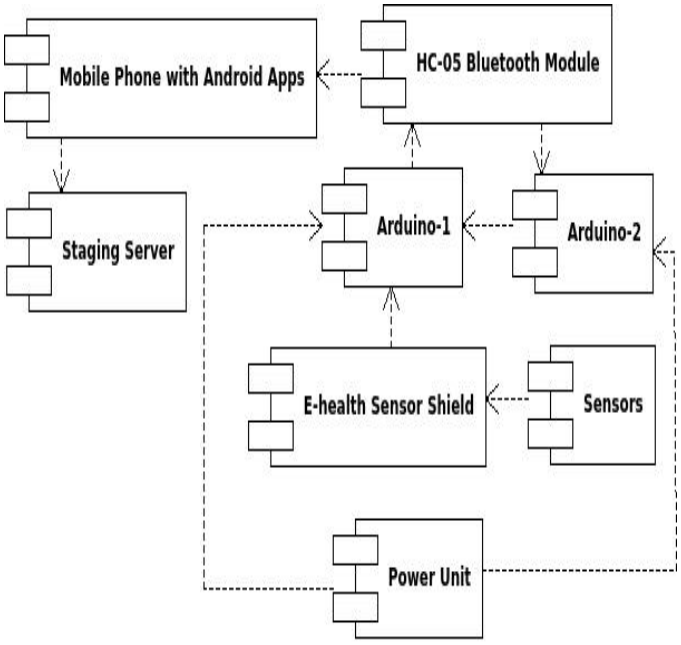


Fig. 6. Physical Arrangements of the Components of the Developed Toolkit.

We have developed a communication module for data processing, visualization and uploading. In order to do this work, we have developed an android application which will interact with our developed portable tool kit through Bluetooth. A secured valid connection is required to transfer patient data for processing and other works. We have done required preprocessing on the collected data to be used for telemedicine services. In this case, we have used AES encryption algorithm for preventing unauthorized access and securely data up-loading to the staging server. Fig. 7 shows the user interface of our developed android application for our system.

In this research, we have used staging server for storing the field test data to be used by our health system, health professionals and other health organizations. Local doctors use portable kit and an android app to upload patient-wise lab test record to the server. In the server, we have used another messaging module for the expert doctor. Automated instant messages are delivered to expert doctors when any patients test data are available in the system. Quick remote healthcare services can be given to the patient by using our services. From the staging server, we have collected remote patient's vital information through python script. We have mapped the patient ID and test ID in both health system and staging server to import health record of patients.

From the present telemedicine models, we have seen that there are no health service related record is available in these models. As a result, the facilities of telemedicine services are not utilized fully. Patient health data are not interoperable among health organizations. This creates a lot of problems for us. In our model, we have used a HL7 based open source health system for telemedicine services. Our used health system is inter-operable and provides maximum benefits for the patients. All health service oriented electronic record is organized according to the definition of standard developing organizations. The modules of our used health system are shown in Fig. 8.

From the Fig. 8, we can see that there are mainly nine modules in our health system. Different health services are delivered remotely under these modules through our model. We have tried to reduce the overall cost of the system through the use of Raspberry PI. It is small in size, portable, easily manageable and a cost effective component for our model. We have made our remote log-in module for our developed system through Raspberry PI. Fig. 9 shows the developed remote client log-in module.

Fig. 7. User Interface of the Developed Android Application.

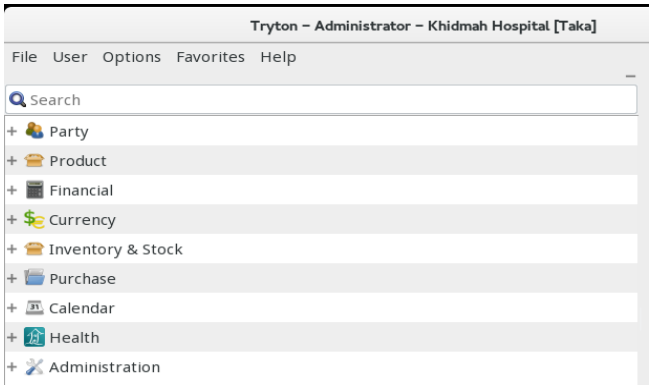


Fig. 8. Modules of Developed Health System.

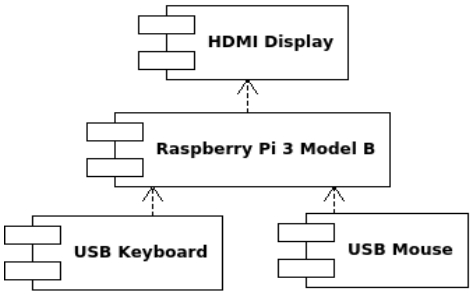


Fig. 9. Developed Raspberry PI based Client Module.

In order to make our model more secured, we have applied the suitable security features of standard and well known security model in our developed model. In our developed model, we have used the functionality of role based security model, multilevel based security model and access control list security model. Fig. 10 shows the security measurement techniques applied to our model. These techniques prevent unauthorized access and provide more privacy of the system.

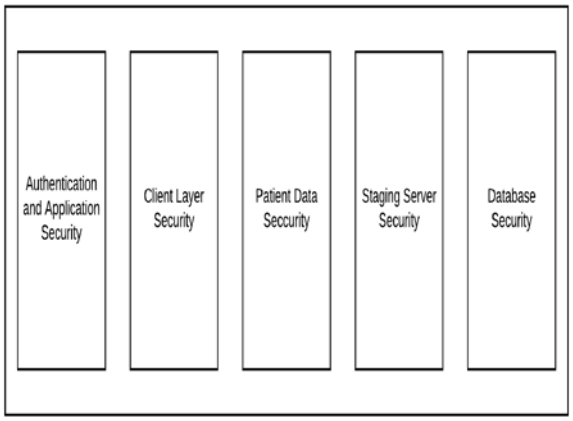


Fig. 10. Security Implementation Stages for the Developed System.

IV. Results

We have developed our components according to the telemedicine model shown in Fig. 4. We have installed the hardware components and successfully up-loaded the code to the microcontroller. We have prepared client module Raspberry Pi to connect with the health system. This log-in module will be used in the pharmacy of a local village. In this research, we have focused the primary healthcare services for the remote patients of Bangladesh. Our developed module works on blood pressure, ECG, height, weight, SPO2, airflow, body position, temperature, and glucometer sensor. Fig. 11 shows the log-in module for doctor to the health system.

Patient registration and history of the patient will be inputted by the administrator of pharmacy. Fig. 12 shows the snapshot from pharmacy module about patient record.

Remote patient will also get the medication plan from the expert doctor. Fig. 13 shows the medication report prepared by expert doctors and delivered from pharmacy module. Doctors can easily get the patients history and previous medication plan any time from the system. This can be a guide for them about the treatment of a patient.

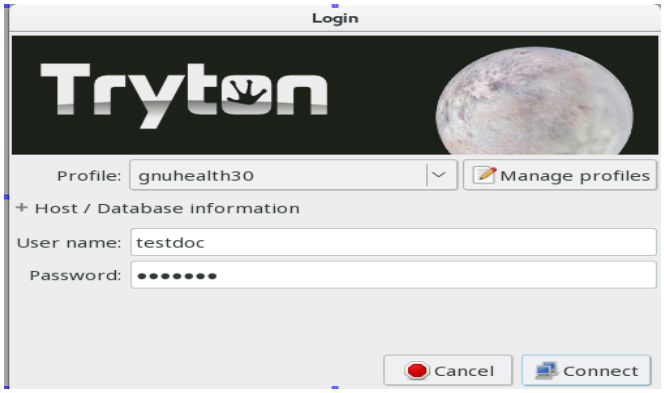


Fig. 11. Doctors Log-in Module of the System.

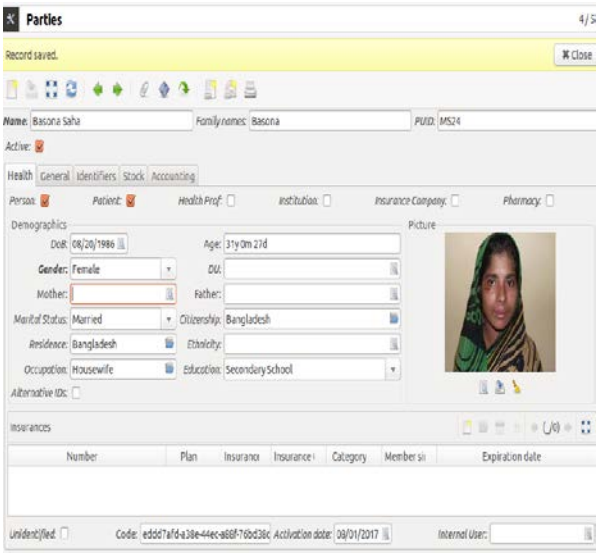


Fig. 12. Outputs from Pharmacy Module about Patient Record.

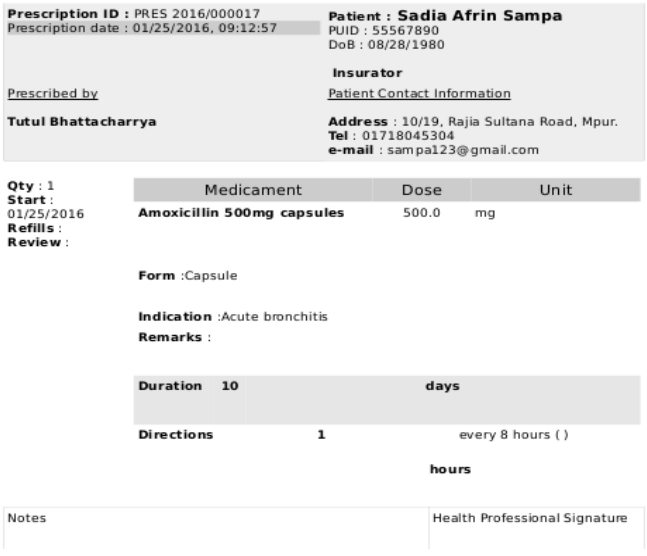


Fig. 13. Medication Report Generated from the Developed System.

Local doctor will use our developed android application to collect vital signs from the tool kit for further analysis. Fig. 14

shows patient-wise data up-loading section to the staging server.

According to the direction of the expert doctor, local doctor will use our developed different sensors for data collection. Fig. 15 shows the output generated from our communication module for body position and glucometer sensor. Expert doctors will communicate with local doctors in case of emergencies for the better treatment of the patients. This communication will be done through our used health system.

Expert doctors are one of the vital actors in our model. When rural patient data are available through our communication module, the doctor is instantly notified through our messaging module. On line and off-line messaging are available in our system. Both e-mail and mobile number is used for the messaging of expert doctors. Doctors can easily log-in to the health system through our developed module. Doctors can instantly check the status of the patients and provide expert opinion to the remote patient. Fig. 16 shows the expert doctor's lab test requests for rural patients.

Warning messages can also be given in the report as well as real time message to expert doctor for the patients to take advance preparations. Fig. 17 shows the section of how the messages can be given to the patients and doctors for lab test results.

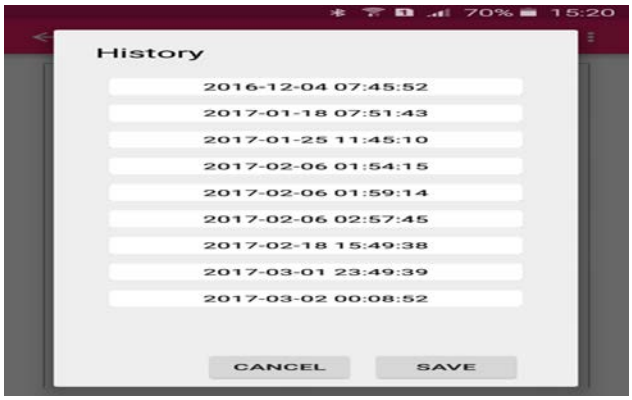
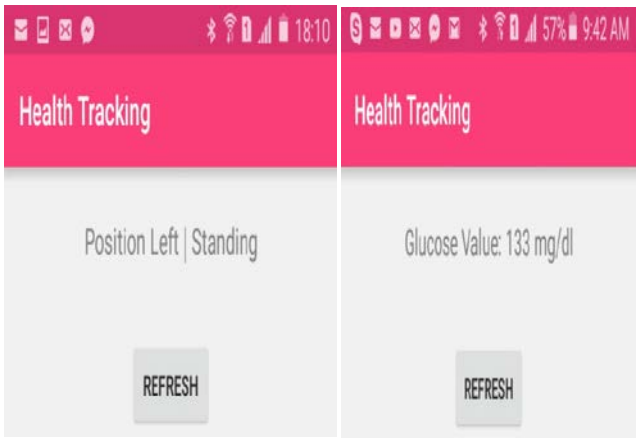


Fig. 14. Patient-wise History from Staging Server.



(a) Body Position Sensor Output. (b) Glucometer Sensor Output.

Fig. 15. Outputs from Body Position and Glucometer Sensor.

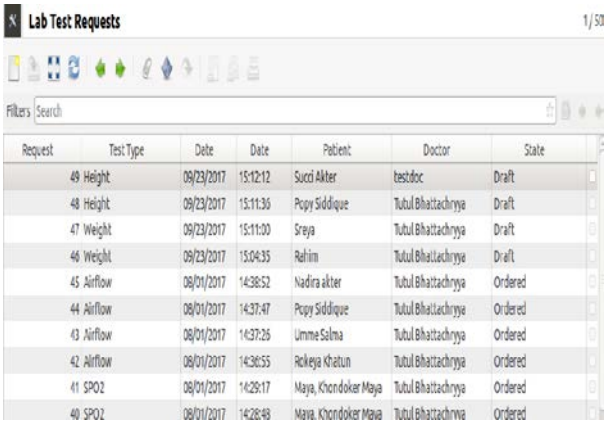


Fig. 16. Lab Test Requests for Patients from Expert Doctor.

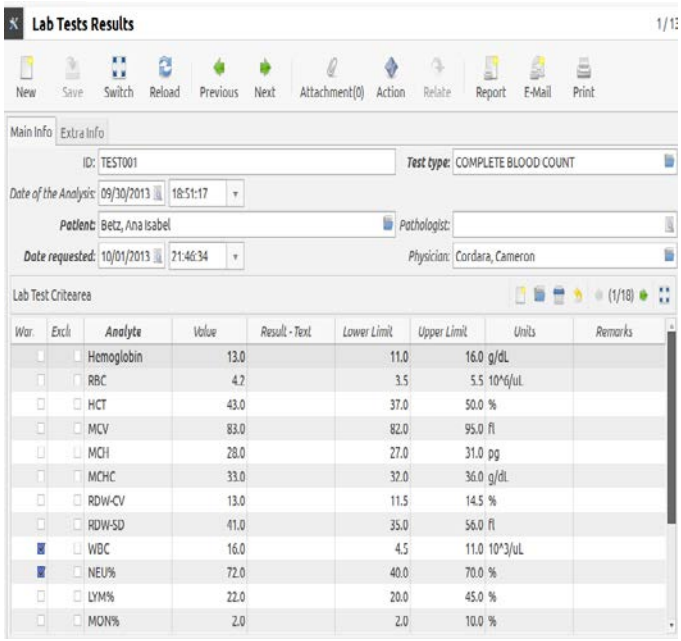


Fig. 17. Warning Message Generated from the System for the Patients.

V. Discussions

In this paper, we have focused on the primary healthcare services for the huge number of remote rural people of Bangladesh. In order to meet their basic healthcare facilities, we have successfully developed the low cost components of the telemedicine service considering the financial status of these people. A recent study was done by Uzzal Kumar Prodhan and others in 2017 on the telemedicine services of Bangladesh. From the survey result, we came to know that the telemedicine operating cost should be in the range of 200-300 taka for the people of Bangladesh. Focusing on these financial proposals, we have developed a cost effective model so that remote people can afford the cost and our telemedicine model will be sustainable in-terms of cost.

Most of the telemedicine systems were in project basis in Bangladesh. Some of the models use real time based facilities such as video conferencing and some of the models use store and forward based techniques. But in our developed advanced model, we have included the facilities of both methods for the

acceptance of our system. In the case of local doctors, they are getting the rural patient information through our communication module in real time. Then local doctors send the vital information to expert doctor through our health system by using store and forward based method. Expert doctors also provide the medication plan through store and forward based. By using this approach, we can deliver fast healthcare services to the remote patients.

From the survey results published in 2016, we came to know that rural patient costs minimum 1500 taka to get one time treatment from the expert doctors from urban areas. Sometimes it is also difficult or impossible for a remote patient to travel to a long distance. In order to assist those patients, we have developed this model which can be easily deployable in the pharmacy of a remote village. Patients can get the expert doctors opinion without travelling to urban areas from pharmacy. This arrangements saves rural patients time, cost and get immediate primary healthcare services at low cost.

Telemedicine service center setup cost is very high to introduce telemedicine services. In Bangladesh, these costs vary from project to project. Largest telemedicine projects run by Director General of Health Services (DGHS), Bangladesh costs about 30 lacks taka to setup a center. All the models were in hospital based. Patients will have come to the urban telemedicine center to get the service. We want to deploy our advanced telemedicine model to the remote areas of Bangladesh. So, we have considered their financial strength. Normally, a pharmacy needs to use a Raspberry Pi based client module and an Internet connection to run the center. This module will be connected to the centralized health system. This module costs around 10000 taka only. Pharmacy owner's has already showed their interest to invest this small amount for the health services of the poor people of Bangladesh.

In this research, local doctors are equipped with our developed tool kit for doing our different primary tests according to the direction of the expert doctor. We have focused on the primary health check-up of rural patients. By using our health kit, we can collect nine vital signs of patients and check the patient's primary vital signs in real time. We have included a HC-05 Bluetooth module to make our kit for sending data for further diagnosis. In this kit, we have used low cost microcontroller named ardunio uno and e-health sensor shield. The cost of the kit is also very low compared to the available existing tool kit used in our country. This tool kit has extensibility capabilities. We can add more sensors with our developed tool kit by updating the microcontroller code to diagnosis more vital parameters of patients.

In order to test our developed tool kit, we have conducted a field test with the patients of Marie Stopes Bangladesh Hospital, Dhaka, Bangladesh. Our sample size was twenty. We have tested our system with the close supervision of the expert doctors in the hospital. Doctors have checked the collected results from our device and verified the results. The results are shown in the result section.

This research worked on nine vital signs of patients only. In future we can easily extend this research work to collect more vital signs of patients. This study conducted only one field test with the Marie Stopes hospital. More field test can be conducted for the validity of the developed system.

Every health organizations in Bangladesh have its own customized health systems. Currently, health systems of Bangladesh do not interact with each other for health services. As a result, patients will have to cost more for the disparity of present health systems. They will have to do their same health check-up repeatedly. There is no central database of their health record. All these lacking make the total health system inefficient. In this study, we have solved these problems by following the recommendations of standard developing organization for health. In order to offer telemedicine services, we have followed Health Level Seven International (HL7) based open source health system. This health system solves data interoperability problem and assists health professionals to make a decision about a patient. This approach helps health organizations to manage their patient's records efficiently.

VI. Conclusions

We have successfully implemented and tested our developed telemedicine model for primary healthcare services of remote patients. In this research, we also developed the required hardware's at low cost. We have used Bluetooth and mobile network for collecting rural patient data to be used in our advanced telemedicine model. As mobile network is available in all the remote villages of Bangladesh, we have used these features in remote client module for telemedicine services. Remote client module can be operated with low energy. As power crises exist in Bangladesh, we can run this module without electricity for a long time. Pharmacy will be the service center for this model where patients will receive their services. Local doctors, patients and expert doctors are the vital components of the advanced telemedicine model. Remote patients can get the improved healthcare services at low cost without travelling to a long distance. Expert opinion can be easily and instantly delivered to the patients in emergency situation. Finally from the point of view of Bangladesh, we can say that this model will be a cost effective and sustainable telemedicine model for the telemedicine services for the unreached people of Bangladesh.

Acknowledgment

Authors would like to thanks to the Information and Communication Technology Division of Ministry of Posts, Telecommunications and Information Technology, ICT Tower, Agargoan, Dhaka, Bangladesh for their fellowship of PhD program. Authors would also like to give especial thanks to the officials of Computer and Electronics labs of Jahangirnagar University, Savar, Dhaka, Bangladesh for their assistance to make this research successful.

References

[1] Caroline Free, Gemma Phillips, Louise Watson, Leandro Galli, Lambert Felix, Phil Edwards, Vikram Patel, and Andy Haines, "The effectiveness of mobile-health technologies to improve health care service delivery processes: A systematic review and meta-analysis," PLOS, Vol. 10, No. 1, 2013, e1001363.

[2] Steinhubl SR, Muse ED, and Topol EJ, "Can mobile health technologies transform health care?," JAMA, Vol. 310, No. 22, 2013, pp. 2395-2396.

[3] Steven R. Steinhubl, Evan D Muse, and Eric J. Topol, “The emerging field of mobile health,” Science Translational Medicine, 7(283):283rv283, 2015.

[4] Ashir Ahmed, Lutfe Kabir, Eiko Kai, and Sozo Inoue, “Gramhealth: A bottom-up approach to provide preventive healthcare services for unreached community,” in Annual International Conference of the IEEE EMBS, No. 35, IEEE, 2013, pp. 1668-1671.

[5] Jecinta Kamau, Andrew Reberio-Hargrave, Hiroaki Satto, Emran Abdullah, Hiroshi Okajima, and Ashir Ahmed, “Social services on wheels: A sustainable model to improve access in unreached communities,” in IST-Africa 2014 Conference, 2014, pp. 1-8.

[6] Umme Sayma Busra and Mohammad Zahidur Rahman, “Mobile phone based telemedicine service for rural Bangladesh:ECG,” in International Conference on Computer and Information Technology, No. 16, IEEE, 2014, pp. 203-208.

[7] Aditi Kaushik and Pooja Sabherwal, “Monitoring electrocardiogram using android based smart phone,” in Annual IEEE India Conference, IEEE, 2015, pp. 1-6.

[8] Salman Ahmed, Sabrin Millatand, Md. Aymanur Rahman, Sayeda Naeyna Alam, and Md. Saniat Rahman Zishan, “Wireless health monitoring system for patients,” in IEEE International WIE Conference on Electrical and Computer Engineering, IEEE, 2015, pp. 164-167.

[9] Md.Nazam Al Hossain, Aprojit Pal, and SK Alamgir Hossain, “A wearable sensor based elderly home care system in a smart environment,” in International Conference on Computer and Information Technology, No. 18, IEEE, 2015.

[10] Ananda Mohon Ghosh, Debashish Halder, and SK Alamgir Hossain, “Remote health monitoring system through IoT,” in International Conference on Informatics, Electronics and Vision, No. 5, IEEE, 2016, pp. 921-926.

[11] V.Vijendra and Meghana Kulkarni, “ECG signal filtering using dwt haar wavelets coefficient techniques,” in International Conference on Emerging Trends in Engineering, Technology and Science, IEEE, 2016, pp 1-6.

[12] Laxmi Shetty, “Electrocardiogram preprocessing using Weiner filter and least mean square algorithm,” International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, Vol. 3, No.1, 2015, pp. 82-85.

[13] Wilmer W Nichols, Charalambos Vlachopoulos and Michael O Rourke, “Mc-Donald Blood Flow in Arteries, Sixth Edition: Theoretical, Experimental and Clinical Principles,” CRC Press, Editions 6, 2011.

[14] M L Antonova, “Noninvasive determination of the arterial elastogram and blood pressure part i: Arterial elastogram and volume pulsations,” Blood Pressure Monitoring, Vol. 18, 2013, pp. 32-40.

[15] José Antonio de la O Serna, Wendy Van Moer, and Kurt Barbé, “Using alternating kalman filtering to analyze oscillometric blood pressure waveforms,” in IEEE Transaction Instrument Measurement, Vol. 62, IEEE, 2013, pp. 2621-2628.

[16] N. Dey, A. S. Ashour, F. Shi, S. J. Fong, and R. S. Sherratt, “Developing Residential Wireless Sensor Networks for ECG Healthcare Monitoring,” in IEEE Transactions on Consumer Electronics, Vol. 64, No. 4, 2017, pp. 442-449.

[17] Shaikh Farhad Hossain, “Real time activity recognition framework for health monitoring support in home environments,” in 3rd International Conference on Electrical Information and Communication Technology (EICT), Khulna, IEEE, 2017, pp. 1-6.

[18] J. L. Minoi and A. W. Yeo, “Remote Health Monitoring System in a Rural Population: Challenges and Opportunities,” in IEEE Conference on Biomedical Engineering and Sciences, 2014, pp. 895-900.

[19] R. S. B. Rosli and R. F. Olanrewaju, “Mobile Heart Rate Detection System for EarlyWarning of Potentially-Fatal Heart Diseases,” in International Conference on Computer and Communication Engineering, IEEE, 2016, pp. 422-427.

[20] W. Yan, "Technologies for Primary Health Care Help Meet Global Goals," in IEEE Pulse, Vol. 10, No. 3, pp. 15-18, May-June 2019, doi: 10.1109/MPULS.2019.2911822.

[21] A. Jayanthilladevi, K. Sangeetha and E. Balamurugan, "Healthcare Biometrics Security and Regulations: Biometrics Data Security and Regulations Governing PHI and HIPAA Act for Patient Privacy," 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2020, pp. 244-247,doi:10.1109/ESCI48226.2020.9167635.

[22] S. Chakraborty, S. Aich and H. Kim, "A Secure Healthcare System Design Framework using Blockchain Technology," 2019 21st International Conference on Advanced Communication Technology (ICACT), PyeongChang Kwangwoon_Do, Korea (South), 2019, pp. 260-264, doi: 10.23919/ICACT.2019.8701983.

[23] G. K. Getele, T. Li and J. T. Arrive, "The Role of Supply Chain Management in Healthcare Service Quality," in IEEE Engineering Management Review, vol. 48, no. 1, pp. 145-155, 1 Firstquarter,march 2020, doi: 10.1109/EMR.2020.2968429.

[24] K. Wan and V. Alagar, "Resource-aware Service-oriented Approach for Elderly Healthcare," 2018 5th International Conference on Systems and Informatics (ICSAI), Nanjing, 2018, pp. 1199-1205, doi: 10.1109/ICSAI.2018.8599300.

Towards a Standardization of Learning Behavior Indicators in Virtual Environments

Benjamin Maraza-Quispe[1], Olga Melina Alejandro-Oviedo[2]
Walter Choquehuanca-Quispe[3], Nicolás Caytuiro-Silva[4], José Herrera-Quispe[5]
Facultad de Ciencias de la Educación, Universidad Nacional de San Agustín de Arequipa, Arequipa-Perú[1, 2, 3]
Facultad de Ciencias e Ingenierías Físicas y Formales, Universidad Católica de Santa María, Arequipa-Perú[4]
Facultad de Ingeniería de Producción y Servicios, Universidad Nacional de San Agustín de Arequipa, Arequipa-Perú[5]

***Abstract*—The need to analyze student interactions in virtual learning environments (VLE) and the improvements this generates is an increasingly emerging reality in order to make timely predictions and optimize student learning. This research aims to implement a proposal of standardized learning behavior indicators in virtual learning environments (VLE) to design and implement efficient and timely learning analytics (LA) processes. The methodology consisted of a data management analysis that was carried out in the Moodle platform of the Faculty of Education Sciences of the National University of San Agustin of Arequipa, with the participation of 20 teachers, where qualitative online questionnaires were used to collect the participants' perceptions. The results propose a standard in terms of indicators of behavior in the teaching-learning process in EVA as they are: Preparation for learning, progress in the progress of the course, resources for learning, interaction in the forums and evaluation of resources. These were evaluated through learning analytics and show the efficiency of the proposed indicators. The conclusions highlight the importance of implementing standardized behavior indicators that allow us to efficiently develop learning analytics processes in VLE in order to obtain better predictions to make timely decisions and optimize the teaching-learning processes.**

Keywords*—Indicators; behavior; learning; analytical; environments; virtual

I. Introduction

Currently most students make use of VLE. We often do not consider the large amount of data left by students as a result of their interactions with these VLE's. Some of its advantages include improving educational decision making by identifying at-risk students and providing timely intervention to help students achieve success, improving instructional course designs, optimizing learning assessments, improving competency determination, and mapping curriculum and recommendations for learning improvement [1].

Currently in the scientific literature we do not find indicators of learning behavior in virtual learning environments that are standardized in order to be used to develop learning analytics. In this context, the following question is asked: To what extent is it possible to standardize learning behavior indicators in virtual learning environments?

According to [2] learning analytics can be defined as the measurement, collection, analysis, and presentation of data about learners and their contexts in order to understand and optimize learning and the environments in which it occurs. [3] learning analytics is the collection, analysis, use, and appropriate dissemination of actionable student-generated data for the purpose of creating appropriate cognitive, administrative, and effective support for students. [4] In other words, learning analytics is a tool that takes a snapshot of a given course by extracting data from the content management system for later analysis. In the same way [2] and [5] highlights LA's major contributions to literature. It includes the latest theories, findings, strategies, tools, and case studies, and focuses on the following uses:

- How to improve the performance of students and teachers.
- How to improve students' understanding of the course material.
- Assessing and Addressing the Needs of Students with Disabilities.
- How to Improve Rating Accuracy.
- How to allow instructors to assess and develop their own strengths.
- How to encourage more efficient use of resources at the institutional level.

The author in [4] highlights that learning analytics provides a centralized space for information across semesters, sections, instructors, students, and assignments. In this sense, the learning analysis process is unique in that it links large amounts of student-generated data to produce metrics or visualizations that can be used to improve the educational experience [6]. In accordance with [7] One of the main technologies used for the management of distance learning courses is the Learning Management System (LMS). Many higher education institutions have adopted the LMS; however, the negative perception of the LMS by faculty diminishes its potential for a system-wide approach to implementing a learning analysis strategy.

However, more recent research in the area of learning analytics, attempts to understand the reactions of students and learners when visualizing data and presenting it in the dashboard. For example, [8] consider that students and teachers are not able to interpret the information provided by the most common dashboards and, therefore, the effects on their

learning are non-existent and sometimes even negative. Therefore, as some researchers point out [9], [10], it is necessary to focus more on how to understand the meaning of the data provided by the dashboards, so that there is a positive influence on the student's own learning scenario. However, research on the pedagogical use of data visualization and its impact on the student's learning process is scarce. And this is precisely the aim of this research: To propose standards in terms of indicators of learning behavior to be analyzed in VLE in order to develop better predictions to optimize the teaching-learning processes.

II. State of the Art

A. Learning Analytics Models

Using a general morphological analysis methodology, analyze the discussions in the learning analytics communities; through this method they raise a model, which focuses on its fundamental structure [11]. This model has six dimensions: Actors, objectives, data, instruments, external constraints and internal limitations. From this model it can be highlighted that the fundamental objectives of learning analytics are reflection and prediction.

Another aspect that is worth highlighting, and that is related to its link with other related fields, is that the instruments dimension of learning analytics, is oriented to the conversion of the Big Data from its original state (that is, unstructured, complex, etc.), into useful information. Under this premise, learning analytics is nourished by analytical instruments from various fields, such as Machine Learning or classic statistical analysis.

A common factor, among the models presented by [12], [13], [14], is that they delimit the life cycle of the learning analytics process, from a common starting point, such as data collection, to a final stage that varies according to the model. Table I presents the stages that have been followed to make the example of sample analysis of the data obtained considering our proposal.

The model proposed by [14] has a systemic approach where, in addition to the analytical process, support resources are systematized. From this model it is highlighted that information collection is subject to the purpose of analysis, which can incorporate purely research, academic, institutional, management, etc. motivations. As for the analysis, the techniques and tools are varied and include social network analysis, neurolinguistic programming, prediction, risk assessment, support search or concept development, among others.

TABLE I. Stages of the Learning Analytics Cycle [14]

Model	Stages of the Learning Analytics Cycle
Siemens (2013)	1. Data collection and acquisition 2. Storage 3. Data Cleaning 4. Integration 5. Analysis 6. Representation and visualization 7. Action (intervention, optimization, alerts, etc.) 8. Restarting the process (loop)

III. Method

The methodology used presents a quantitative approach because we use data collection and analysis to answer research questions and test a previously established hypothesis and rely on numerical measurement, counting, and statistics are used to accurately establish indicators of learning behavior in virtual learning environments.

A. Objective

To propose standardized learning behavior indicators in Virtual Learning Environments in order to design and implement learning analytics processes to develop timely predictions and optimize learning.

B. Context and Sample

The research was conducted using the Moodle platform of the National University of San Agustín de Arequipa: https://aulavirtual.unsa.edu.pe/aulavirtual/ in the Faculty of Educational Sciences, where the courses are currently taught in the online mode. However, taking into account that the indicators of learning behavior to be observed should be standardized in order to carry out more precise analyses regarding the teaching-learning processes in the online modality, the research was developed with 20 teachers of this modality. The selection criteria were carried out through a simple random sampling of a total population of 70 teachers.

C. Procedure

Online questionnaires were used for the collection of qualitative data which were answered by the 20 teachers of the Faculty. Based on the observation of the data obtained through the IntelliBoard module and the survey developed, the indicators of learning behavior in EVA were standardized.

To validate the proposal, data collection was carried out through the IntelliBoard module installed in the university's Moodle platform and an analysis of this data was made considering the learning behavior indicators developed in our proposal.

D. Data Collection Instrument

The open-ended questionnaire was designed following the framework proposed by [15] and aims to understand teachers' perceptions of the most appropriate learning behavior indicators to be standardized. The questionnaire consists of 10 questions. The data obtained were analyzed according to previously stated analysis criteria. Also, data collection was done through the IntelliBoard module that was installed in the Moodle platform. IntelliBoard offers analytical and reporting services for education communities and institutions using the Moodle platform. IntelliBoard extracts static data collected in Moodle and presents consistent data in graph and report formats. With the power to convert this analytical data into simple, easy-to-read reports, IntelliBoard.net becomes the primary reporting tool [16].

IV. Analysis and Results

A. Selection of Learning behavior Indicators in EVA

According to the results obtained, the indicators of learning behavior directly affect the accuracy and credibility of the

prediction of student performance. Therefore, scientifically selecting effective learning behavior indicators is an important part of predicting student performance [17]. Because of the diversity of online learning behaviors, and the complexity of the correlation between behaviors, not all indicators of learning behavior can affect learning and data can be collected in a quantitative way. Therefore, based on existing research results, combining the characteristics of the online course, were classified into five stages: Preparation of learning, progress of the curriculum, resources for learning, interaction in the forums and evaluation of activities, in different dimensions and content, through them are obtained the indicators of learning behaviors related to online learning activities on the platform. Sixteen learning behavior indicators were selected for the study as shown in Table II.

In Table II, the review of resources is calculated by the total time of the learning resources (recommended time) divided by the time the student spends on the learning resources, reflecting the completion of learning. The amount of resource review refers to the number of resource review time divided by the difference in time between the last resource view and the first view, reflecting the students' concentration. The number of submissions is calculated by the number of test submissions divided by the difference in time between the last submission and the first submission.

B. Standardization of Learning behavior Indicators in Virtual Environments

Table III shows the standardization of the learning behavior indicators implemented from the data provided by the 20 teachers in 20 different courses.

Achievement levels are categories that classify students according to their performance where belonging to each of these levels of achievement describes the knowledge and skills developed by students. These achievement levels were implemented based on the International Standard Classification of Education according to UNESCO [19].

TABLE II. INDICATORS OF E-LEARNING BEHAVIOR OBSERVED

Learning process	Dimension	Content	Indicator
Preparation	Before class	Before class	Number of views of the course presentation
		Assistance	Number of accesses to the course
Curriculum progress	Learning Objectives	Related to	Number of revised course pages
			Time to review resources
Learning Resources	Gathering information	Importance of learning resources	Completion of resource review
			Number of resources reviewed
	Information processing	Persistence in reviewing resources	Number of resources reviewed repeatedly
			Number of resources reviewed after completion of course
	Collection and processing of information	Other learning resources	Access number to other resources
Interaction in the forums	Information about the publication	Interactive participation	Number of searches in the forum
			Number of publications in the forum
			Number of responses in the forum
Evaluation of activities	Information used	Amount of information completed	Number of published activities
			Average of activities developed
		Successful completion	time difference between the uploaded resource and the start of the activity
			Quantity of the shipment

TABLE III. INDICATORS STANDARDIZATION OF LEARNING BEHAVIOR INDICATORS IN VIRTUAL ENVIRONMENTS

Dimension	Learning behavior indicators	Achievement levels	Standardization of levels in %
Before class	Number of views to the course presentation	In start	0% - 20%
		In process	21% - 50%
		Satisfactory	51% - 80%
		Excellent	81% - 100%
	Number of accesses to the course	In start	0% - 30%
		In process	31% - 60%
		Satisfactory	61% - 80%
		Excellent	81% - 100%
Learning objectives	Number of revised course pages	In start	0% - 20%
		In process	21% - 50%
		Satisfactory	51% - 80%

		Excellent	81% -100%
Information gathering	Completion of resource review	In start	0% - 30%
		In process	31% - 60%
		Satisfactory	61% - 80%
		Excellent	81% -100%
	Amount of resources reviewed	In start	0% - 30%
		In process	31% - 60%
		Satisfactory	61% - 80%
		Excellent	81% -100%
Information processing	Number of resources reviewed repeatedly	In start	0% - 20%
		In process	21% - 50%
		Satisfactory	51% -80%
		Excellent	81% - 100%
	Number of resources reviewed after finishing the course	In start	0% - 20%
		In process	21% - 50%
		Satisfactory	51% - 80%
		Excellent	81% - 100%
Information collection and processing	Access number to other resources	In start	0% - 30%
		In process	31% - 60%
		Satisfactory	61% - 100%
		Excellent	81% - 100%
Publication information	Number of searches in the forum	In start	0% - 20%
		In process	21% - 50%
		Satisfactory	51% - 80%
		Excellent	81% - 100%
	Number of forum posts	In start	0% - 20%
		In process	21% - 50%
		Satisfactory	51% - 80%
		Excellent	81% - 100%
	Number of responses in the forum	In start	0% - 20%
		In process	21% - 50%
		Satisfactory	51% - 80%
		Excellent	81% - 100%
Information used	Number of published activities	In start	0% - 20%
		In process	21% - 50%
		Satisfactory	51% - 80%
		Excellent	81% - 100%
	Average of activities carried out	In start	0% - 30%
		In process	31% - 60%
		Satisfactory	61% - 80%
		Excellent	81% - 100%
Information used	time difference between the uploaded resource and the start of the activity	In start	0% - 20%
		In process	21% - 50%
		Satisfactory	51% - 80%
		Excellent	81% - 100%
	Number of shipments	In start	0% - 20%
		In process	21% - 60%
		Satisfactory	61% - 80%
		Excellent	81% - 100%

V. Validation of the Proposal

In order to validate the proposal in terms of indicators of learning behavior, an analysis of data obtained through the technique of linear regression, which is a supervised learning algorithm used in Machine Learning and statistics. In its simplest version, what we will do is "draw a line" that will indicate the trend of a set of continuous data. In statistics, linear regression is an approach to model the relationship between an independent scalar variable "X", and one or more independent explanatory variables "Y".

In Fig. 1, taking into account the indicator of learning behavior: Number of visits to the presentation or introduction of the course, we have that the highest number of visits to the introduction of the course is concentrated in an average of 10 students and the lowest number of visits is concentrated between 30 to 40 students, which will allow making timely decisions in order to improve the number of visits to the presentation of the course.

In Fig. 2, taking into account the indicator of learning behavior: Number of pages reviewed in the course, we have an average of 2.5 out of a total of 100 students have reviewed all the pages of the course, which will allow making timely decisions in order to improve the number of pages reviewed in the course.

In Fig. 3, taking into account the learning behavior indicator: Number of assignments sent, an average of 18 students has sent the highest number of assignments, while an average of 6 students has not sent any assignment, which will allow making timely decisions in order to improve the number of assignments sent.

In Fig. 4, taking into account the learning behavior indicator: Number of accesses to the course, 15 students have made their access an average of 75 times to the course and 5 students have made their access an average of 174 times, which represents a minimum amount, which will allow us to make timely decisions in order to improve the number of accesses to the course.

In Fig. 5, taking into account the learning behavior indicator: Time spent reviewing resources, a minimum average of students dedicates 12.5 hours to the review of resources, while a higher average of students dedicates a lower average of time to the review of resources; which will allow us to make timely decisions in order to improve the time each student dedicates to the development of the course.

In Fig. 6, taking into account the learning behavior indicator: Number of posts in the forum, we have that an average of 37 students make a greater number of posts, while an average of 6 students do not make posts; this will allow us to make timely decisions in order to improve the number of posts in the forum.

According to the proposed behavior indicators, the graphs show which values are concentrated among most records to be analyzed. Regarding behaviors, students interact with the elements of the learning context: Forums, resources, submissions. In the interaction of the digital elements the students generate traces. These traces identify the behavior of each student only and only in these digital elements. Somehow, this behavior forms patterns. The student adopts these patterns as a solution to learning situations. So, analyzing these patterns allows the extraction of learning indicators. One of the objectives of learning behavior indicators in VLE is to break or alter learning behavior patterns so that the student adopts healthier ones. The indicators help the teacher to know the learning patterns that the students adopt. At the same time, it allows to check which students leave the standard, do different actions and how, for example, to approach the tutorial actions. Establishing indicators offers a learning opportunity for both the student and the teacher. The teacher must be aware that behavior X is not the cause of effect Y. However, there may be a correlation. Discovering patterns of behavior helps the teacher to alert the student, test new behaviors, and check results [1].

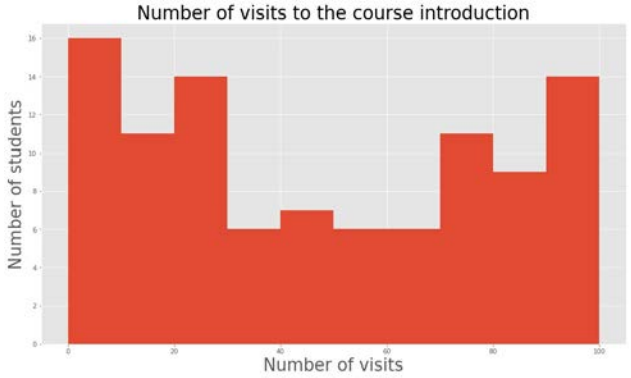


Fig. 1. Number of Visits to the Course Introduction.

Fig. 2. Number of Pages Reviewed in the Course.

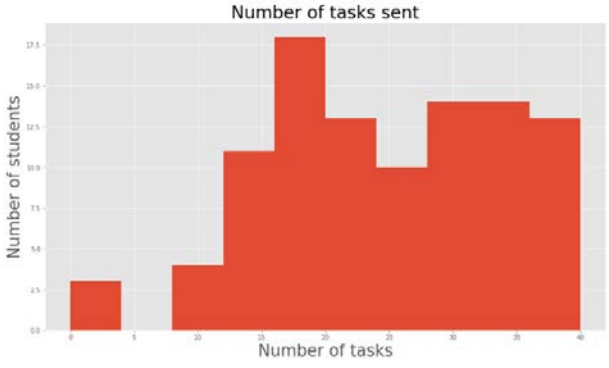


Fig. 3. Number of Assignments Sent.

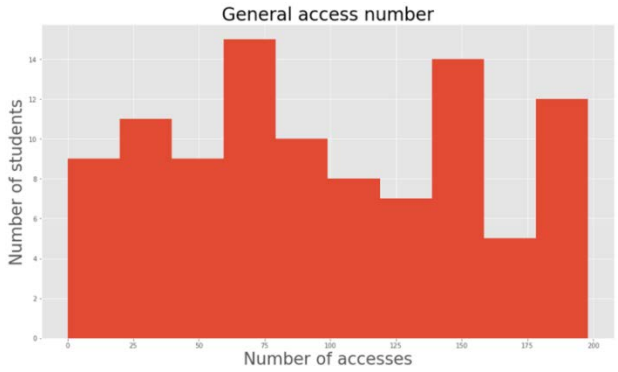


Fig. 4. Number of Accesses to the Course.

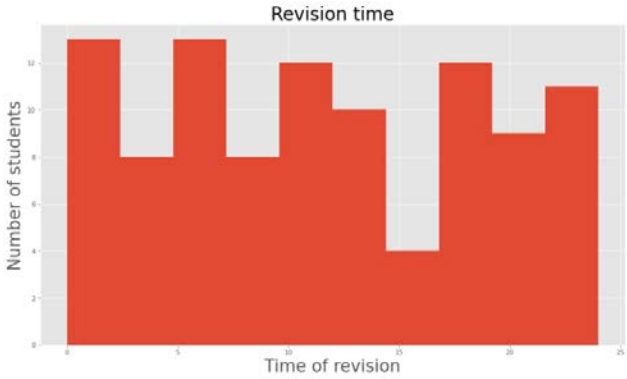


Fig. 5. Time Spent Reviewing Resources.

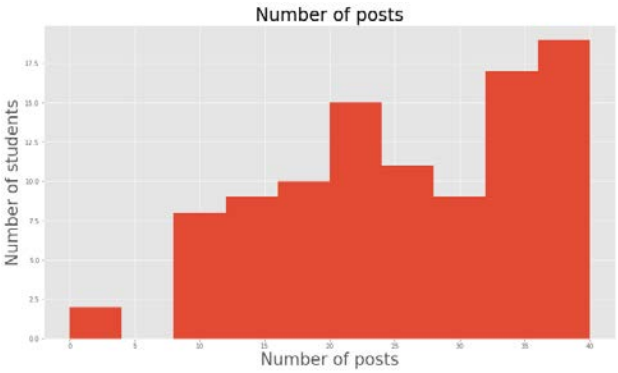


Fig. 6. Number of Posts in the Forum.

Table IV shows the data obtained by considering these standardized indicators where behaviors can be aggregated into patterns. The patterns can be aggregated into indicators. Indicators can be aggregated into predictions. Predictions will allow us to make timely decisions, [18]. A prediction result is a summary of interactions, patterns and behaviors. A prediction facilitates a reflection that allows us to follow, guide and tutor the student in his learning process.

In Fig. 7, we can see an example of Prediction in simple linear regression developed with the data obtained taking into account the standardized learning behavior indicators, which show us that if we take into account these indicators, reliable predictions can be made to make timely decisions regarding the improvement of the teaching-learning processes in virtual learning environments. [20].

TABLE IV. SUMMARY OF DATA OBTAINED CONSIDERING THE PROPOSED BEHAVIORAL INDICATORS

	Number of Resources reviewed	**Number of posts in the Forum**
Count	100.000000	100.000000
Mean	23.310000	23.640000
Std	9.465189	9.478865
Min	5.000000	3.000000
25%	15.000000	15.000000
50%	22.500000	23.000000
75%	32.000000	32.000000
Max	42.000000	40.000000

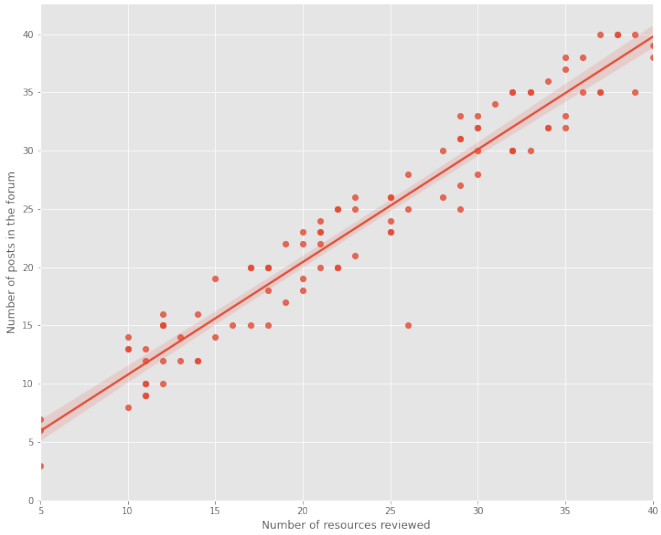


Fig. 7. Example of Data Analysis using the Linear Regression Technique using the Proposed Indicators.

VI. DISCUSSION

Although it is true that in the reviewed literature there is no information regarding the standardization of learning behavior indicators in virtual learning environments, it has been possible to standardize 16 online learning behavior indicators according to achievement levels: Initially, in process and satisfactory. 100% of the teachers surveyed express their agreement with these indicators, which can be applied to different contexts according to needs.

The results obtained through data management in the Moodle platform with the IntelliBoard module and the survey developed to teachers prioritize the importance of visualizing standardized educational data in order to develop learning analytics with more accurate predictions to optimize the teaching-learning process of students in a meaningful way.

VII. CONCLUSIONS

These indicators of learning behaviors in VLE are very important for self-regulation and reflection of students and teachers within their teaching and learning context. Likewise, teachers could provide very effective feedback by knowing the indicators of learning behavior in which they have weaknesses. That is why teachers considered that these data could help in the redesign of their courses.

After a validation process, it is concluded that there are 16 indicators that are prioritized for an effective analysis of the data using learning analytics techniques, which would allow the implementation of better predictions in order to make timely decisions to optimize the teaching-learning processes in virtual learning environments.

Learning Analytics (LA) has attracted a great deal of attention. In recent years as educational institutions and researchers are increasingly seeing the potential that LA has to support the learning process. LA approaches share a movement from data to analysis to action to learning. LA is an interdisciplinary field in which several related research areas converge.

In the future it is recommended to collect data from a larger sample of teachers, also collect qualitative data from students and then compare them between the two roles. In addition, it is required to perform more in-depth analysis tests using learning analytics techniques combined with Artificial Intelligence.

REFERENCES

[1] Maraza, B. (2016). Towards Personalized Learning in Virtual Environments. Virtual Campuses, 5(1), 20-29. From http://www.uajournals.com/ojs/index.php/campusvirtuales/article/viewFile/111/100.

[2] Calvet, L., & Juan, Á. (2015). Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. Revista de Universidad y Sociedad del Conocimiento, 98-112.

[3] Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. American Behavioral Scientist, 57, 1510-1529.

[4] Lindsey, B. (2016). Incorporating Learning Analytics into Basic Course Administration: How to Embrace the Opportunity to Identify Inconsistencies and Inform Responses. Journal of the Association for Communication Administration, 2-13.

[5] Larusson, J., & White, B. (. (2014). Learning Analytics: from Research to Practice. New York: Springer Science+Business Media.

[6] Clow, D. (2012). The learning analytics cycle: Closing the loop effectively. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 134-138. J. A.

[7] Bollenback, D., & Glassman, A. (2018). Big data in higher education: Adjunct faculty perceptions of learning analytics and their uses. Issues in Information Systems, 71-80.

[8] Corrin, L., & de Barba, P. (2015). How do students interpret feedback delivered via dashboard? Paper presented at the International Conference on Learning Analytics and Knowledge, Poughkeepsie, NY.

[9] Dawson, S., Gasevic, D., Siemens, G., & Joksimovic, S. (2014). Current state and future trends: a citation network analysis of the learning analytics field. Paper presented at the International Conference on Learning Analytics and Knowledge, Indianapolis, IN.

[10] Wise, A. F. (2014). Designing pedagogical interventions to support student use of learning analytics. Paper presented at the International Conference on Learning Analytics and Knowledge, New York, NY.

[11] Greller, W. y Drachsler, H. (2012). Translating Learning into Numbers: A Generic Framework for Learning Analytics. Educational Technology & Society, 15(3), 42-57.

[12] Martin, T. y Sherin, B. (2013). Learning Analytics and Computational Techniques for Detecting and Evaluating Patterns in Learning: An Introduction to the Special Issue. Journal of the Learning Sciences, 22(4), 511-520. doi: 10.1080/10508406.2013.840466.

[13] Picciano, A. (2012). The Evolution of Big Data and Learning Analytics in American Higher Education. Journal of Asynchronous Learning Networks, 16(3), 9-20.

[14] Siemens, G. (2013). Learning Analytics: The Emergence of a Discipline. American Behavioral Scientist, 57(10), 1380-1400. doi: 10.1177/0002764213498851.

[15] Pardo, A., Jovanovic, J., Dawson, S., Gasevic, D., & Mirriahi, N. (2017). Using learning analytics to scale the provision of personalised feedback. British Journal of Educational Technology. https://doi.org/10.1111/bjet.12592.

[16] Intelliboard. (2019). Intelliboard. Retrieved October 20, 2019, from https://intelliboard.net/about.

[17] Villagrá-Arnedo C J, Gallego-Durán F J, Llorens-Largo F. 2016. Improving the expressiveness of black-box models for predicting student performance[J]. Computers in Human Behavior, 72.

[18] Maraza-Quispe, B., Alejandro-Oviedo, O., Fernández-Gambarini, W., Cisneros-Chavez, B., & Choquehuanca-Quispe, W. (2020). Analysis of YouTube as a documentary research tool in higher education students. Publications, 50(2), 133-147. doi:10.30827/publicaciones.v50i2.13949.

[19] Instituto de Estadística de la Unesco. 2011. Clasificación Internacional Normalizada de Educación. CINE 2011. From https://unesdoc.unesco.org/ark:/48223/pf0000220782.

[20] Maraza-Quispe, B., Alejandro-Oviedo, O., Choquehuanca-Quispe, W., Hurtado-Mazeyra, A., Fernandez-Gambarini, W. (2019). e-Learning Proposal Supported by Reasoning based on Instances of Learning Objects. International Journal of Advanced Computer Science and Applications, Vol. 10, No. 10. From https://thesai.org/Downloads/Volume10No10/Paper_35-E_Learning_Proposal_Supported_by_Reasoning.pdf.

Ensemble Learning for Rainfall Prediction

Nor Samsiah Sani[1], Abdul Hadi Abd Rahman[2]
Afzan Adam[3]
Center for Artificial Intelligence Technology
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia, Bangi, Malaysia

Israa Shlash[4], Mohd Aliff[5]
Ministry of Agriculture
Babil Province Branch, Baghdad, Iran[4]
Instrumentation and Control Engineering
Malaysian Institute of Industrial Technology
Universiti Kuala Lumpur, Kuala Lumpur, Malaysia[5]

***Abstract*—Climate change research is a discipline that analyses the varying weather patterns for a particular period of time. Rainfall forecasting is the task of predicting particular future rainfall amount based on the measured information from the past, including wind, humidity, temperature, and so on. Rainfall forecasting has recently been the subject of several machine learning (ML) techniques with differing degrees of both short-term and also long-term prediction performance. Although several ML methods have been suggested to improve rainfall forecasting, the task of appropriate selection of technique for specific rainfall durations is still not clearly defined. Therefore, this study proposes an ensemble learning to uplift the effectiveness of rainfall prediction. Ensemble learning as an approach that combines multiple ML multiple rainfall prediction classifiers, which include Naïve Bayes, Decision Tree, Support Vector Machine, Random Forest and Neural Network based on Malaysian data. More specifically, this study explores three algebraic combiners: average probability, maximum probability, and majority voting. An analysis of our results shows that the fused ML classifiers based on majority voting are particularly effective in boosting the performance of rainfall prediction compared to individual classification.**

***Keywords*—Ensemble learning; classification; rainfall prediction; machine learning**

I. Introduction

Time-series forecasting has recently gained the research interest and has been explored in multiple domains like stock market, finance and climate change studies. Time-series forecasting refers to the analysis process of a sequence of data points containing successive measurements that are made within a specific time interval. The domains mentioned above are currently heavily reliant on time-series data [1-3] Climate change research is one of the domains that utilizes time-series forecasting to analyse the varying weather patterns statistically for a particular period [4]. The nature of climate change data representation across time is the key characteristic of climate change [5].

Weather forecasting is a subset of climate change research that predicts the atmosphere's state at a future time and location [4]. An important application of weather forecasting is rainfall prediction, which is heavily used in various large-scale activities such as food production planning, water resource management and others that rely on water. It is therefore crucial to ensure that rainfall predictions can be further improved, especially with respect to their accuracy and predictive performance, so that the proper preparation and planning of large-scale activities can be worked out beforehand.

Machine learning has made dramatic improvements and is a core sub-area of artificial intelligence. It also enables computers to discover themselves without being explicitly programmed. A set of machine learning algorithms can be used to obtain meaningful insights into the data that help make effective detection on phishing websites. However, it is still very far from reaching human performance. The machine still needs human assistance to predefine the algorithms on initialization. Several machine learning approaches for rainfall forecasting have been studied for various locations such as South Africa, China and other countries [6-9]. The classifiers that are used for rainfall prediction include the Naïve Bayes, decision tree, support vector machine, neural networks, random forest, genetic algorithm, support vector regression, M5 rules, radial basis neural networks, M5 model trees, and k-nearest neighbours [10-14].

This paper highlights the phishing webpage detection mechanism based on machine learning classification techniques. The rest of the paper is organized in the following manner: Section II presents the rainfall prediction methodology, Section III presents the utilization of machine learning classification techniques, Section IV presents the utilization of ensemble machine learning techniques, and Section V presents the experimental results gained after the implementation of the ensemble classification methods in the rainfall datasets.

II. Methodology

Machine learning is one of the most exciting recent technologies. Machine learning had been positioned to address the shortages of human cognition as well as information processing, specifically in handling large data, their relations and the following analysis [15-16]. In general, machine learning studies the research and algorithms construction that can learn from, and derive predictions about data [17-18]. Therefore, the machine learning approach is selected to predict the rainfall.

The research methodology we used in our study can be segregated into four distinct phases. The first phase is the dataset phase, in which we manually identify the data for this study by analysing their sources, amount, and other details. Next, the pre-processing phase prepares the data for further processing by cleaning the data (i.e., addressing missing

values) and normalizing the data to limit the values to specific ranges. The pre-processed data are then used in the third phase to comparatively analyse the five ML techniques to identify the best technique from the five ML classifiers that are noted above. The fourth and final phase focuses on configuring the ensemble method to carry out assessment on the performance of the entire algorithm. Each of these four phases is described further in the subsections that follow.

A. Dataset

The dataset was obtained from the Drainage and Irrigation Department, and the Malaysian Meteorological Department. The dataset consists of 1,581 instances and was organized into two classes. The first is the 'active rainfall' class, containing 428 instances, and the remaining instances are grouped as 'no rainfall'.

The obtained data on description and location are illustrated in Table I. The features displayed in the dataset include the relative humidity, rainfall, temperature, flow, and water level. The feature details are as described in Table II. Table III provides the detailed measurements for each feature.

TABLE I. DETAILS OF THE DATASET

Source	Daily data	Station number	Station name
Malaysian Meteorological Department	24 hour mean temperature	48650	KLIA Sepang
	24 hour mean relative humidity	2917401	Sungai Langat at Kajang Selangor
	Daily total rainfall	2917112	Kajang at Hulu Langat
	Daily means water level	2917401	Sungai Langat at Kajang Selangor

TABLE II. FEATURE DETAILS

Feature	Valid records	Missing values
Temperature	1581	0
Relative humidity	1572	9
Flow	1464	117
Rainfall	1569	12
Water level	1464	117

TABLE III. MEASUREMENT FEATURE DETAILS

Attribute name	Attribute type	Attribute metre
Temperature	Continuous	°C
Humidity	Continuous	Percentage of relative humidity, %
Rainfall	Continuous	mm
River flow	Continuous	m3/s
Water level	Continuous	ms
Class	Nominal	Rainfall – yes Rain off - no

B. Pre-processing

As noted above, the pre-processing phase ensures that the available data are prepared for further processing in subsequent phases. Here, raw data are usually negatively impacted by noise or incomplete information. The pre-processing phase is a crucial stage in enhancing the improvement of the prediction process by ensuring the data are regularized and filtered beforehand [15], [19]. Therefore, we applied two rather common pre-processing subtasks: cleaning and normalization. In this study, Waikato Environment for Knowledge Analysis (Weka) is used as a tool to perform the pre-processing task. Weka is java-based machine learning software that is developed by the University of Waikato, New Zealand, and it has various types of machine learning algorithms and operates on an open source license. It also provides various visualization tools for data analysis as well as predictive modelling.

C. Cleaning

In the cleaning task, the data obtained are found to contain missing values, which are represented by characters such as '?'and '*'. In fact, such missing values can cause errors in the prediction process. Therefore, these missing values must be addressed. Table IV illustrates a sample of data containing missing values. A mean average mechanism is then used to populate the missing values. The mean average functions are obtained by summing all instances of an attribute that is selected and then dividing the sum by the number of records. in the second attribute (humidity), for example, the missing values are filled by firstly adding all instances (87.6, 88.9, 84.7, 85.2, 88.3, and 84.2), and then dividing the results by the total number of instances, which in this case is 6. Table V shows the mean average for each attribute.

TABLE IV. DATA WITH MISSING VALUES

Temperature	Humidity	Rainfall	Flow	Water level
27.9	85.3	?	3.94	22.37
27.3	86.2	?	3.82	22.36
27.8	83.6	*	3.67	22.34
27.7	*	*	10.68	22.54
27.3	84.2	11.4	11.93	22.61
27.4	82.8	40	14.6	22.69
27.3	82.3	8.9	20.24	22.89
26.8	85.8	7.7	14.04	22.68
27.3	81.4	*	11.1	22.57
24.7	90.3	*	10.62	22.54
26.0	86.2	*	10.23	22.53
27.7	-1.1	*	8.73	22.45
28.6	73.4	?	?	?
29.3	68.3	?	?	?
29.1	67.8	5.7	?	?
28.8	67.9	11.3	?	?
28.9	64.1	10.9	?	?

TABLE V. AVERAGE FEATURES

Attribute	Average
Humidity	27.528
Rainfall	81.265
River flow	5.477
Water level	11.837

D. Normalization

In the normalization task, values are limited within a specific interval, in which the interval facilitates the prediction since the values are reduced into specific ranges. Normalization is crucial for particular algorithms like ANN and SVM. Table VI illustrates values prior to normalization. As shown in Table VI, the values are found to vary greatly, although these values are seen to decompose around the 20s and 80s for the first two features and around the 10s for the remaining three features. To unify these values, we chose an interval range of -1 to +1 and use the normalization mechanism that was introduced by [20] as defined in (1):

$$\gamma = \frac{(y_{max} - y_{min}) \times (x - x_{min})}{(x_{max} - x_{min})} + y_{min} \quad (1)$$

In Eq. (2.1), x is the data that requires to be normalized. X_{min} is the minimum value for all data, and X_{max} refers to the maximum value of all the input data. On the other hand, Y is the normalized data, while Y_{min} is the ideal minimum value. Y_{max} refers to the desired maximum value. Following the normalization task, all values for the five features are converted to be in the range of -1 to +1. As illustrated in Table VII, the data have been normalized to prepare for further processing.

TABLE VI. VALUES BEFORE NORMALIZATION

Temperature	Humidity	Rainfall	Flow	Water level
22.3	87.6	2.31	2.78	2.79
26.4	88.9	5.74	4.29	5.74
22.9	84.7	1.68	6.78	1.25
27.8	85.2	5.03	5.46	4.56
24.1	88.3	5.03	4.29	4.56
26.5	86.4	5.03	4.29	4.56
26.9	86.4	2.69	1.64	6.47
29.3	84.2	10.4	2.14	8.46
21.2	86.4	5.03	4.65	4.56

TABLE VII. NORMALIZATION TASK

Temperature	Humidity	Rainfall	Flow	Water level
-0.728	0.446	-0.855	-0.556	0.572
0.283	1	-0.068	0.031	0.245
-0.580	-0.756	-1	1	-1
0.629	-0.512	-0.392	1	-1
-0.283	1	-0.392	0.760	-1
0.308	1	-0.392	0.760	-1
0.407	1	-1	-1	-0.02

E. Evaluation Metrics

For the purpose of evaluating the method proposed, the common information retrieval metrics are employed. The evaluation is carried out through the use of the common information retrieval metrics of recall, F-measure, and precision. Our model predicts 2 classes (rain or not), so sensitivity or recall can reflect the ratio of rain and no-rain correctly identified by the model. The R2, SSE, and MSE are better for continuous values, while our model does not predict such an output. Precision evaluates the true positives (TP) that are classified correctly and the false positives (FP) that are entities classified incorrectly, which could be computed using (2):

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|} \quad (2)$$

The recall parameter is used in assessing the true positives (TP) with respect to the false negatives (FN), which are unclassified entities. This evaluation is calculated as shown in (3):

$$\text{Recall} = \frac{|TP|}{|TP| + |FN|} \quad (3)$$

Lastly, the average of the recall and precision, which is the F-Measure is computed as follows:

$$\text{F-measure} = 2 \cdot \frac{precision \times Recall}{precision + Recall} \quad (4)$$

III. MACHINE LEARNING MODELS

Numerous learning methods are selected in this study to benchmark the rainfall prediction performance. These are NB, C4.5, SVM, ANN, and RF, which are all supervised learning methods. A notable aspect of the supervised machine learning methods is that they select suitable methods together with parameters and features that are deemed suitable [21-25]. Two main experiments were carried out in order to evaluate the performance of the classifiers. The first identifies the best parameterization set of each classification model to be employed, since the model has a few alternatives as well as options, which would affect the method's success. Different tuning parameters are used to tune every classifier in order to yield highly accurate results. A series of experiments were carried out to obtain the optimal values of each classifier. The performances between the five classifiers are then evaluated and compared. The second experiment analyse the true performance of the classifiers for rainfall prediction.

A. C4.5 Algorithm

In this section, the J48 decision tree, which is included in Weka, is formed based on the C4.5 decision algorithm. C4.5 is one of the most effective classification methods [26]. Table VIII shows the pseudo code of the algorithm. A decision tree is generated by C4.5 in which every node splits the classes with reference to the information. Splitting criteria is selected based on the attribute having the highest normalized information gain. For example, our dataset contains temperature, humidity, rainfall, river flow, and water level. The C4.5 techniques first explores these features to determine which feature is the best for splitting data (a feature with high information). The feature is then used to split the data into the next feature until it reaches the last destination.

TABLE VIII. C4.5 PSEUDO CODE

Input: Dataset D
1. Tree = {a}6
2. If D is 'pure' OR other stopping criterion met then
3. Terminate
4. End if
5. For all attribute a ϵ do
6. Compute information-theoretic criterion if split on a
7. End for
8. a_{best}= Best attribute according to above computed criterion
9. Tree = Create a decision node that tests a_{best} in the root
10. D_u= Induced sub-datasets from D based on a_{best}
11. For all $D_u$do
12. $Tree_u$= J48 (D_u)
13. Attach $Tree_u$to the corresponding branch of Tree
14. End for
15. Return Tree

The evaluation was performed using confidence factor, *MinNumObj*, and *Numfolds* parameters. The splitting mechanism splits the dataset into 60% for training and 40% for testing, and the evaluation was performed with the use of the common information retrieval metrics of the recall, precision, and F-Measure. Table IX illustrates the algorithm results. As shown in Table IX, several parameter values are used. The best results are achieved when the parameters are (*confidence factor* = 0.5, *MinNumObj* = 4, and *Numfolds* = 5), which results in a precision of 71.3%, a recall of 74.2% and 72.7% F-Measure.

B. Naïve Bayes

Naïve Bayes is classified as supervised machine learning method that belongs to the probabilistic classifiers family which applies Bayes theory to the independence assumption between features [27]. As a matter of fact, Naïve Bayes identifies the probability of every feature by calculating the assumptions. Table X depicts the pseudo code of the Naïve Bayes algorithm.

This section evaluates the Naïve Bayes technique being applied using Weka. For every known class value, NB computes every attribute conditional probability on the class value. Later, it obtains the joint conditional probability for the attributes using the product rule. This process is followed by the use of Bayes rule to obtain the class variable's conditional probabilities. After completing this process for each class value, the class having the highest probability is reported.

The parameter tuning experiment was carried out in identifying the best parameters from a few different options available. There are two parameters that affect the performance of the NB classifier: debug and use Kernel Estimator. In this study, debug and use Kernel Estimator are tested on two different values (True and False) for choosing the optimal parameter of the NB classifier. When use Kernel Estimator=True, it means the NB model employs a kernel estimator for numeric attributes as opposed to a normal distribution. Moreover, if the debug parameter is set to False, it means that the classifier may not output any extra information to the console. As shown in Table XI, the best outcomes are attained when the parameter (*debug* = False, use *Kernel Estimator* = True) obtains 65.5% precision, 71.5% recall and 65.5% F-Measure.

TABLE IX. RESULTS FOR THE C4.5 PARAMETER TEST

C4.5 Parameter			**Precision**	**Results Recall**	**F-Measure**
Confidence factor	**MinNumObj**	**Numfolds**			
0.25	2	3	70.1%	73.4%	70.1%
0.5	4	5	71.3%	74.2%	72.7%
0.7	6	7	70%	73.4%	71.3%

TABLE X. NAÏVE BAYES PSEUDO CODE

Input: Dataset D
For each Feature f Compute the assumptions of f values based on class label 1 End for For each Feature f Compute the assumption of f values based on class label 2 End for Prediction class = Maximum (assumption label 1, assumption label 2) Repeat for all features

TABLE XI. RESULTS FOR THE NAÏVE BAYES PARAMETER TEST

NB Parameter		**Precision**	**Recall**	**F-measure**
debug	**Use Kernel Estimator**			
False	True	65.5%	71.5%	65.5%
False	False	62.9%	68.4%	64.4%

C. Support Vector Machine

This section discusses the evaluation of the support vector machine method by using the libSVM package in Weka. Some parameters have to be fitted to the data to avoid errors due to the SVM being very sensitive to the presence of any inappropriate parameters. The support vector machine is a method that divides data into two sections with the use of a hyperplane[12]. This division process independently addresses every class label, and this could be carried out through classifying the data into class *x* and not class *x*, and then further categorizing the data into class y and not class *y*, where *x* and *y* are the two class labels. The classification is carried out by calculating the distance between every data point and the hyperplane's margin. Table XII contains the description of the algorithm.

TABLE XII. SUPPORT VECTOR MACHINE PSEUDO CODE

Initialize $y_i = y_1$ for iϵI
Repeat
Compute SVM solution w, b for dataset with imputed labels
Compute outputs f_i= (w, x_i) + b for all $x_i$in positive bags
Set y_i= sgn(fi) for every iϵI and $y_1 = 1$
For (every positive bag Bi)
If ($\sum_{i\epsilon I}^{Bi} \frac{1+y_i}{2} = 0$)
Compute I = $argmax_{i\epsilon I} f_i$
Set $y_i = 1$
End
End
While (imputed labels have changed)
Output (w, b)

The SVM algorithm uses a kernel that is a set of mathematical functions to allow for data classification in a higher dimensional space when such data could not be linearly separated in a lower dimensional space. Various kernel functions are available to govern the above, like linear, polynomial, nonlinear, radial basis function (RBF) and sigmoid. The SVM can be further classified into two categories, namely, the C SVM and the *nu* SVM. C and *nu* refer to the regularization of the parameters that aids in implementing a penalty on the misclassifications that happen when the classes are separated. C ranges from 0 to infinity and *nu* is always between 0-1.

The parameter evaluation was performed using the SVM type and kernel type parameters. The default parameter value for the SVM type is set as *nu*-SVC, which uses a range between 0 and 1 to represent the lower and upper bounds of the number of examples that are support vectors which lie at the wrong side of the hyperplane. Moreover, a number of different parameterization combinations are tested under the kernel type. Here, each kernel type parameter is changed one at a given time while SVM type is remained the same consistently so that any differences due to the kernel type parameter change can be recorded. Four techniques have been established in the literature already, namely, the linear, radial basis function (RBF), polynomial, and sigmoid, and these four are extensively tested in our experiments. Table XIII illustrates the results of the algorithm. The best results are reported when the parameters are (SVM type = *nu*-SVC, kernel type=RBF), where the precision is reported to be 71.1%, the recall is 72.8% and the F-Measure is 69.1%.

D. Neural Networks

Neural networks were originally motivated by modelling machines that replicate the brain's functionality. Every neural unit is linked to many others. Links could either be inhibitory or enforcing in nature, with regards to their activation state effect of the connected neural units. Every individual neural unit could have a summation function combining its input values [28]. This algorithm is used in regression, classification, prediction and clustering [28]. Table XIV depicts the pseudo code of the algorithm.

There are two parameters that significantly affect the neural network performance classifiers: the number of hidden layers and the value of learning rate. To get the optimal hidden layer value, a range of values are tested from 2 to 10 (at an increment of 2), and the learning rate is tested on five different values from 0.02 to 0.10 (at an increment of 0.02). As shown in Table XV, the best outcomes are attained when the parameters (learning rate = 0.02 and the hidden layer =4) obtain 72.7% precision, 74.5% recall and 73.2% F-Measure. Therefore, the optimal parameters of the ANN are set as the learning rate = 0.02 and the hidden layer=4. The parameter evaluation was performed with 100 iterations.

E. Random Forest

Random forest is a method employed for many purposes, including regression, classification, and prediction. This method is an ensemble of decision trees aiming to construct, within the training data, a multitude of decision trees and generate the class as the output. Table XVI depicts the algorithm's pseudo code.

The random forest classifier is tuned using the maximum depth of the tree (Max Depth) and the number of features to randomly investigate (Num Features) and the number of tree (Num Tree) parameters. The Max Depth, Num Feature and Num Tree is tested on three different values which are (1, 5 and 10), (0, 3 and 5) and (10, 12 and 15), respectively. Experimental outcomes reveal the classification performance of the RF classifier is increased when the depth, the number of features and the number of tree increase. The obtained parameter tuning result is reported in Table XVII. The best results are reported when the parameters are (Max Depth=10, Num Feature=5, Num Tree=15), where the precision is reported to be 71.3%, the recall is 74.4% and the F-Measure is 70.7%. Thus, the optimal values are Max Depth=10 Num Feature=5 and Num Tree=15.

TABLE XIII. RESULTS FOR THE SUPPORT VECTOR MACHINE PARAMETER TEST

SVM parameter		**Precision**	**Recall**	**F-Measure**
SVM type	**Kernel type**			
nu-SVC	Linear	53%	71.8%	61.3%
nu-SVC	Polynomial	53%	68%	61.3%
nu-SVC	RBF	71.1%	72.8%	69.1%
nu-SVC	sigmoid	69.5%	70%	65.2%

TABLE XIV. NEURAL NETWORKS PSEUDO CODE

1. For iteration = 1 to T
2. For e = 1 to N (all examples)
3. X = input for example e
4. Y = output for example e
5. Run x forward through network, computing all $\{a_i\}$, $\{in_i\}$
6. For all weights (j, i)
7. Compute $\Delta_i = \{(y_i - a_i) \times g'(in_i)\, g'(in_i) \sum_k^i w_{i,k}\, \Delta_k\}$
8. Repeat

TABLE XV. RESULTS FOR THE NEURAL NETWORKS PARAMETER TEST

ANN Parameter		**Precision**	**Recall**	**F-measure**
Learning rate	**Hidden Layer**			
0.02	2	72.7%	74.5%	73.2%
0.04	4	72.3%	73.6%	72.8%
0.06	6	72.3%	73.3%	72.7%
0.08	8	72.4%	73.3%	72.6%
0.10	10	72.4%	73.5%	72.9%

TABLE XVI. RANDOM FOREST PSEUDO CODE

1. For simple Tree T
2. For each node
3. Select m a random predictor variable
4. If the objective function achieved (m=1)
5. Split the node
6. End if
7. End for
8. Repeat for all nodes

TABLE XVII. RESULTS FOR THE RANDOM FOREST PARAMETER TEST

RF Parameter			Precision	Recall	F-Measure
Max Depth	Num Feature	Num Tree			
1	0	10	68%	69.9%	68.7%
5	3	12	53%	72.8%	61.3%
10	5	15	71.3%	74.4%	70.7%

F. The Performance Evaluation of Different Classifiers

Extensive parameterization tests were performed in quantifying each parameter's influence for the optimization of the classification models. Numerous parameters crucial to maximizing the model's performance are selected from the tests. Whereas, other parameters are classified as less sensitive. Based on the finalized parameters, the classification models are properly executed to quantify their performance on rainfall prediction. Five machine learning methods are identified in the study. They are the Naïve Bayes (NB), C4.5, neural network (ANN), support vector machine (SVM), and random forest (RF). This phase quantifies the performance of the mentioned machine learning methods and determines the best overall methods for rainfall prediction. From the observations during the predictive studies, the results revealed the most effective classification to be the ANN and that the NB yielded the weakest result. Hence, the ANN this time is benchmarked against the ensemble method. Table XVIII shows the comparison of the five classifiers.

TABLE XVIII. RESULTS FOR THE CLASSIFICATION MODELS

Name of Classifier	Precision	Recall	F-Measure
SVM	71.1%	68%	69.1%
C4.5	71.3%	74.2%	71.3%
ANN	72.7%	74.5%	73.2%
NB	65.5%	71.5%	65.5%
RF	71.3%	74.4%	70.7%

The performance levels of ML-based predictions vary between the studies, although a neural network classification technique has a slight performance advantage compared to other classifiers. With individual ML classifier techniques for rainfall predictions already extensively documented, the fusion of various ML classification techniques based on an ensemble methodology presents an opportunity to tap any possibility to improve prediction performance. Apart from that, the varying performance levels of such techniques create space for improvement through the combination of various methods or improving techniques.

IV. ENSEMBLE METHODS FOR MODEL PREDICTION

The combination of multiple classifiers that result in one subsequent model is known as the ensemble model. Recently, ensemble techniques have been increasingly utilized to improve the prediction performance of classification tasks [29]. In general, there are three common issues faced in most single classification techniques that can be improved when using multiple classifier instances.

1) Statistical reasoning: In the event the training data amount is not sufficient, a learning algorithm extracts a weak hypothesis. The combinations of many classifiers, however, have a tendency to find a stronger hypothesis.

2) Computational reasoning: An appropriate hypothesis for an individual classifier (such as neural networks) may be more difficult as well as time consuming. Combining multiple classifiers (experts) with an appropriate parameterization (considering speed, efficiency and accuracy) may provide a better hypothesis while reducing the computation time through the enforcement of each classifier's strength in this case.

3) Representational reasoning: An individual classifier at times could not represent true hypothesis in the hypothesis space. In the case of ensemble methods, the formation of weighted sum of the hypotheses from the hypothesis space expands the hypothesis space in providing a hypothesis that is more presentable [29].

Ensemble methodology basically works by weighting numerous individual classifiers and later combining them in order to obtain a new classifier that theoretically outperforms their individual instances. The use of different classifiers from various different learning algorithms is an effective method in addressing the diversity among classifiers since it has the potential in minimizing errors or increasing the prediction performance by basing on diverse approaches [29]. The purpose of ensemble methodology is therefore to build a predictive model through the integration of multiple models.

Various fields of study have reported successful outcomes from the use of the ensemble method, such as in healthcare, information retrieval and statistics. Research of the ensemble method has greatly increased from the 1990s onwards. In order to improve the single model's predictive performance. Based on the suggestion [30], an ensemble of neural networks configured similarly. Reference [31] laid out the foundations for the award-winning AdaBoost. Reference [31] and [32] algorithm revealed that through the combination of weak classifiers, a strong classifier in the probably approximately correct (PAC) sense can be generated. Reference [33] proposed a novel ensemble health care decision support method to assist an intelligent health monitoring system that utilizes a Meta classifier voting system made up of three base classifiers, which are the C4.5, random forest and random tree algorithms [32]. Reference [32] employed the ensemble neural network for breast cancer diagnosis, where the researchers combined several neural network outputs that are fused to construct an ensemble output using the simple averaging algorithm. From the study, they found that the ensemble neural network improved the generalization with less false positive malignant diagnoses while accelerating the learning process. Furthermore, [34] adapted multiple neural networks as a method to improve the robustness of predictions. They used several methods such as linear combinations and stacked generalizations to combine member networks. From the study, they found that two combination methods, i.e., selective combination and network combination with various structures, are the best performers that greatly improved model performance.

Rainfall forecasting gained the attention of many researchers' as it has interesting challenges represented by the complexity that lies beneath predicting specific factors that are linked to rainfall like wind and humidity [21-23], [35-36]. Current techniques for rainfall prediction utilize only individual classifiers such as neural networks [21], the k-nearest neighbours [22], support vector machine [22] and others. Based on encouraging results from the ensemble methods application in various fields, the ensemble classification technique is applied to rainfall prediction by leveraging three linear algebraic combiners: majority voting, average probability, and maximum probability. Since there have not been any applications of the ensemble method for rainfall forecasting in Malaysia, this research should serve in addressing numerous supervised learning methods for the above case.

To assess the performance of our ensemble classification for rainfall prediction, we first fused all five machine learning models (i.e., the Naïve Bayes, C4.5, neural network, support vector machine, and random forest). Their combinations are based on three linear algebraic combiners, which are majority voting, average probability, and maximum probability. Equation (7), (8) and (9) describe their mathematical derivations. Voting is essentially the general blueprint for combining classifiers into ensembles. Voting schemes are divided into unweighted as well as weighted voting schemes. Unweighted schemes are including maximum probability, minimum probability, the product of probabilities, majority voting, and the average of probabilities. Whereas, weighted schemes encompass simple weighted voting, best–worst weighted voting, rescaled weighted voting, and quadratic best–worst weighted voting [32-34].

This study only focuses on unweighted voting schemes. In principal, the *k* base classifier's binary outputs are being combined such that the output of the ensemble is chosen based on the highest number of votes. Any of the unweighted schemes are used to guide the classifiers. Equation (5) highlights the basic form of the classification ensemble calculation.

$$H(X) = argi = 1 \dots n\ max\big(LCi(X)\big) \quad (5)$$

Here,

i = 1..., *n* is the number of classes and

J = 1...,*m* is the number of classifiers that are contained in ensemble method.

LCi(X) is thus represented as any combination scheme in determining the final output of the classifier ensemble.

A. Majority Voting

A classification of unlabelled instances is performed in this combining scheme based on the class having the highest (the most frequent) number of votes. This method is employed most of the time as a combination technique to compare newly proposed techniques [37]. The majority voting is defined by (6) as follows:

$$H(X) = arg\ i = 1 \dots n\ max\ \{Si = \sum_{j=1}^{m} I(hi(X)) = Y \quad (6)$$

B. Average of Probabilities

Every classifier output a probability distribution vector over all classes that are relevant in the probabilistic approach, as shown by (7). The individual probability values are averaged (or summed) by all classifiers for every class, and the class that yields the maximum value is chosen at last [32].

$$LCi(X) = \frac{1}{m}\sum_{j=1}^{m} Pj(wi \vee X) \quad (7)$$

C. Maximum Probabilities

The maximum probability approach is almost identical to the average probability approach described above [32]. Here, the only difference is in the selection of a probability with the maximum value, as highlighted in (8) below:

$$LCi(X) = maxj = 1 \dots m\{Pj(wi \vee X)\} \quad (8)$$

V. Results and Discussion

Table XVIII shows the comparison of the five classifiers using the three metrics based on the test dataset. From the table, the neural network outperforms the other techniques with a precision of 72.7%, a recall of 74.5%, and an F-Measure of 73.2%. The predictive results obtained from the neural network will be compared to the ensemble rainfall prediction approach.

We further discuss the experimental results that are obtained by applying the ensemble method with the three unweighted combiners (majority voting, average probability, and maximum probability) for rainfall prediction. This combination can be used to combine any of the five classifiers (Naïve Bayes, C4.5, support vector machine, neural networks, and random forest). The ensemble model works by combining classifiers from both groups of 'weak' and 'strong' classifiers, thereby forming an ensemble. Thus, in ensemble terms, the classifiers are weak learners, while the ensemble model is a strong learner. The evaluation of the ensemble methods is performed by the use of common information retrieval metrics as follow: recall, precision and F-Measure. The outcomes are based on a similar splitting mechanism for the dataset of 60% as training data and 40% as testing data.

Tables XIX, XX and XXI highlight the test data results from multiple combinations and the three ensemble methods based on the selected metrics. The single ANN classifier is also included in the comparison for benchmark purposes. For the precision metric, Table XIX demonstrates that the combination of the SVM, C4.5 and ANN methods via the majority voting scheme yielded the highest precision accuracy at 76%. This result is followed closely by the same combined ML tools but with the average probability scheme close to 75% accuracy. There is generally a 2% to 3% increase in precision accuracy if the best ensemble methods are to be compared to the single ANN classifier. On the other spectrum, the full or 4 ML tool combination that was tested resulted in disappointing precision accuracy for most of the weighting schemes, and was far below the 73% threshold that was posed by the single ANN classifier, except for the full combined ML tools with the maximum probability ensemble scheme, which scored an unusual 71% accuracy for such combinations.

TABLE XIX. COMBINATION OF TOOLS AND ENSEMBLE RESULTS BASED ON PRECISION METRIC

Combination	Precision		
	Majority Voting	Average Probability	Maximum Probability
Combination of (SVM, ANN, NB, C4.5, and RF)	53%	53%	71%
Combination of (ANN, NB, C4.5, and RF)	53%	53%	53%
Combination of (SVM, C4.5, and ANN)	76%	75%	71%
Combination of (SVM, C4.5, and NB)	74%	73%	71%
Combination of (NB and ANN)	70%	67%	67%
Single classifier (ANN)	73%		

For the recall accuracy, Table XX highlights a similar pattern to the precision metric, whereby the combination of the SVM, C4.5 and ANN tools and the use of the majority voting and average probability schemes scored the highest recall accuracy at 77%. This is a 2% increase from the single ANN classifier, which was 75%. Except for the two combination or ensemble schemes that were mentioned above, the remaining combination or ensemble schemes all performed slightly worse than the ANN classifier. Table XXI highlights the F-Measure accuracy for the employed classifiers and combiners. Again, the same combination of the SVM, C4.5 and ANN tools based on both the majority voting and average probability ensemble schemes yield the best accuracy as compared to the other classifiers. It is noted that the remaining fusion classifiers and ensemble schemes scored well below the 73% of the ANN classifier.

From the experiments, the fusion of the classifiers was shown to generally boost prediction diversity without compromising the individual prediction strengths of the individual classifiers. However, care needs to be considered, as not every fusion strategy works at improving performance over single classifiers. This is demonstrated by the fact that in all 3 performance metrics tests, the fusion of the 4 classifiers and all classifiers degrade the performance accuracy regardless of the ensemble scheme that is chosen. On the other hand, the selected fusion classifiers based on the majority voting scheme are superior to the single classifiers. Particularly, a combination of three tools with the minimum presence of the SVM and C4.5 algorithms ensure that superior performance can be achieved. Table XXII illustrates the confusion matrix for a two-class classifier (i.e., Rain and No Rain). The matrix is a summary of the prediction results that are obtained from the best ensemble method (i.e., majority voting) for the rainfall classification problem on the test dataset (i.e., 632 instances). In the context of our study, the entries in the confusion matrix carry the following meaning: true positive (TP) indicates the number of instances that correctly predict that it will rain, which is equivalent to 438 days; true negative (TN) shows the number of instances that correctly predict that it will not rain, which is equivalent to 39 days; and false positive (FP) shows the number of instances that incorrectly predict that it will rain, which is equivalent to 133 days. FP is also known as false positive predictions. Finally, false negative (FN) indicates the number of instances that incorrectly predict that it will not rain, which equivalent to 22 days and is otherwise known as false negative predictions.

TABLE XX. COMBINATION OF TOOLS AND ENSEMBLE RESULTS BASED ON RECALL METRIC

Combination	Recall		
	Majority voting	Average probability	Maximum probability
Combination of (SVM, ANN, NB, C4.5, and RF)	73%	73%	68%
Combination of (ANN, NB, C4.5, and RF)	73%	73%	73%
Combination of (SVM, C4.5, and ANN)	77%	77%	68%
Combination of (SVM, C4.5, and NB)	76%	73%	68%
Combination of (NB and ANN)	73%	72%	72%
Single classifier (ANN)	75%		

TABLE XXI. COMBINATION OF TOOLS AND ENSEMBLE RESULTS BASED ON F-MEASURE METRIC

Combination	F-Measure		
	Majority voting	Average probability	Maximum probability
Combination of (SVM, ANN, NB, C4.5, and RF)	61%	61%	69%
Combination of (ANN, NB, C4.5, and RF)	61%	61%	61%
Combination of (SVM, C4.5, and ANN)	76%	75%	69%
Combination of (SVM, C4.5, and NB)	63%	73%	69%
Combination of (NB and ANN)	70%	68%	68%
Single classifier (ANN)	73%		

TABLE XXII. CONFUSION MATRIX ON MAJORITY VOTING WITH 40% TEST SET

Number of instances = 632		Predicted		
		Class: No rain	Class: Rain	Total
Actual	Class: No rain	39	133	172
	Class: Rain	22	438	460
	Total	61	571	632

VI. Conclusions

Rainfall forecasting is a process to predict potential rainy locations by considering numerous factors like humidity, wind speed, level of water, and temperature. The common methods employed in rainfall forecasting are supervised machine learning methods in which predefined example data are first trained and then followed by the prediction using the testing data. The key challenge of these methods is in identifying suitable mechanisms, the sensitivity of the objective functions, and the dependency on treating features. These differences lead to inconsistent performances, making the selection process of a suitable method for rainfall prediction a task that is challenging. This paper proposes an improved method to develop long-term (i.e., monthly) and short-term (i.e., daily) weather forecasting models for rainfall predictions using the ensemble technique. Therefore, this paper proposes an improved method to develop long-term (i.e., monthly) and short-term (i.e., daily) ensemble weather forecasting models for rainfall predictions by using three linear algebraic combiners (i.e., majority voting, average probability and maximum probability) for combining five rainfall prediction models (i.e., the Naïve Bayes, C4.5, neural network, support vector machine, and random forest). By leveraging daily meteorological data in Selangor, Malaysia, over a period of 6 years (from 2010 to 2015), 1581 instances were obtained and organized into two classes. The first is the 'active rainfall' class containing 428 instances, while the remaining instances are grouped as 'no rainfall'. We have experimented a group of base algorithm models including the NB, SVM, ANN, RF and C4.5. From the analysis, all five ML techniques that were mentioned are shown to perform very well, although the ANN technique in particular is generally found to perform the best, while the NB technique is relatively the weakest. The study further explored the ensemble's potential for further upper-bound improvements in the rainfall prediction model. The ensemble experiment analysis revealed that the rainfall prediction is indeed enhanced using the ensemble method. Particularly, the ensemble method based on majority voting is shown to provide better predictive performance with high precision, recall, and F-Measures compared to other experimented algebraic combiners. Overall, the combiners have been demonstrated to be superior to single classification methods. Such results complement previous findings on ML methods in rainfall prediction and hence, our recommendation is to use ensemble ML algorithms as an effective approach for the above. It is hoped that the outcomes of this study may help to find suitable machine learning techniques that improve the performance of rainfall forecasting predictions.

Acknowledgment

The authors would like to extend their gratitude and thank Universiti Kebangsaan Malaysia (UKM) and Ministry of Education Malaysia (MOE) under the Research University Grant (project code: GUP-2019-060 and FRGS/1/2018/ICT02/UKM/02/6) for funding and supporting this research.

References

[1] C. W. J. Granger, and P. Newbold, Forecasting Economic Time Series. Cambridge, MA: Academic Press, 2014.

[2] A. S. Weigend, Time Series Prediction: Forecasting the Future and Understanding the Past, 1st ed., Milton Park, Abingdon: Routledge, 2018.

[3] G. Nan, S. Zhou, J. Kou, and M. Li, "Heuristic bivariate forecasting model of multi-attribute fuzzy time series based on fuzzy clustering," Int. J. Inf. Tech. Decis., vol. 11, pp. 167–195, January 2012.

[4] M. E. Mann, and P. H. Gleick, "Climate change and California drought in the 21st century, Proc. Natl. Acad. Sci., vol. 112, pp. 3858–3859. March 2015.

[5] M. J. Lee, I. Park, J. S. Won, and S. Lee, "Landslide hazard mapping considering rainfall probability in Inje, Korea, Geomatics, Nat. Hazards Risk, vol. 7, pp. 424–446, January 2016.

[6] C. C. Stephan, N. P. Klingaman, P. L. Vidale, A. G. Turner, M. E. Demory, and L. Guo, "A comprehensive analysis of coherent rainfall patterns in China and potential drivers. Part I: Interannual Variability, Clim. Dyn., vol. 50, pp. 4405–4424, June 2017.

[7] N. A. B. Klutse, B. J. Abiodun, B. C. Hewitson, W. J. Gutowski and M. A. Tadross, "Evaluation of two GCMs in simulating rainfall inter-annual variability over Southern Africa," Theor. Appl. Climatol., vol. 123, pp. 415–436, February 2016.

[8] K. Sittichok, A. G. Djibo, O. Seidou, H. M. Saley, H. Karambiri, and J. Paturel, "Statistical seasonal rainfall and streamflow forecasting for the Sirba watershed, West Africa, using sea-surface temperatures," Hydrol. Sci. J., vol. 61, pp. 805–815, April 2016.

[9] J. Wu, J. Long, and M. Liu, "Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm," Neurocomputing, vol. 148, pp. 136–142, January 2015.

[10] J. Sulaiman, and S. H. Wahab, "Heavy rainfall forecasting model using artificial neural network for flood prone area," in Proc. of IT Convergence and Security 2017, H. C. Kim, and K. J. Kim, Eds. Singapore: Springer, 2018, pp. 68–76.

[11] B. T. Pham, D. T. Bui, M. B. Dholakia, I. Prakash, and H. V. Pham, "A comparative study of least square support vector machines and multiclass alternating decision trees for spatial prediction of rainfall-induced landslides in a tropical cyclones area," Geotech. Geol. Eng., vol. 34, pp. 1807–1824, December 2016.

[12] S. Cramer, M. Kampouridis, A. A. Freitas, and A. K. Alexandridis, "An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives," Expert Syst. Appl, vol. 85, pp. 169–181, November 2017.

[13] M. Kim, Y. Kim, H. Kim, W. Piao, and C. Kim, "Evaluation of the k-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant," Front. Env. Sci. Eng, vol. 10, pp. 299–310, April 2016.

[14] S. Zainudin, D. S. Jasim,, and A. Abu Bakar, "Comparative analysis of data mining techniques for Malaysian rainfall prediction," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 6, pp. 1148–1153, December 2016.

[15] N. S., Sani, I. I. S. Shamsuddin, S. Sahran, A. H. A. Rahman, and E. N. Muzaffar, "Redefining selection of features and classification algorithms for room occupancy detection", Int. J. Adv. Sci. Eng. Inf. Technol., vol. 8, pp. 1486–1493, October 2018.

[16] P. S. Maya Gopal, "Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms," Appl. Artif. Intell., vol. 33, 621–642, June 2019.

[17] J. D. Holliday, N. Sani, and P. Willett, "Calculation of substructural analysis weights using a genetic algorithm," J. Chem. Inf. Model., vol. 55, pp. 214–221, February 2015.

[18] J. D. Holliday, N. Sani, and P. Willett, "Ligand-based virtual screening using a genetic algorithm with data fusion," MATCH Commun. Math. Comput. Chem., vol. 80, pp. 623–638, June 2018.

[19] N. S. Sani, M. A. Rahman, A. A. Bakar, S. Sahran, and H. M. Sarim. 2018. Machine learning approach for bottom 40 percent households (B40) poverty classification. Int. J. Adv. Sci. Eng. Inf. Technol., vol. 8, pp. 1698–1705, September 2018.

[20] K. K. Htike, and O. O. Khalifa, "Rainfall forecasting models using focused time-delay neural networks," in IEEE International Conference Computer and Communication Engineering (ICCCE), pp. 52–57, May 2010.

[21] R. V. Ramana, B. Krishna, S. Kumar, and N. Pandey, "Monthly rainfall prediction using wavelet neural network analysis," Water Resour. Manag., vol. 27, pp. 3697–3711, August 2013.

[22] M. Devak, C. T. Dhanya, and A. K. Gosain, "Dynamic coupling of support vector machine and K-nearest neighbour for downscaling daily rainfall," J. Hydrol., vol. 525, pp. 286–301, June 2015.

[23] D. Gupta, and U. Ghose, "A comparative study of classification algorithms for forecasting rainfall, in 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), pp. 1–6, September 2015.

[24] S. Shabudin, N. A. Sani, K. A. Z. Ariffin, and M. Aliff, "Feature selection for phishing website classification," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 11, pp. 587–595, January 2020.

[25] A. Abu, R. Hamdan, and N. S. Sani, "Ensemble learning for multidimensional poverty classification," Sains Malays., vol. 49, pp. 447–459, 2020.

[26] Chauhan, H., and A. Chauhan, "Implementation of decision tree algorithm c4.5," Int. J. Sci. Res., vol. 3, pp. 1–3, October 2013.

[27] M. Mayo, and E. Frank, "Improving naive bayes for regression with optimized artificial surrogate data," Appl. Artif. Intell., vol. 34, pp. 484–514, May 2020.

[28] Y. Wang, and L. Najjar, "Factor neural network theory and its applications," Int. J. Inf. Tech. Decis., vol. 14, pp. 239–251, March 2015.

[29] Y. Ren, L. Zhang, and P. N. Suganthan, "Ensemble classification and regression-recent developments, applications and future directions," IEEE Comput. Intell. Mag., vol. 11, pp. 41–53, January 2016.

[30] L. K. Hansen, and P. Salamon, "Neural network ensembles," IEEE Trans. Pattern. Anal. Mach. Intell., vol. 12, pp. 993–1001, October 1990.

[31] Y. Freund, and R. E. Schapire, "Experiments with a new boosting algorithm," in Proceedings of the Thirteenth International Conference on International Conference on Machine Learning (ICML), vol. 96, pp. 148–156, July 1996.

[32] A. J. Sharkey, N. E. Sharkey, and S. S. Cross. "Adapting an ensemble approach for the diagnosis of breast cancer, in International Conference on Artificial Neural Networks, pp. 281–286, September 1998.

[33] A. S. M. Salih, and A. Abraham, "Novel ensemble decision support and health care monitoring system," Int. J. Netw. Innov. Comput., vol. 2, pp. 041–051, 2014.

[34] Z. Ahmad, and J. Zhang, "A comparison of different methods for combining multiple neural networks models," in IEEE Proceedings of the 2002 International Joint Conference on Neural Networks, IJCNN'02 (Cat. No. 02CH37290), vol. 1, pp. 828–833, May 2002.

[35] J. Joseph, and T. K. Ratheesh, "Rainfall prediction using data mining techniques," Int. J. Comput. Appl., vol. 83, pp. 11–15, January 2013.

[36] S. S. Monira, Z. M. Faisal and H. Hirose, "Comparison of artificially intelligent methods in short term rainfall forecast," in 13th International Conference Computer and Information Technology (ICCIT), pp. 39–44, December 2010.

[37] A. Onan, S. Korukoğlu, and H. Bulut, "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification," Expert Syst. Appl., vol. 62, pp. 1-16, November 2016.

Presenting and Evaluating Scaled Extreme Programming Process Model

Muhammad Ibrahim[1], Shabib Aftab[2], Munir Ahmad[3], Ahmed Iqbal[4], Bilal Shoaib Khan[5]
Muhammad Iqbal[6], Baha Najim Salman Ihnaini[7], Nouh Sabri Elmitwally[8]
Department of Computer Science, Virtual University of Pakistan, Lahore, Pakistan[1, 2, 4]
School of Computer Science, National College of Business Administration & Economics, Lahore, Pakistan[2, 3, 6]
Department of Computer Science, Minhaj University Lahore, Lahore, Pakistan[5]
Department of Computer Science, College of Science and Technology, Wenzhou Kean University, China[7]
Department of Computer Science, College of Computer and Information Sciences, Jouf University, KSA[8]
Department of Computer Science, Faculty of Computers and Artificial Intelligence, Cairo University, Egypt[8]

***Abstract*—Extreme programming (XP) is one of the widely used software process model for the development of small scale projects from agile family. XP is widely accepted by software industry due to various features it provides such as: handling frequent changing requirements, customer satisfaction, rapid feedback, iterative structure, team collaboration, and small releases. On the other hand, XP also holds some drawbacks, including: less documentation, less focus on design, and poor architecture. Due to all of these limitations, XP is only suitable for small scale projects and doesn't work well for medium and large scale projects. To resolve this issue many researchers have proposed its customized versions, particularly for medium and large scale projects. The real issue arises when XP is selected for the development of small scale and low risk project but gradually due to requirement change, the scope of the project changes from small scale to medium or large scale project. At that stage its structure and practices which works well for small project cannot handle the extended scope. To resolve this issue, this paper contributes by proposing a scaled version of XP process model called SXP. The proposed model can effectively handle such situation and can be used for small as well as for medium and large scale project with same efficiency. Furthermore, this paper also evaluates the proposed model empirically in order to reflect its effectiveness and efficiency. A small scale client oriented project is developed by using proposed SXP and empirical results are collected. For an effective evaluation, the collected results are compared with a published case study of XP process model. It is reflected by detailed empirical analysis that the proposed SXP performed well as compared to traditional XP.**

Keywords*—Extreme Programming Process Model; XP; modified XP; scaled XP; customized XP; empirical comparison; empirical analysis

I. Introduction

Agile process models replaced the conventional and traditional software development methodologies due to effective features which were not available in conventional models [34]. These features include: emphasis on customer satisfaction, team collaboration and managing changing requirements [20],[45],[50]. Agile models follow an iterative and incremental way of development which delivers a high quality software [2-3], [32],[46]. Agile process models are backed by Agile Manifesto which is considered a parent document of agile family. This document explains the foundations of agile software development in the form of 12 basic principles and practices. These basic principles are about frequent team communication, customer satisfaction, managing frequent changing requirements even at later stages of development and early delivery of partial working software [1],[28],[31],[33],[35],[47],[48]. Many agile process models are used by the software industry now a days such as: Extreme Programming (XP), Scrum, Feature Driven Development (FDD), and Dynamic System Development Method (DSDM) [3],[5],[11],[37]. Extreme programming (XP) is one of popular agile process models for the development of small scale projects as well as widely used by the software industry [5],[41],[42],[43],[51]. XP is a light weight approach for software development, designed and developed by Kent Beck in 2000 [6]. It develops a qualitative software in limited time and lower cost by using some of the best engineering practices, principles and values in a disciplined way. The XP development life cycle (Fig. 1) has six phases: "Exploration phase, Planning phase, Iteration to release phase, Productionizing phase, Maintenance phase and Death phase" [7],[12],[13],[38]. Exploration phase deals with the requirement gathering and it is also responsible for the selection of particular architecture for development. Project planning phase deals with the overall planning, including: number of iterations, no of requirements to be implemented in each iteration, cost and time etc. Iteration to release phase deals with the development of a workable software, this phase may consists of one or more iterations. Productionizing phase deals with the testing of developed module. Maintenance phase deals with the addition of any new functionally (if required) by keeping the old ones intact and finally death phase deals with the completion of software as per client's requirement and ends with the release of software product. All of these phases are backed by twelve best practices of software engineering, including: "planning game, small releases, metaphor, simple design, continuous testing, refactoring, pair programming, collective ownership, continuous integration, 40-hour work per week, on-site customer and particular coding standards [8],[9],[10]. The structure of XP process model along with the umbrella of these 12 practices is best suited for small scale project and also can handle frequently changing requirements very

effectively [4],[39],[40],[44]. However the structure of XP cannot handle medium, complex and large scale projects [13],[36],[49],[52]. To handle this issue, many researchers have introduced improved versions of XP process models. However real issue arises when XP is selected for the development of small scale project and gradually the scope of the project extends to medium or large scale project due to clients requirements. To resolve this issue, this paper presents Scaled Extreme Process (SXP) Model. The proposed model can be used as an effective alternative of XP which can handle small as well as medium and large scale projects. Moreover, in the situation of sudden change in requirements and extension of the scope of small scale project to medium or large scale project, SXP can be effective as well. This research also evaluates the proposed SXP with an empirical case study in which a real time client oriented project is developed and results are compared with another published case study where XP is used for the development of client oriented small project. Comparative analysis reflects the effectiveness of proposed SXP process model.

This paper is further organized in the following sections. Section II highlights and discusses some of the related studies. Section III elaborates the problem definition. Section IV presents the proposed SXP process model. Section V empirically evaluates the proposed SXP. Section VI presents the critical analysis on results. Section VII finally concludes the paper.

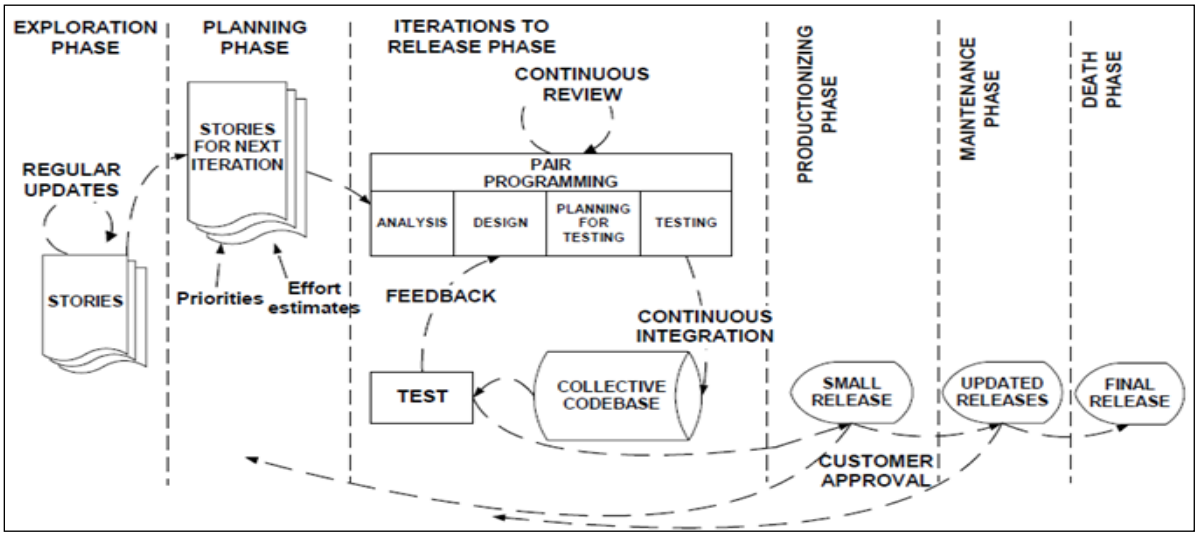


Fig. 1. XP Process Model.

II. Related Work

Many researchers have proposed modified versions of XP process models to reduce its limitations, some of the related studies are discussed here. Researchers in [14], customized the conventional XP process model in order to resolve: design, documentation and quality related issues. The proposed customized model performs the activities regarding non-functional requirements in a separate iteration. The proposed version of XP has some drawbacks including the need of extra staff members to refine each deliverable in each phase. Moreover, this process model is not good for that project which has higher interdependencies among subsystems. This research also lacks the empirical validation of proposed model. In [15], the researchers, proposed an extended iterative maintenance life cycle using XP practices for software maintenance. This approach uses RC (Request for Change) stories and old software as an input and performs all the phases & produce upgraded software. This study is validated through academic projects. However, academic projects are less complex than real-time oriented projects. In [16], the researchers proposed a hybrid model named DXPRUM by combining three agile process models DSDM, XP, and Scrum. The DXPRUM is proposed in order to achieve various features in one framework including: business solution, project management, agile team management, and core engineering practices. The proposed model is validated through the empirical case study in [17]. The DXPRUM process model performed much better as compared to DSDM. Researchers in [18], adopted XP for the development of large scale distributed project and introduced some new practices such as: stand-up meetings, code control, visual indicators, adaptive planning, XP project management, and code gallery. In [19], the authors proposed a new solution for development. It is a combination of seven principles of SOA and XP practices. The proposed solution did not resolve the issue of SOA complexities and did not sustain the agility of XP. In addition, the proposed solution has lacked empirical proof. In [20], the authors used Analytical Hierarchy (AHP) for the CRC cards prioritization process. The APH helps the developer to identify the most significant classes for simple designs. However, the proposed model is not evaluated on real time test cases. In [21], the author studied 40 different teams that use extreme programming for the development of small scale projects. This study provided comprehensive factors and practices which provide positive effects on team performance including: release planning, planning game, on-site customer, small releases, and stand-up meeting. On the other hand the researchers have also highlighted that unit testing, acceptance-testing, test-first design (using TTD), pair programming, and refactoring impacts negatively. In [22], the authors proposed an integrated XP process model. This model has the best engineering practices & management practices of XP, Scrum

and DSDM process model. They suggested a new role named "Technical Writer", who writes effective documentation that enhances understandability and future maintenance. In [23], the authors customized the XP process model for medium and large scale projects. This proposed solution is appropriate for parallel and incremental project development. Extended-extreme programming is Omni-direction in nature and it has five phases including the risk management phase. However, this study did not provide statistical proof regarding large scale project about parallel development. In [24], the author proposed an optimization model that assists in the activity of release planning in XP. The proposed solution essentially supports the development team of XP and the client in the release planning phase. In this model, the release plan is developed based on stories and their relations along with the priorities. However, it consumes a lot of time for data collection which ultimately loses the agility.

In [25], the authors presented a controlled empirical case study of XP and Waterfall methodology. Same project was developed multiple times over five years. The purpose of this research work is to evaluate the efficiency of the XP and Waterfall process models. This research shows unexpected same results of both process models. However, this research has a lack of diversity of data source and data characteristics. In [26], the researchers proposed a hybrid process model named eXRUP for small to medium scale projects by integrating XP and RUP. The proposed solution has been validated through a controlled case study. However, the proposed eXRUP has minimal interaction of programmers with customers and needs higher management. In [27], the authors proposed tailored extreme programming (TXP), a simple version of the XP process model. The author removed unnecessary practices and phases of the XP model to modify it for small teams and small projects with predefined requirements. In [29], the authors identified the need of software process improvement (SPI) in small firms. These small firms face the same SE challenges as large software industries face about SPI. This research develops an SPI structure for small firms by using XP process model. In [30], the researcher introduced a modified XP for medium and large scale projects with large team size. The proposed solution extends the capability of the conventional XP by resolving the design, and documentation related issues. A new phase named "Analysis and risk management" is also introduced to handle failure risks. The new XP model is validated through two case studies on two independent software houses.

III. Problem Definition

XP process model was designed for small teams to develop small scale projects having limited scope. In XP, the collection of good engineering practices and simple SDLC steps help to produce high quality software product within scheduled time but with limited scope (small project). Many researchers have explored the capabilities of XP and customized its practices for various projects types (such as medium and large) and nature (simple and complex) [30]. XP is ideal for small scale project however issue arises when the requirements of client constantly changes with the gradual passage of time which increases the scope of project from small to medium and large scale projects. In such cases, the features of XP like simple design, less documentation and limited testing and absence of proper change management activities can create hurdles to manage the quality as well as delivery of the product within specified time. To handle such issue, this research presents Scaled XP process model which works well for small scale to medium and large scale projects. Proposed SXP can tackle the issue of change management in such cases where project starts with limited scope but gradually extends to medium and large scale projects due to client's frequently changing and increasing requirements.

IV. Proposed Solution

This research proposes a customized XP process model called Scaled Extreme programming (SXP) for small to medium and large scale projects. The proposed solution is equally suitable for small, medium and large scale software projects, unlike the conventional XP process model. The SXP personalized the current practices of XP model to eliminate its limitations without effecting its agility. These limitations are eliminated by adding new phases and practices in SXP. Some effective activities which are included in the proposed model, include: managing the Risk Register, addition of UML artifacts, Effective Testing Mechanism, Formal Refinement Techniques and a formal procedure of Requirement Change Management (RCM). The RCM provides management support to the development team and customers to produce software products in a controlled & monitored environment. The workflow of SXP consists of seven phases as shown in Fig. 2. These phases are named as: Start Phase, Planning & Analysis Phase, Design Phase, Development phase, Acceptance Phase, Refinement Phase, and Release Phase.

A. Start Phase

The first phase of the SXP model is similar to the first phase of XP Model. In this phase, requirements are gathered from the clients by writing user stories. Writing story card is a very effective XP practice. User stories provide a high-level summary of the requirements for the desired system, and these are used as a primary input into estimating and scheduling. However, these user stories do not contain any technical detail of the desired software. In addition, Non Functional Requirements (NFR) are also explored with customer by keeping in view the Functional Requirements.

The extraction of NFR is also vital to the success of the project as these are extracted in order to get rid of undesirable results like unsatisfied client, as well as schedule & budget overruns, etc.

B. Planning and Analysis Phase

This phase consists of very important activities for the initiation of project and initiates with the input of detailed requirements which are further explored to estimate the risk, time, cost, budget and effort. Essential decisions regarding planning are made & documented including: Iteration plan, team size, estimation of cost, budget & effort. Activities of RCM are assigned to a team. Identification of the potential risk, Monitoring the risk and perform any actions required to mitigate the risk are included in the activities of RCM. Risk registers are used to document the complete actions during the process of risk management.

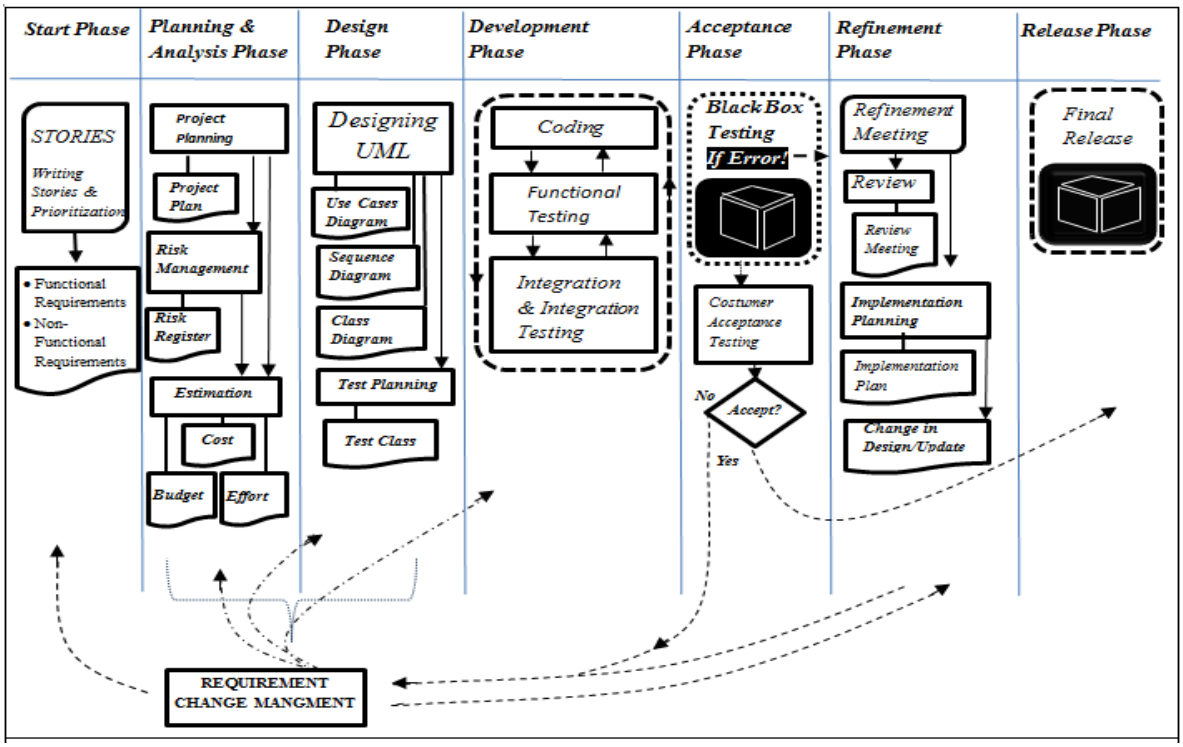


Fig. 2. Scaled Extreme Programming.

C. Design Phase

The design phase of SXP initiates with the development of UML diagrams including: Use Case, Sequence, and Class. These diagrams aim to help the developers to understand the functionalities of system in an easy way. These diagrams also help in change management activity. Moreover, by keeping in view the developed UML diagrams, test planning is completed. During test planning, test classes are written that verifies whether the certain pieces of code & classes are properly working or not.

D. Development Phase

In this phase, designed artifacts are transformed into working software or modules. The UML artifacts and test classes are the input of this phase. The design artifacts help developers to code in object oriented languages. The basic activities performed in this phase include: Coding, Functional Testing, and Integration & Integration Testing. After completion of each component, functional testing is performed to check and analyze the working of that developed component. If the developed component is working fine then it is integrated with previously developed component and then integration testing is performed in order to check whether the integrated components are working fine or not.

E. Acceptance Phase

It is a short phase in which testing is done by the tester in the presence of customer. The tester is a member of the development team who is assigned the task of testing the product externally. Black-Box testing is performed to examine the functionality and features of the system to meet client requirements. If the tester found any error during the test then this phase is aborted and refinement phase is initiated. However, if the tester passes the software then, the product is ready for the acceptance testing which is performed by the customer. This activity is essential and crucial in order to satisfy the customer. Moreover in this phase, the feedback is collected on the software and if the new request or dissatisfaction is reflected by the customer then it will be further handled and catered through RCM.

F. Refinement Phase

The phase initiates if the issues are found in Black-Box testing. The refinement starts with a formal meeting in which a detail review is performed to check the stories, developed artifacts, test classes, and codes in order to identify the issues. At the end of review meeting, identified issues are documented and resolved through an implementation plan. RCM takes necessary action against the refinement decision. In addition, some important documents are also updated by RCM like, Risk register, Change request register and design.

G. Release Phase

This is the last phase, in this stage software is ready to release or deliver to the client. The team moves to this stage when all user stories are implemented, and the customer is satisfied with the software. In addition, training, and documentation are provided to the client after deployment.

H. Role and Responsibilities of RCM

Requirement Change Management (Fig. 3) is a supporting framework of SXP. It is introduced to cater the change requests properly without dropping the team productivity. The RCM can consist of one or more team members.

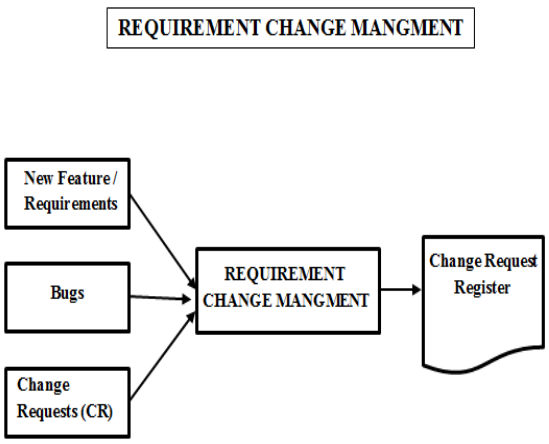


Fig. 3. Requirement Change Management.

Some key duties of RCM are as follows.

- It directly interacts with the customer and the team.
- It collects requirements from the customers in case of change request or requirement for new feature.
- It provides support to the customer and the development team in the entire project cycle.
- It records in change request register if 1) client wants new features 2) clients want to change existed features 2) client is not satisfied by acceptance testing.
- If errors are found in the acceptance phase by tester then it records the changes in change request register.

V. Empirical Evaluation of Proposed SXP

This research proposed a modified version of XP process model named SXP in order to resolve the issues of conventional XP. To analyze the performance of proposed model, an empirical case study is conducted in which a real time client oriented project is developed by using the proposed SXP. The case study was conducted in a software house situated in Karachi, capital of Sindh Province, Pakistan. All of the team members, who participated in the case study have higher degrees in computer science discipline (BS or MS) along with software development experience of at least one year. During the development, data of various software metrics are collected for empirical analysis. For effective evaluation, the results of SXP case study are compared with another case study presented in [26], where conventional XP process model was used to develop a small scale project. Characteristics of both the case studies are reflected in Table I.

XP case study was conducted by the students of computer science programs of BS and MS level with no or little knowledge of agile development. However a training session of 10 days was organized before the development. The reason of choosing the XP case study for comparison with the case study of proposed SXP is to critically analyze the performance of SXP with empirical data, so that any gap or deficiency can further be identified for further improvement. This study compares the proposed SXP with XP process by providing detailed empirical results of all of the iterations as per the guidelines provided by [10]. Both of the case studies (SXP and XP) are empirically compared in Table II. First column of the table shows the serial no whereas second column reflects the particular metrics for which the data was collected during the development for comparison. These metrics include: development time, cost, productivity etc. All of these measures are considered as an effective way to analyze the quality of any process model. The columns (Release 1 to Release 4) shows release wise measures/values of the attributes of column 2 and finally the last columns reflect the aggregated/average values of all the releases.

TABLE I. Selected Case Studies

Characteristics	**SXP**	**XP**
Product Type	Social Media platform	Real Estate Management
Size	Small	Small
Iterations	4	3
Programming Approach	Object Oriented	-
Language	JavaScript	PHP
Documentation	Ms Office & JS Doc	MS Office
Testing	Desktop & Mobile browser testing	-
Project Complexity Type	Average	Average
Team Size	4 Members	3 Members
Development Environment	Visual Studio, Ionic SDK & SQL	Macromedia Dream Viewer and Net Beans
Other Tools	MS Visio	MS Visio

TABLE II. EVALUATION OF XP AND SXP PROCESS MODEL

No	Parameters	Release 1		Release 2		Release 3		Release 4	Total	
		XP	SXP	XP	SXP	XP	SXP	SXP	XP	SXP
1	Completion time (Week)	2	1	1	1	1	1	1	4	4
2	Number of Modules	2	8	1	5	1	7	3	4	23
3	No of User Stories	17	1	13	2	11	2	2	41	7
4	Budget Effort in (h)	240	128	120	128	120	128	128	480	512
5	Actual Effort in (h)	210	120	90	120	90	96	120	390	456
6	No. of user Interface	2	2	1	3	1	2	2	4	9
7	No of class designed	46	10	34	8	30	4	4	110	26
8	Total Line of Code	4500	5300	3200	3900	3300	2900	3200	11000	14700
9	KLOC	4.5	5.3	3.2	3.9	3.3	2.9	3.2	11	14.7
10	No of integration	20	14	12	12	12	18	6	44	50
11	Post Release Defects	2	3	2	1	4	2	1	8	7
12	Post Release Defects per KLOC	0.44	0.56	0.62	0.25	1.212	0.68	0.31	0.727	0.47
13	Productivity= LOC / Actual effort	21.4	44.1	35.6	32.5	36.7	30.2	26.6	28.2	32.2
14	No of prerelease change requests	3	2	2	3	2	1	3	7	9
15	Total change requests per KLOC	0.66	0.37	0.62	0.76	0.60	0.34	0.93	0.636	0.61
16	Time to implement changes in hour	4	2	3	2	1	1	5	8	10

VI. CRITICAL ANALYSIS

Some significant differences are reflected in the performance of both models (XP and SXP) in Table II. Complexity level of both the projects is different as reflected by: KLOC, No of code integration, No of modules, and No of interfaces. “KLOC” and “Actual Effort” both are considered as important software metrics to analyze the performance of software process models. KLOC developed during SXP case study and XP case study is reflected in Fig. 4. Actual Effort (h) in both the case studies is reflect in Fig. 5. Release wise Actual Effort in both the case studies is also shown in Fig. 6. In XP project, 11 KLOC were produced with 390 hours of actual effort. However during the development of SXP project, 14.7 KLOC were produced with 456 hours of actual effort. It can be seen that the proposed SXP model slightly performed better in these metrics as 3.7 more KLOC were developed with 66 more hours of actual effort as compared to XP. However, it should also be noted that in XP project, there were 3 team members whereas in SXP, there are 4 team members.

During XP case study, 41 requirements were implemented however in SXP case only 7 requirements are implemented (Fig. 7). It should also be noted here that no of modules designed and developed during the implementation of 41 requirements of XP case study were only 4 as compared to 23 in SXP case study. Moreover, 4 interfaces were developed in XP case study as compared to 9 in SXP. Another important metric which should be discussed along with “No of implemented requirements” is the “No of code integrations”. There were 44 integrations in XP as compared to 50 in SXP. So it can be analyzed that only the no of requirements implemented in a project cannot reflect the performance of a model as client can write only one requirement which might have many modules, interfaces and backend integrations. The performance of SXP is better in all of these metrics as the team of SXP done more work as compared to the team of XP.

The “No of designed classes” is also an important software metric to analyze the performance of teams especially when this metric is analyzed along with developed KLOC. In XP case study, 110 classes were designed as compared to 26 in SXP (Fig. 8). As the development approach in SXP was object oriented so the very less no of classes with higher no of KLOC is justified. Object oriented principles used in SXP case study is also one of the reasons of good performance as it increases the re usability of code with an effective and efficient way.

The defects which are discovered by the client after the release is an important quality parameter which also reflects the customer satisfaction. The software application developed with XP reflected 8 defects as compared to 7 in SXP case study (Fig. 9). This metric raises the questions on the quality assurance activity and particularly the testing strategy of software process model. In SXP, efforts are made to produce the quality software even it performed better than XP (reflected from the implemented case study) but 7 defects after the release are not acceptable. However, this issue can be raised if the testing mechanism of the model is not implemented properly by the team.

Software productivity is a crucial metric to analyze the performance of any software process model. It reflects the effort of whole development team during the project. It shows the amount of effort, the team has put to complete the project within defined time. However in order to analyze the effectiveness and efficiency of the model, this single parameter is not enough, instead all of the software metrics

shown in Table II collectively reflect the performance of model. The team of XP reflected the productivity 28.2 as compared to 32.2 in SXP (Fig. 10, Fig. 11). The SXP model produced more lines of code in less time. If the Productivity is analyzed by keeping in view the complete list of parameters in each of the given iterations then it can be said that the proposed SXP performed well.

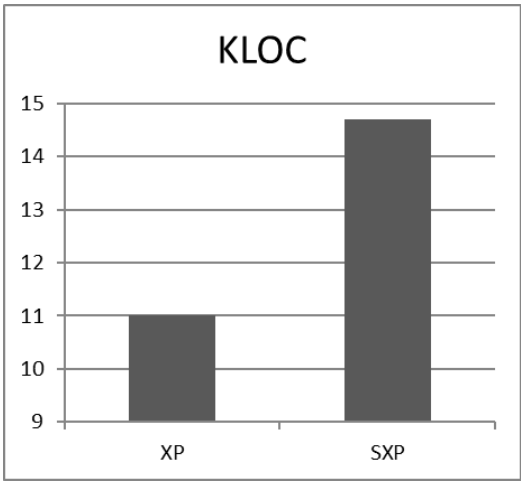


Fig. 4. KLOC.

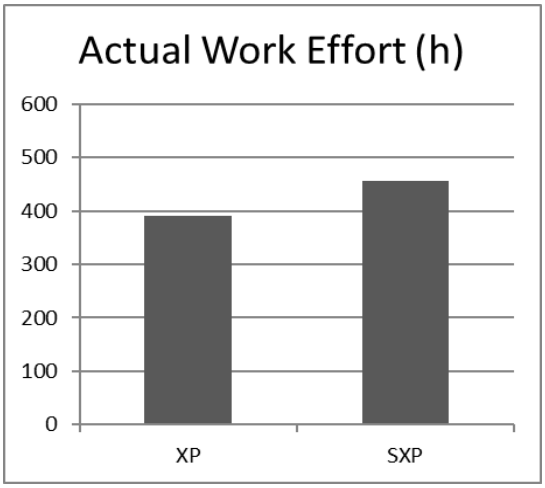


Fig. 5. Actual Work Effort.

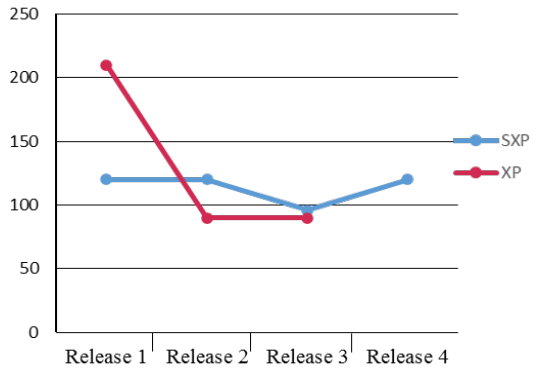


Fig. 6. Releasee Wise Actual Work Effort.

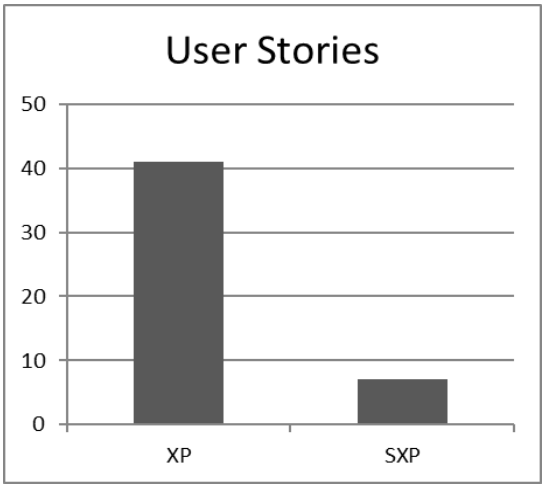


Fig. 7. User Stories.

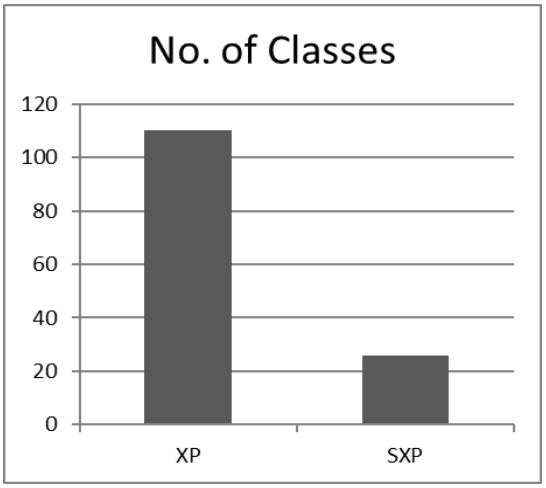


Fig. 8. No of Classes.

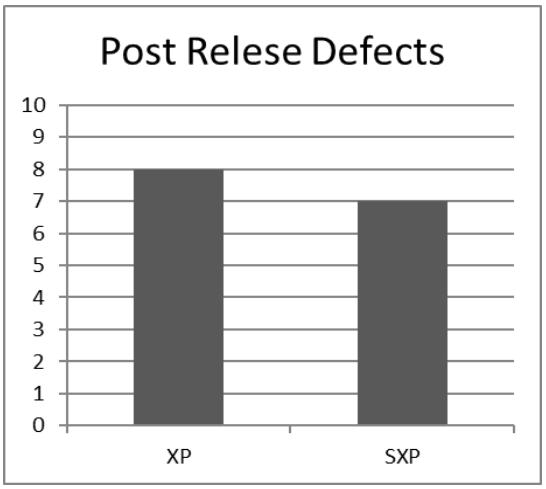


Fig. 9. Post Release Defects.

The projects implemented in both the case studies were same in nature (web based) but different in complexity level. The project implemented with SXP was complex as compared to XP project. Moreover development language, project size and no of team members were also different. Results of all the

quality parameters reflect the effectiveness of proposed SXP model however 7 defects after the release of complete software product raises the question on quality assurance activities. There might be other reasons of these defects including the human error of testing personnel or lack of quality test cases etc.

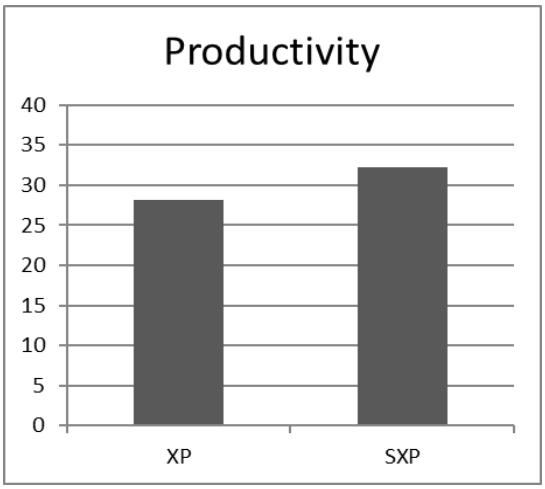


Fig. 10. Productivity.

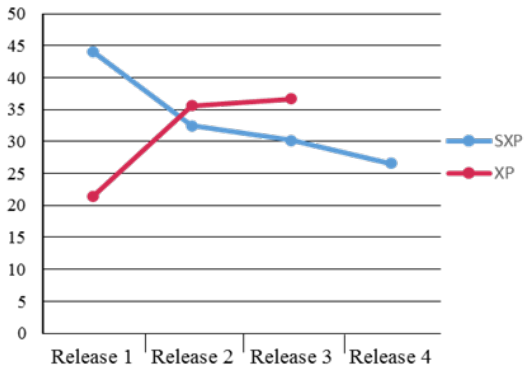


Fig. 11. Release wise Productivity.

VII. Conclusion

XP is considered as one of the widely used agile process model by the software industry for the development of small scale projects. Its practices include: quick response to changing requirements, customer satisfaction, rapid feedback, team collaboration, and small releases. However, besides the featured practices, XP has a major drawback as well and that is: its ability to handle only small projects. To resolve this issue many researchers have proposed its customized versions specifically to handle medium and large scale projects. However real problem arises when the conventional XP process is selected for a small scale and low risk project but with the gradual passage of time, the frequently changing requirements due to modern business change drags the scope of project from small scale to medium or even large scale. At this stage, some characteristics of conventional XP don't let its life cycle to handle medium or large projects. The characteristics include: poor architectural structure, lack of documentation, less focus on design and absence of proper change management procedure. This research has proposed a scaled version of XP process model which can handle such situations very effectively. Moreover, the proposed model can be equally effective for small, medium and large scale projects. In the proposed model, more focus is given on designing, testing and particularly on change management procedure. Due to these features, SXP can handle any extension in the scope of the project. An empirical evaluation is also performed in order to analyze the effectiveness of proposed SXP. For this purpose, a case study is conducted in which a real time client oriented project is developed. Empirical results of software quality metrics are collected during the development and then compared with another published case study in which XP was used for the development of a client oriented project. A detailed empirical analysis is performed and it is observed that the proposed SXP performed well almost in every important quality parameter. However to further evaluate the proposed model, medium or large complex project should be chosen for development.

References

[1] S. Ashraf and S. Aftab, "Latest Transformations in Scrum: A State of the Art Review," Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 7, 2017.

[2] L. Williams, "Agile software development methodologies and practices," in Advances in computers, vol. 80, pp. 1–44, 2010.

[3] C. R. Kavitha and S. M. Thomas, "Requirement gathering for small projects using agile methods," IJCA Spec. Issue Comput. Sci. Dimens. Perspect. NCCSE, 2011.

[4] J. Newkirk, "Introduction to agile processes and extreme programming," in Proceedings of the 24th International Conference on Software Engineering. ICSE 2002, 2002, pp. 695–696.

[5] A. Begel and N. Nagappan, "Usage and perceptions of agile software development in an industrial context: An exploratory study," in First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007), pp. 255–264, 2007.

[6] K. Beck, Extreme programming explained: embrace change. addison-wesley professional, 2000.

[7] F. Anwer, S. Aftab, S. S. M. Shah, and U. Waheed, "Comparative Analysis of Two Popular Agile Process Models: Extreme Programming and Scrum," Int. J. Comput. Sci. Telecommun., vol. 8, no. 2, pp. 1–7, 2017.

[8] M. N. Swamy, L. M. Rao, and M. P. KS, "Component Based Software Architecture Refinement and Refactoring Method into Extreme Programming," architecture, vol. 5, no. 12, 2016.

[9] S. Ashraf and S. Aftab, "Scrum with the Spices of Agile Family: A Systematic Mapping," Mod. Educ. Comput. Sci., vol. 9, no. 11, 2017.

[10] S. Aftab, Z. Nawaz, F. Anwer, M. Ahmad, A. Iqbal, A. A. Jan, and M. S. Bashir, "Using FDD for small project: An empirical case study," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 3, pp. 151–158, 2019.

[11] J. A. J. Builes, D. L. R. Bedoya, and J. W. B. Bedoya, "Metodología de desarrollo de software para plataformas educativas robóticas usando ROS-XP," Rev. Politécnica, vol. 15, no. 30, pp. 55–69, 2019.

[12] T. Saeed, S. S. Muhammad, M. A. Fahiem, S. Ahamd, M. T. Pervez, and A. B. Dogar, "Mapping Formal Methods to Extreme Programming (XP)–A Futuristic Approach," Int. J. Nat. Eng. Sci., vol. 8, no. 3, pp. 35–42, 2014.

[13] F. Anwer and S. Aftab, "SXP: Simplified Extreme Programing Process Model," Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 6, p. 25, 2017.

[14] M. R. J. Qureshi and J. S. Ikram, "Proposal of Enhanced Extreme Programming Model," Int. J. Inf. Eng. Electron. Bus., vol. 7, no. 1, p. 37, 2015.

[15] J. Choudhari and U. Suman, "Extended iterative maintenance life cycle using eXtreme programming," ACM SIGSOFT Softw. Eng. Notes, vol. 39, no. 1, pp. 1–12, 2014.

[16] M. Fahad, S. Qadri, S. S. Muhammad, and M. Husnain, "Software Quality Assurance of Medium Scale Projects by using DXPRUM Methodology," Int. J. Nat. Eng. Sci., vol. 8, no. 1, pp. 42–48, 2014.

[17] M. Fahad, S. Qadri, S. Ullah, M. Husnain, R. Qaiser, S. Ahmed, W. A. Qureshi, and S. S. Muhammad, "A Comparative Analysis of DXPRUM and DSDM," IJCSNS, vol. 17, no. 5, p. 259, 2017.

[18] E. Abdullah and E.-T. B. Abdelsatir, "Extreme programming applied in a large-scale distributed system," in 2013 International Conference On Computing, Electrical And Electronic Engineering (Icceee), 2013, pp. 442–446.

[19] F. Carvalho and L. G. Azevedo, "Service agile development using XP," in 2013 IEEE Seventh International Symposium on Service-Oriented System Engineering, pp. 254–259, 2013.

[20] S. Alshehri and L. Benedicenti, "Prioritizing CRC cards as a simple design tool in extreme programming," in 2013 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 2013, pp. 1–4.

[21] S. Wood, G. Michaelides, and C. Thomson, "Successful extreme programming: Fidelity to the methodology or good teamworking?," Inf. Softw. Technol., vol. 55, no. 4, pp. 660–672, 2013.

[22] S. Sultana, Y. H. Motla, S. Asghar, M. Jamal, and R. Azad, "A hybrid model by integrating agile practices for pakistani software industry," in 2014 International Conference on Electronics, Communications and Computers (CONIELECOMP), 2014, pp. 256–262.

[23] M. R. J. Qureshi, "Agile software development methodology for medium and large projects," IET Softw., vol. 6, no. 4, pp. 358–363, 2012.

[24] G. van Valkenhoef, T. Tervonen, B. de Brock, and D. Postmus, "Quantitative release planning in extreme programming," Inf. Softw. Technol., vol. 53, no. 11, pp. 1227–1235, 2011.

[25] F. Ji and T. Sedano, "Comparing extreme programming and Waterfall project results," in 2011 24th IEEE-CS Conference on Software Engineering Education and Training (CSEE&T), 2011, pp. 482–486.

[26] G. Rasool, A.Shabib, H.Shafiq, and S.Detlef, "eXRUP: A Hybrid Software Development Model for Small to Medium Scale Projects.", Journal of Software Engineering and Applications, vol. 6, pp. 446-457, 2013.

[27] F. Anwer, S. Aftab, and I. Ali, "Proposal of Tailored Extreme Programming Model for Small Projects," Int. J. Comput. Appl., vol. 171, no. 7, pp. 23–27, 2017.

[28] S. Alam, S. Nazir, S. Asim, and D. Amr, "Impact and Challenges of Requirement Engineering in Agile Methodologies: A Systematic Review," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 4, pp. 411–420, 2017.

[29] M. Y. Al-Tarawneh, M. S. Abdullah, and A. B. M. Ali, "A proposed methodology for establishing software process development improvement for small software development firms," Procedia Comput. Sci., vol. 3, pp. 893–897, 2011.

[30] M. R. J. Qureshi, "Estimation of the New Agile XP Process Model for Medium-Scale Projects Using Industrial Case Studies," Int. J. Mach. Learn. Comput., vol. 3, no. 5, pp. 393–395, 2013.

[31] S. Ashraf and S. Aftab, "Pragmatic Evaluation of IScrum & Scrum," Int. J. Mod. Educ. Comput. Sci., vol. 10, no. 1, pp. 24–35, 2018.

[32] F. Anwer, S. Aftab, U. Waheed, and S. S. Muhammad, "Agile software development models tdd, fdd, dsdm, and crystal methods: A survey." Int. J. Multidiscip. Sci. Eng., vol. 8, no. 2, pp. 1–10, 2017.

[33] S. Aftab, Z. Nawaz, M. Anwar, F. Anwer, M. S. Bashir, and M. Ahmad, "Comparative Analysis of FDD and SFDD," Int. J. Comput. Sci. Netw. Secur., vol. 18, no. 1, pp. 63–70, 2018.

[34] S. Ashraf, "IScrum: An Improved Scrum Process Model," Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 8, pp. 16–24, 2017.

[35] Z. Nawaz, S. Aftab, and F. Anwer, "Simplified FDD Process Model," Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 9, pp. 53–59, 2017.

[36] S. Aftab, Z. Nawaz, F. Anwer, M. S. Bashir, M. Ahmad, and M. Anwar, "Empirical evaluation of modified agile models," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 6, pp. 284–290, 2018.

[37] F. Anwer and S. Aftab, "Latest Customizations of XP: A Systematic Literature Review," Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 12, pp. 26–37, 2017.

[38] F. Anwer, S. Aftab, M. S. Bashir, Z. Nawaz, M. Anwar, and M. Ahmad, "Empirical Comparison of XP & SXP," Int. J. Comput. Sci. Netw. Secur., vol. 18, no. 3, pp. 161–167, 2018.

[39] M. R. J. Qureshi, "Empirical Evaluation of the Proposed eXSCRUM Model-Results of a Case Study," Int. J. Comput. Sci. Issues, vol. 8, no. 3, pp. 150–157, 2011.

[40] Z. Mushtaq and M. R. J. Qureshi, "Novel Hybrid Model: Integrating Scrum and XP," Int. J. Inf. Technol. Comput. Sci., vol. 4, no. 6, pp. 39–44, 2012.

[41] M. R. J. Qureshi, "An Evaluation of the Improved XP Software Development Process Model," Strategy, vol. 20, no. 2, pp. 79–82, 2008.

[42] S. Kazi, M. S. Bashir, M. M. Iqbal, Y. Saleem, M. R. J. Qureshi, and S. R. Bashir, "Requirement change management in agile offshore development (RCMAOD)," Sci. Int., vol. 26, no. 1, pp. 131–138, 2014.

[43] M. R. Jameel Qureshi, "Evaluating the Quality of Proposed Agile XScrum Model," Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 11, pp. 41–48, 2017.

[44] A. I. Khan, M. R. J. Qureshi, and U. A. Khan, "A Comprehensive Study of Commonly Practiced Heavy & Light Weight Software Methodologies," no. February, 2012.

[45] M. R. J. Qureshi and S. A. Hussain, "An Improved XP Software Development Process Model," SCIENCE INTERNATIONAL-LAHORE, vol. 20, no. 1, 2012.

[46] M. R. Jameel Qureshi and M. Kashif, "Adaptive Framework to Manage Multiple Teams Using Agile Methodologies," Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 1, pp. 52–59, 2017.

[47] R. J. Qureshi, M. O. Alassafi, and H. M. Shahzad, "Lean Agile Integration for the Development of Large Size Projects," Int. J. Mod. Educ. Comput. Sci., vol. 11, no. 5, pp. 24–33, 2019.

[48] R. J. Q. M and Z. Abass, "Long Term Learning of Agile Teams," Int. J. Softw. Eng. Appl., vol. 8, no. 6, pp. 01–18, 2017.

[49] M. R. J. Qureshi and A. Barnawi, "Kinect Based Electronic Assisting System to Facilitate People with Disabilities Using KXPRUM Agile Model," Life Sci. J., vol. 11, no. 10, pp. 56–62, 2014.

[50] M. Rizwan Jameel Qureshi, "Comparison of Agile Process Models to Conclude The Effectiveness for Industrial Software projects," Sci. Int., vol. 28(5), no. November-December, pp. 5119–5123, 2016.

[51] M. R. J. Qureshi and S. A. Hussain, "A reusable software component-based development process model," Adv. Eng. Softw., vol. 39, no. 2, pp. 88–94, 2008.

[52] S. U. Nisa and M. R. J. Qureshi, "Empirical Estimation of Hybrid Model: A Controlled Case Study," Int. J. Inf. Technol. Comput. Sci., vol. 4, no. 8, pp. 43–50, 2012.

Measuring Impact of Traffic Parameters in Adaptive Signal Control through Microscopic Simulation

Fatin 'Ayuni Bt Aminzal[1]
Faculty of Civil Engineering and Built Environment
Universiti Tun Hussein Onn Malaysia
Batu Pahat, Malaysia, 86400

Munzilah Binti Md Rohani[2]
Associate Professor
Faculty of Civil Engineering and Built Environment
Universiti Tun Hussein Onn Malaysia
Batu Pahat, Malaysia, 86400

***Abstract*—This paper aims to exploit the traffic parameters setting in adaptive traffic control. In this study, is known as Dynamic Timing Optimiser (DTO). DTO is an online algorithm, uses real-time optimisation in estimating cycle length according to fluctuations arrival flow registered from the detector. DTO cycle time estimation is also incorporated with preset parameters including saturation flow rate (s) and lost time (L). However, these traffic flow parameters commonly inputted as one deterministic value which adopted for the whole day. For example, presumed constant of saturation flow rate (s) do not accurately represent an actual oversaturated condition. The effects of employing inaccurate saturation flow rate (s) lead to the underestimation of cycle length. Therefore, a set of parameters value is applied and tested encompass of default value and adjusted value that implied a heaviest traffic condition through microscopic simulation. This resulted in outcomes of intersection performance in terms of intersection delay, travel time and throughput. According to simulation result, saturation flow rate (s) parameters show a great influence in cycle length optimisation compared to lost time (L) parameter. Employing a realistic saturation flow rate (s) while inputting parameters in DTO according to real traffic conditions contribute to a less intersection delay. In addition, the study revealed that a longer lost time (L) configured in the signal system, a longer cycle length generated by DTO algorithm. As predicted, high delay occurs during long cycle length yet benefited in allowing a higher throughput.**

Keywords*—Adaptive signal control; optimal cycle length; saturation flow rate; lost time; microsimulation

I. INTRODUCTION

Dynamic Timing Optimiser (DTO) cycle length model has been developed to produce an optimal cycle length in real-time traffic condition. DTO is an adaptive signal control system implemented in Malaysia mainly in Kuala Lumpur City Centre. During installation of DTO, two traffic parameters setting were identified in DTO cycle length model; saturation flow rate (s) and lost time (L).

Saturation flow rate (s) is defined as vehicle queue discharge at a more or less constant rate during effective green [1]. Saturation flow rate (s) can be measured from site by collecting similar vehicle headway after four of five vehicles depart from stop line, or about 10 seconds after the start of green period [2]. Calculation of saturation flow rate (s) can be measured by averaging saturation headway (h) in the field as summarised in Eq. 1 [3].

$$s = \frac{3{,}600}{h} \quad (1)$$

Where s = saturation flow rate, 3,600 = number of seconds per hour and h = saturation headway.

Generally, vehicle headway is varying according to site condition, including speed limit, gradient of the intersection and number of through lanes [4]. A reduced saturation flow rate resulted a longer cycle length [5]. For the cycle length estimation in fixed time control, Highway Capacity Manual has suggested a standard saturation flow rate is fixed to 1,900 pcu/hour/lane [6].

Another traffic parameters involved in optimising cycle length by DTO is a lost time (L) in regards to the total of signal phasing. The estimation of lost time (L) is crucial in determining optimal cycle length. Lost time estimation methods that has been distinguished by USA, Japan, Germany and Australia is equal to the sum of the intergreen times (I) [7]. An increasing of lost time may effects on the increasing of optimal cycle length, and as a consequences, the intersection delay increases [8].

The traffic parameters in DTO is structured to be configured as one deterministic value which adopted for the whole day without differentiate off-peak and highest peak condition. Engineers are typically inputted an average value to applied during all day. As explained on above, saturation flow rate (s) presumed constant do not accurately represent an actual oversaturated condition. In addittion, a small variation of lost time may lead to inaccurate calculation of cycle length [9]. Therefore, saturation flow rate need to be calibrated particularly when demand reached the capacity [10].

This study is intended to examine a scenario if the parameters can be adjusted according to the peak condition instead of using a default value. The scenarios are basically exploting the value of 1) saturation flow rate of 1,714 pcu/hr/lne and 2,250 pcu/hr/lane 2) lost time of 16s, 24s and 36s. The evaluation was conducted using vehicle flow obtained from detector, signal timing generated from DTO traffic engine and travel time collected from site survey. This research is expected to serve future reference on relationship of paramaters in cycle length model theory especially when adopted in real-time calculation of adaptive signal control.

This paper is organised as follows. "Background" introduces on the concept of optimal cycle length and DTO adaptive control. "Methodology" briefed on the study area along with the data collection, DTO parameters configuration and microsimulation modelling process. Findings and result are presented in "Result and Discussion". Finally, "Conclusion" conclude this research study with future works explained.

II. Background

Signalised intersection acquired a probabilistic calculation of cycle length. Since the beginning, there are a few cycle lengths models that are adopted in controlling the movement of vehicular traffic at intersection, for example Webster method and HCM 2000 model. Webster method estimates the green time of the phases is in the ratio of their flow ratio [11]. This can be shown in the Equation 2.

$$C_0 = \frac{1.5L+5}{1-Y} \quad (2)$$

Where C_0 = optimal minimum delay cycle length (s); L = total lost time within the cycle (s); and Y = sum of critical phase flow ratios.

In the Highway Capacity Manual 2000 document, estimation of cycle length is available in Appendix A, Chapter 10 [12].

$$C = \frac{L}{1-\frac{\min(CS,RS)}{RS}} \quad (3)$$

Where C= cycle length (s); L = total lost time (s/cycle); CS = sum of the critical phase traffic volumes (veh/hr); RS = reference sum flow rate (s x PHF x fa), (veh/hr); PHF = peak hour factor; s = saturation flow rate and fa = area type of adjustment factor [0.90 if central business district and 1.0 for otherwise].

Eventhough the above model have been widely used at isolated pre-timed signalised intersection, they have some limitations when implemented for cycle length estimation. For example, Webster method is not applicable when the traffic condition is near-saturated or saturated condition. When the critical flow ratio is equal or more than one, the model becomes infeasible where it generates unreasonbaly large cycle length [13]. Whereas the HCM 2000 model is targeting on the expected saturation in generating their cycle length which may contributing to the augmented value [14]. Webster method is efficient use at under saturated condition and HCM 2000 model is suitable at oversaturated condition [15]. Both model were developed to suit pre-timed signal operation where the traffic pattern is well distinguish between peak hour and off peak condition.

As time and technology emerged, an adaptive traffic control has been introduced to overcome the limitations of pre-timed control. Adaptive control offered an independent strategy in calculating their signal timings according to actual traffic demand. Sena Traffic Systems Sdn Bhd (STS), a local made company in Malaysia has took the challenge in developing own adaptive signal algorithm known as Dynamic Timing Optimiser (DTO). The purpose of DTO is targeted to minimize wasting green and delay time as well as to increase capacity of the intersection. DTO signal control requires detector to detect vehicle passage/presence to register presence of vehicles after the termination of green in estimating next cycle green time.

In the beginning of DTO signal timing process, the real-time vehicle detection at site is transmitted to the controller. The information received by the controller is subsequently directed to a server by using a wired or wireless communication. After that, DTO algorithm performed a traffic data processing by considering dependant variables such as peak hour factor and congestion index. DTO are then doing customisation of the timing according to the algorithm decision branches that has been developed according to traffic conditions.

DTO cycle length model has incorporated a traffic parameters similar with Webster and HCM model such as saturation flow rate (s) and lost time (L). This traffic parameters are configured offline during pre-installation and applied throughout the day. This traffic parameters are essential for DTO traffic engine in providing most accurate signal timing according to the situation on site. Fig. 1 demonstrates the summary of DTO optimisation process as explained above.

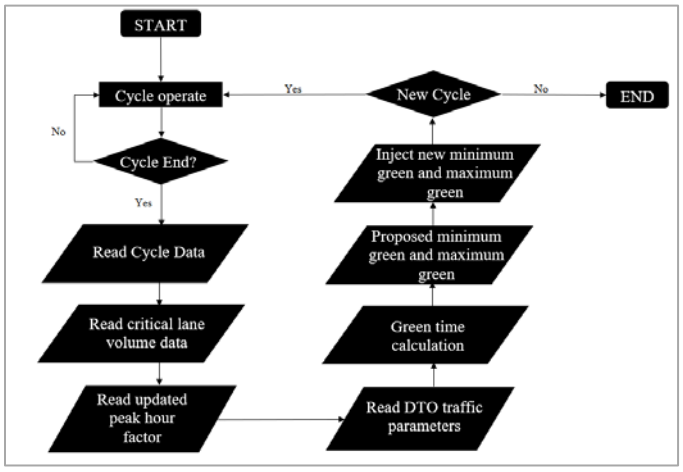


Fig. 1. Flowchart of DTO Optimisation Process.

III. Methodology

A. Description of Study Area

A signalised four-legged intersection has been selected as a case study to achieve the research aim. This intersection is on the four-lane dual carriageway on Jalan Jalil Perkasa 1, Kuala Lumpur. Most of the traffic traverse this sub-arterial road going towards and from KL city centre. Fig. 2 is a map showing the location of the study intersection and peak hour traffic condition.

B. Field Data Collection

Traffic volume, intersection geometry, queue length data and travel time data were obtained on typical weekday. A morning peak traffic condition were chosen as study duration as to simulate a high traffic condition. These data were used to calibrate the simulation models.

Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia; Ministry of Higher Education (MOHE), Malaysia.

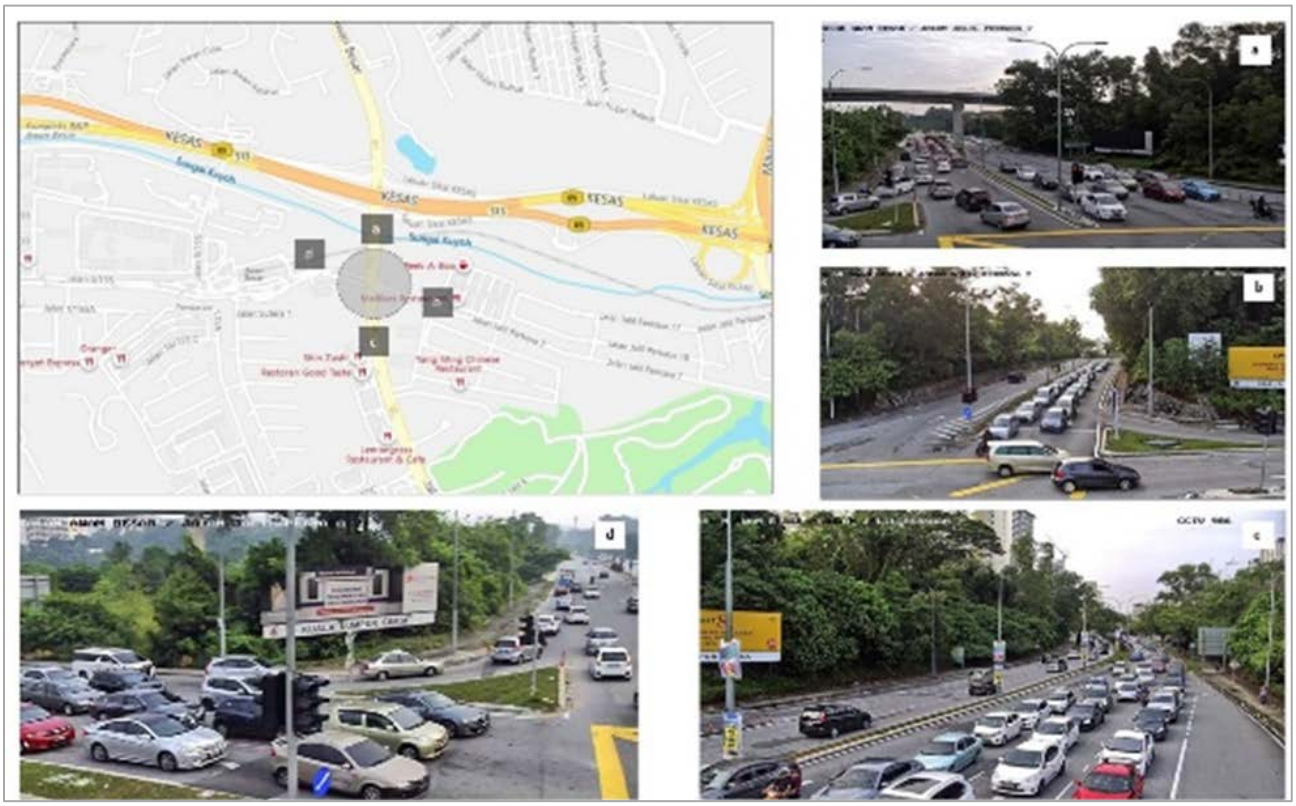


Fig. 2. Junction Location and Site Condition.

- Traffic Volume

Traffic volume from detector at stop line were obtained through STS control centre software. Highest traffic volume during morning peak between 7.00am to 10.00am has been identified. While turning movement volume and vehicle composition were collected during manual count survey. Table I tabulates the traffic volume on 8.00am to 9.00am. It is noted the traffic composition during morning peak hour dominated by car (98.2%), medium lorries (1.3%) and busses (0.5%).

- Intersection Geometry

The geometric data were extracted from Google Earth and validated by field measurement. The entry approach lane width and short lane length were identified. Fig. 3 show the existing layout of the intersection.

- Queue Length and Travel Time

The maximum queue was identified based on speed profile particularly at southbound approach. The selection of this direction is due to the high traffic volume from this approach. The vehicle is consider in a queue state when the speed is below than 10km/hr. While travel time survey was estimated using floating car technique in a test vehicle. This survey is conducted by traverse from start at end point of the defined segment from southbound approach. Travel time within segment was determined based on the average of seven runs of a floating car. Fig. 4 shows a typical identified queue length based on speed profile.

TABLE I. TRAFFIC VOLUME OF SITE STUDY (VEH/HR)

Approaches	**Traffic Volume (veh/hr)**		
	LT	TM	RT
Westbound	464	191	251
Northbound	101	389	243
Eastbound	55	248	213
Southbound	237	1268	312
Total Throughput	3,926		

LT: Left Turn, RT: Right Turn, TM: Through Movement

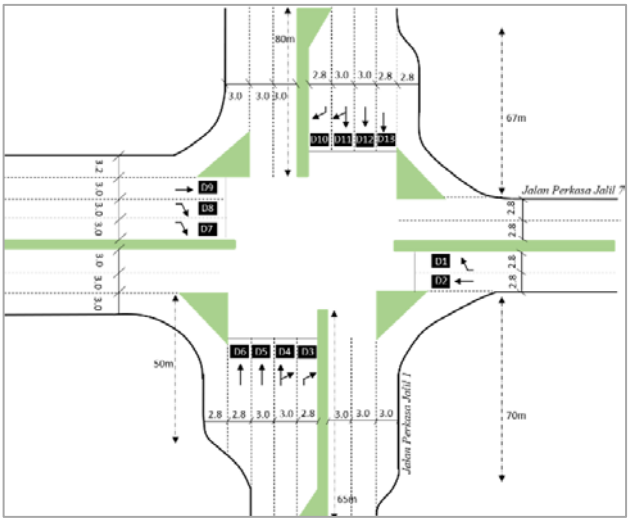


Fig. 3. Intersection Geometry.

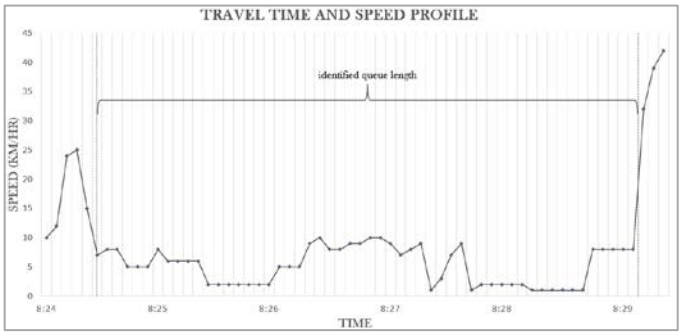


Fig. 4. Travel Time and Speed Profile.

C. DTO Parameters Impact

DTO parameters effect towards intersection performance was investigated by employing different value of saturation flow rate (s), lost time (L) and number of stages (n). As part of the simulation test, a value of 2,250 saturation flow rate is adopted by using a vehicle headway of 1.6s to interpret more aggressiveness of the driver on site during rush hour. This value is obtained from averaging of vehicle headway after four or five vehicles depart from stop line at site. In current configuration, the vehicle headway of 2.1s is adopted for all peak periods.

The intergreen value of existing configuration consist of 2s of amber and 2s of all-red. For the simulation test, a value of 4s amber and 2s that sum up to 6s intergreen is adopted. A longer of intergreen allow vehicle at a distance longer stopping distance allow vehicle to stop comfortably. As mentioned in previous chapter, intergreen value is directed impact of lost time in DTO cycle time calculation.

Besides of DTO parameters abovementioned, the number of signal phasing configuration is impacted on total lost time. The more signal phasing, the higher of total lost time in one cycle. To investigate lost time parameters interaction, existing signal phasing from four stages is configured to be 6 stages. As the number of stages are critical in specifying total lost time, splitting right turn movement and through movement in separate exclusive phase is needed. This overlapping phase is implemented at northbound-southbound (through movement) and eastbound-westbound (right turn) approach. Fig. 5 shows the phase diagram of 6 stages and adjusted value of intergreen time and saturation flow rate.

D. Microsimulation Modelling

- Model Development

The existing four-legged intersection were modelled based on the heaviest peak traffic condition. The Bing Maps available in the VISSIM were used as a background over which the road links and connectors were drawn. Traffic demands and vehicle routing were inputted in the VISSIM. A data collection points, travel time segment and node were defined in the VISSIM.

- Model Calibration

Model calibration is a process of ajusting microsimulation parameters included of driver behaviour characteristic to represent field conditions [16]. According to Wisconsin Department of Transportation (WisDOT), Wiedemennn 74 car following model is preferred to use at urban arterial road [17]. A lower roadway capacity in VISSIM resulted by using higher values of average standstill, additive part of safety distance and multpilcative part of safety distance.

- Model Validation

Validation can be defined as a process to determine the accuracy representation of the model with site condition. A quantitative comparison simulation model output and field observed output can be measure by using a statistical validation [18]. For this study, Root Mean Squared Percentage Error (RMSPE) is applied to penalize large error at higher rate [19]. Statistical measure of RMSPE can be expressed as below.

$$\text{RMSPE} = \sqrt{\frac{1}{N}} \sum_{i=1}^{N} \left(\frac{Mi-Oi}{Oi}\right)^2 \quad (4)$$

Where N = number of samples, M= modeled value, O= observed value.

Output from link volume at stop line and travel time at identified segment were compared with field data. The model is considered validated if the RMSPE value is less than 5% [20].

- Simulation Run Design

The simulation run need to meet a purpose of evaluating DTO cycle length model off using a different DTO parameters value compared to based value from existing junction configuration. Therefore, six (6) scenarios were designed and summarised in Table II. All the simulation runs of each with a different random seed number were performed for 4500 simulation seconds. This simulation period included of 15-minutes warming up simulation period.

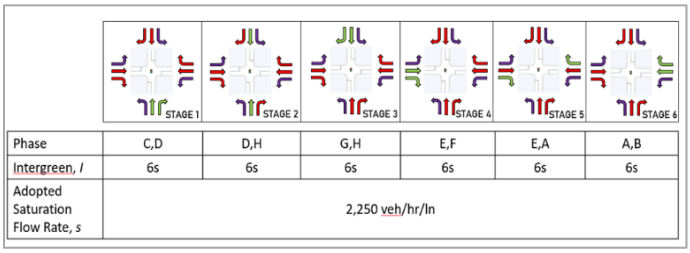


Fig. 5. Adjustment Signal Configuration of Study Site.

TABLE II. SIMULATION SCENARIO DESIGN

Scenario	Sat. Flow Rate, *s*	Intergreen, *I*	No. of Stage, *n*	Total Lost Time, *L*
Base Model	1,714	4	4	16
1	2,250	4	4	16
2	1,714	6	4	24
3	2,250	4	6	24
4	1,714	6	6	36
5	2,250	6	6	36

IV. Results

A. Validation Result

In this study, a validation process involved with comparing modeled and field data in terms of throughput, travel time and maximum queue length. Throughput at each link approach were compared and presented in Fig. 6. Based on the result, the RMSPE is recorded at 0.044 which is less than threshold value. This signifies that the model is replicating site condition in terms of intersection throughput.

As mentioned on previous chapter, travel time and queue length from simulation were compared to the field survey data for validation purposes. The model considers well calibrated if the percentage of difference is less than 10% [21]. Table III tabulates field and simulation result.

B. Intersection Performance

At the end of cycle two, DTO proposed a new cycle length according to the arrival flow at stop line and parameters that has been described on previous chapter. Overall, five (5) alternative scenario of 4,500 seconds were carried out and optimal cycle length generated is shown in Fig. 7. Base model (BM) represented cycle length generated according to current configuration while SC represented an alternative scenario as described in Table II.

Based on the result, Scenario 3, Scenario 4 and Scenario 5 generate large cycle as the number of stage is increase to 6. More number of stages contibuted to the high value of lost time. Meanwhile, Scenario 1 and Scenario 2 generates optimal cycle much similar with the existing configuration scenario.

Performance of the intersection were then evaluated in terms of average delay, travel time and throughput. Fig. 8 demostrate the intersection performance by comparing of average delay and travel time. Scenario 1 and 2 were configured in a same signal phasing (n=4) while Scenario 3, 4 and 5 were run in six (6) signal phasing. Based on the result, Scenario 1,3 and 5 experience less delay compared to the Scenario 2 and Scenario 5 which use an existing parameters value (s = 1,714 veh/hr/ln). Scenario 1, 3 and 5 use a higher value of saturation flow rate (s) to indicates the close headway between vehicle during saturated conditions (s = 2,250 veh/hr/ln).

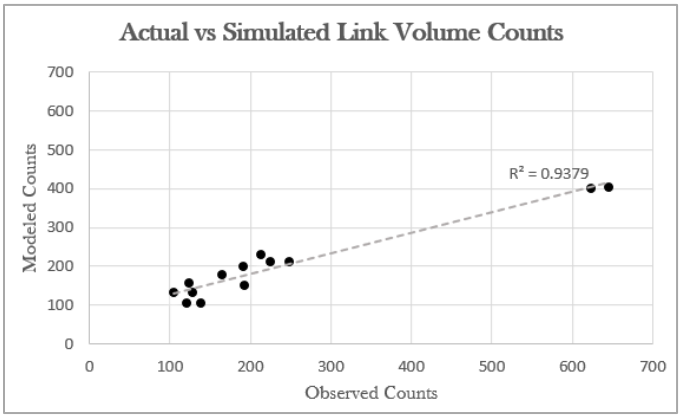


Fig. 6. Optimal Cycle Length Generated by DTO.

TABLE III. Comparison of Field and Simulation Result

Parameters	Data Obtained	
	Field	Simulation
Travel Time (s)	280.00	301.67
Maximum Queue Length (m)	522.19	504.52

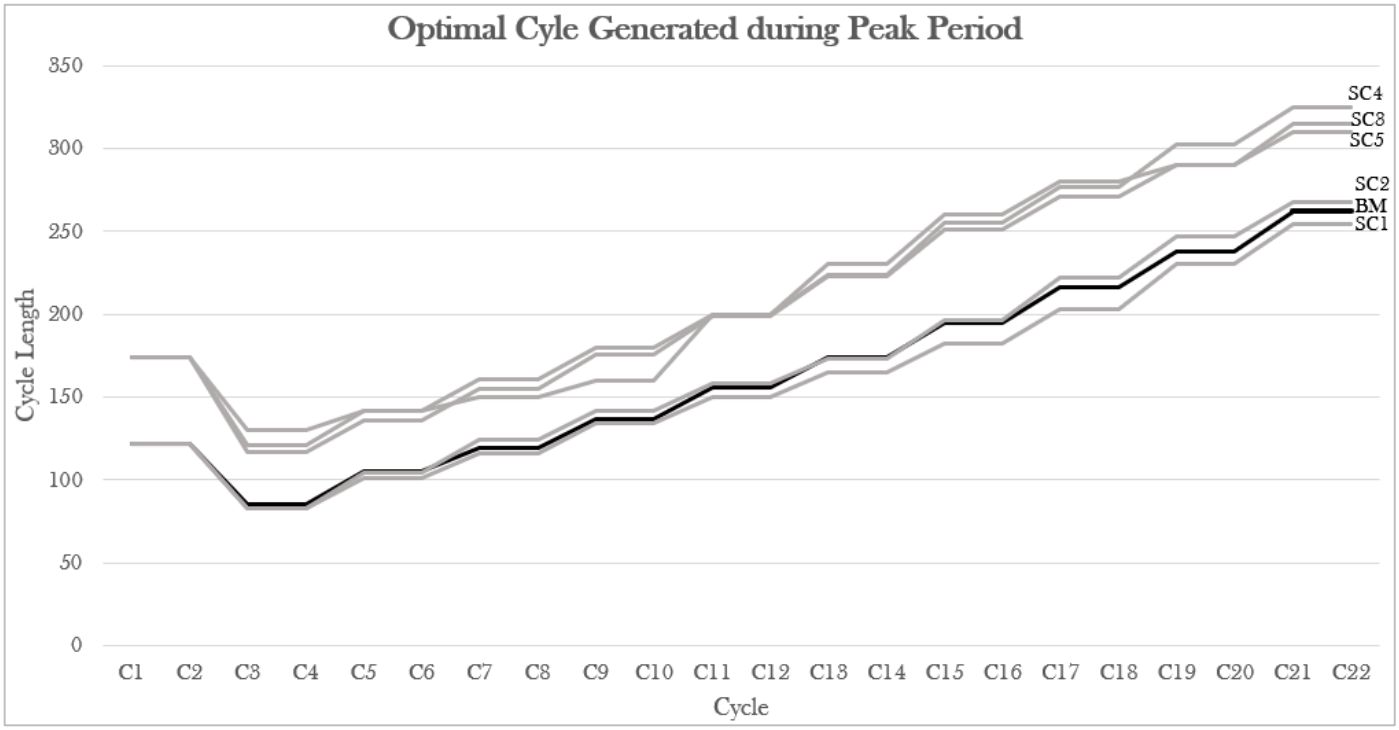


Fig. 7. DTO Optimal Cycle Result.

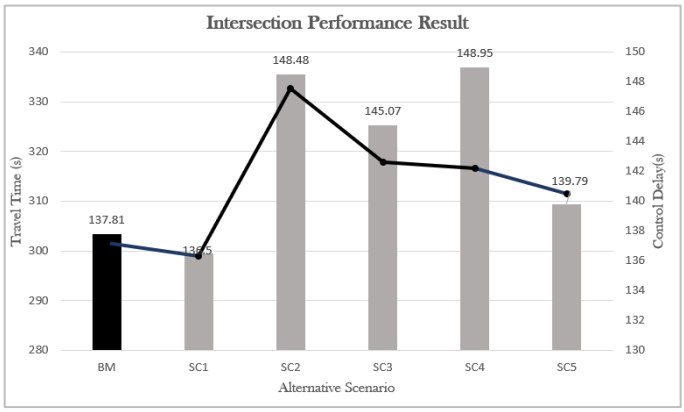


Fig. 8. Intersection Performance Result.

Meanwhile, Scenario 1 has shown a shorten travel time and control delay compared to other scenario included the base model. Scenario 4 experienced a larger delay while Scenario 2 experienced a longer travel time. Both scenario were used a saturation flow rate of 1,714 veh/hr/ln with different signal phasing.

Considering the variability of delay, throughput and average queue length obtained from this study, reliable parameters in DTO should be selected to improve LOS of signalized intersections at study site. It should be noted as HCM highlighted that, the mean control delay is the primary performance measure for signalized intersection. Therefore, Scenario 1 has shown a significant impact to the intersection performance by inputting an appropriate saturation flow rate (s).

Table IV present the maximum throughput, maximum queue length and level of service of each approaches.

Based on the result tabulated in Table IV, Scenario 5 recorded a highest vehicle throughput followed by Scenario 1. Scenario 5 used a highest saturation flow rate (s) and lost time (L) in generating optimal cycle length. This parameter has impacted on generating a longer cycle length in which to allow more vehicle to cross the intersection.

TABLE IV. INTERSECTION PERFORMANCE

Approach	Approach Delay	Approach LOS	Approach Maximum Throughput	Average Queue Length (m)
Scenario 1				
Northbound	69.62	D	686	86.18
Eastbound	130.77	F	485	105.93
Westbound	84.17	F	883	94.78
Southbound	136.19	F	1,089	438.58
Total		F	3,143	
Scenario 2				
Northbound	102.87	F	686	82.91
Eastbound	136.63	F	497	129.00
Westbound	99.12	F	904	91.17
Southbound	144.77	F	999	434.36
Total		F	3,086	
Scenario 3				
Northbound	73.74	E	689	84.00
Eastbound	82.23	F	496	91.55
Westbound	97.48	F	879	80.58
Southbound	177.05	F	1,050	436.24
Total		F	3,114	
Scenario 4				
Northbound	78.4	F	689	78.88
Eastbound	97.64	F	488	93.86
Westbound	123.93	F	839	130.03
Southbound	160.11	F	1,100	438.77
Total		F	3,116	
Scenario 5				
Northbound	84.95	F	696	78.14
Eastbound	107.01	F	490	105.72
Westbound	78.88	E	908	109.04
Southbound	153.61	F	1,102	441.17
Total		F	3,196	

Result shown in the table indicates that most of the approach in every scenario performing at LOS F. This is evident by the long queues and higher delay at most approach.

V. Discussion

This study demonstarates an impact of the traffic parameters in adaptive signal control strategies as known as DTO. Two (2) parameters were introduced during pre-installation of DTO to enhance a cycle length decision.

- Employing a realistic saturation flow rate (s) while inputting parameters in DTO according to real traffic conditions contribute to a better intersection performance. Currently, DTO required one value of saturation flow rate (s) which applied during all peaks (Scenario 1).
- An appropriate saturation flow rate (s) value with extra lost time may contribute for a better intersection performance during peak period (Scenario 3 and 5).
- The intersection performance may not achieve best result if the saturation flow rate (s) were not configured higher to interpret a heavy traffic condition while more signal phasing and intergreen were configured (Scenario 4 and Scenario 2).
- Total maximum throughput is obtained during longer cycle length as more vehicle allow to cross (Scenario 3, 4 and 5). However, longer cycle length impacted on higher delay at intersection as resulted from Scenario 2. This is because Scenario 2 was not interpreting a real vehicle headway where the configuration of saturation flow rate (s) is remained as constant.

VI. Conclusion

As explained in the introduction chapter, saturation flow rate (s) need to be calibrated particularly when demand reached the capacity. Based on the analysis that has been conducted, DTO optimal cycle length is effectively produced by employing a calibrated value of saturation flow rate (s). Further studies will evaluate DTO configuration to consider saturation flow rate (s) as a time-based instead of inputting an average deterministic value in the system. Since this study is part of the DTO ongoing research, intergrating real-time data of vehicle headway from analytic camera is potentially to be part of future study.

This study also discover that optimal cycle length can be produced with the interaction of lost time (L) parameter and calibrated saturation flow rate (s). If the parameters only considered of lost time (L) parameter, a longer cycle length with highest delay is expected to be produced.

Acknowledgment

The authors would like to thank to Research and Development Department, Sena Traffic Systems Sdn Bhd (STS) for their contribution in providing traffic data and signal configuration of the study site. Sincere thanks to the programmer in R&D STS for designing COM API in emulating STS signal control in VISSIM.

References

[1] Transportation Research Board. (2000). Highway Capacity Manual Chapter 10. United States of America: National Academy of Science.

[2] Khisty, C. J. (1990). Transportation Engineering An Introduction . Prentice-Hall.

[3] Shao, C.-q., & Liu, X.-m. (2012). Estimation of Saturation Flow Rates at Signalised Intersections. Discrete Dynamics in Nature and Society.

[4] Bester, C., & Meyers, W. (2007). Saturation Flow Rates . Proceedings of the 26th South African Transport Conference (SATC 2007), (pp. 560-568). Stellenbosch, South Africa.

[5] M.Day, C. (2013). Revisiting the Cycle Length - Lost Time Question with Critical Lane Analysis. Transportation Research Record of the Transportation Research Board.

[6] Hamad, K., & Abuhamda, H. (2015). Estimating Base Saturation Flow Rate for Selected Signalised Intersection in Doha, Qatar. Journal of Traffic and Logistics Engineering, 168-171.

[7] Tang, K., Ono, T., Tanaka, S., & Kuwahara, M. (2012). Re-Examination of Lost Time Estimation and Intergreen Time Design for Right-Turn at Signalised Intersections in Japan. Asian Transport Studies, Volume 1, Issue 4 (2011), (pp. 412-423). Japan.

[8] Zakaria, A. Y., & Rabia, S. I. (2016). Estimating the Minimum Delay Optimal Cycle Length based on a Time-Dependant Delay Formula. Alexandria Engineering Journal, 2509-2514.Akgungor, A. P. (2018, September). Investigating Parameter Interaction with the Factorial Design Method: Webster's Optimal Cycle Length Model. Tehnicki Vjesnik 25(Supplement 2), pp. 391-395.

[9] Webster, F. (1969). Traffic Signal Setting. Road Research Technical Paper No.39. London: Her Majesty Stationary Office.

[10] J.Chen, G. a. (2009). Study on Saturation Flow Rates for Signalised Intersections . International Conference on Measuring Technology and Mechatronics Automation, (pp. 595-601). Zhangjiajie, Hunan.

[11] Akgungor, A. P. (2018, September). Investigating Parameter Interaction with the Factorial Design Method: Webster's Optimal Cycle Length Model. Tehnicki Vjesnik 25(Supplement 2), pp. 391-395.

[12] Transportation Research Board. (2000). Highway Capacity Manual Chapter 10. United States of America: National Academy of Science.

[13] Wanjing, M., & Xinchen, Y. (2013). Optimal Offline Cycle Length Model Based on Online Bus Priority Demand. Journal of Transportation Systems Engineering and Information Technology, 124-129.

[14] Akgungor, A. P., Yildiz, O., & Demirel, A. (2006). A Sensitivity Analysis of the HCM 2000 Model with the Factorial Design Method. Turkish Journal of Engineering and Environmental Sciences, 259-267.

[15] DingXin Cheng, Z. Z. (2005). Development of an Improved Cycle Length Model over the Highway Capacity Manual 2000 Quick Estimation Method. Journal of Transportation Engineering.

[16] Spiegelman, C., & Eun Sug Park, L. R. (2011). Transportation Statistics and Microsimulation. New York: Chapman and Hall/CRC.

[17] Mathew, T. V. (2014). Transportation System Engineering Chapter 16. India: Cell Transmission Model, IIT Bombay.

[18] Law, A. M., & Kelton, W. D. (2000). Simulation Modeling and Analysis. McGraw-Hill.

[19] Balakrishna, R., Antoniou, C., Ben-Akiva, M., Koutsopoulus, H. N., & Wen, Y. (2007). Calibration of Microscopic Traffic Simulation Models. Transportation Research Record 1999, 198-208.

[20] Wisconsin Department of Transportation (WisDOT). (2019, September). Traffic Engineering, Operations & Safety Manual. Retrieved from State of Wisconsin Department of Transportation: https://wisconsindot.gov/dtsdManuals/traffic-ops/manuals-and-standards/teops/16-20.pdf.

[21] Osei, K. K., Adams, C. A., Ackaah, W., & Olver-Commey, Y. (2018). Signalisation Options to Improve Capacity and Delay at Roundabouts through Microsimulation Approach: A Case Study on Arteria; Roadways in Ghana. Journal of Traffic and Transportation Engineering , 1-13.

An Extreme Learning Machine Model Approach on Airbnb Base Price Prediction

Fikri Nurqahhari Priambodo[1], Agus Sihabuddin[2]*
Department of Computer Science and Electronics
FMIPA, Universitas Gadjah, Yogyakarta
Indonesia

***Abstract*—The base price of Airbnb properties prediction is still a new area of prediction research, especially with the Extreme Learning Machine (ELM). The previous studies had several suggestions for the advantages of ELM, such as good generalization performance, fast learning speed, and high prediction accuracy. This paper proposes how the ELM approach is used as a prediction model for Air BnB base price. Generally, the steps are setting hidden neuron numbers, randomly assigning input weight and hidden layer biases, calculating the output layer; and the entire learning measure finished through one numerical change without iteration. The performance of the model is estimated utilizing mean squared error, mean absolute percentage error, and root mean squared error. Experiment with Airbnb dataset in London with twenty-one features as input generates a faster learning speed and better accuracy than the existing model.**

Keywords*—Airbnb; base price prediction; extreme learning machine; fast learning

I. Introduction

Airbnb is a property-sharing marketplace that allows property holders and tenants to put their properties on the web, with the goal that guests can pay to stay in them as part of the hospitality business. In the hospitality domain, evaluating and income, the executives, are two top now and again explored zone because of the theoretical and practical criticality of room estimating. Airbnb has must ace room valuing to build their gainfulness after fulfilling visitor desires [1].

Paid third-party pricing software is available, but generally, the property owners are needed to put their regular daily value (base cost); and the calculation will differ everyday price around that base cost. As a platform provider, Airbnb does not control how their hosts set prices for their postings, yet it gives an assortment of apparatuses to enable their hosts to put their prices all the more adequately. For instance, they permit hosts to set altered day by day rates, end of the week costs, and limits for long haul stays, so the base price determination becomes an essential process [2].

To be able to determine the property price, some researchers use various methods with three components to determine the price: i) a binary classification model predicts the booking probability of each listing night, ii) a regression model predicts the ideal cost for each listing night, iii) personalization reasoning on top of the yield from the resulting model to deliver the last cost suggestions [2].

In contrast to evaluating issues where valuing techniques are applied to an enormous number of indistinguishable items, there are no identical items on Airbnb. Each listing property on the Airbnb platform offers unique qualities and encounters to visitors. A geographically weighted regression (GWR) approach to distinguish a few variables corresponded with Airbnb listing costs has been implemented. The unique nature of Airbnb listing makes it exceptionally hard to appraise a precise interest bend needed to apply traditional income expansion evaluating techniques. The offline and online evaluation result of this pricing model show that the proposed model performs better than an immediate max-fire up evaluating methodology because of max-fire up pricing strategy are likely to suffer from demand curve which is hard to estimate [3].

Several machine learning algorithms such as a ridge, random forest regressor, linear regression, decision tree regressor, and lasso have been used in the forecast of housing selling costs for prediction. The result shows that feature selection is a significant component. Two special exhibitions of machine learning are the precision of forecast and averaging of errors or fitness, which might be influenced by the highlights chose with various gatherings of relationship levels [4].

The ELM has excellent potential for system prediction and modeling, i.e., an ELM based indicator for genuine frequency stability assessment (FSA) of power systems [5], electricity price forecasting [6], sales forecasting [7], security evaluation of wind power system [8], and drying system modeling [9]. However, there is no discussion about the Extreme Learning Machine (ELM) for Airbnb price forecast.

Over the past few years, a simple learning algorithm ELM for a single hidden layer feedforward network (SLFNs) was introduced [10]. ELM has superior faster-learning speed and better generalization performance than traditional feedforward network learning algorithms such as backpropagation (BP) algorithm. ELM achieves similar or preferred speculation execution over Support Vector Machine for regression and binary class, and much better speculation execution for multiclass classification cases [11].

ELM does not have to build up an exact numerical model of the object and appreciate the characteristics of the item. The technique accomplishes deficiency areas only by restricted defect tests for preparing and learning. ELM picks input loads and hidden biases randomly and then analytically computes

*Corresponding Author

yield loads with Moore-Penrose derive pseudo-inverse [12]. The entire cycle is assessed without iterations, so the learning cycle is a very time-productive strategy. ELM Method overcomes numerous issues in gradient-based learning algorithms, for example, learning rate, stopping criterion, reducing local minima, and the number of epochs [13]. The objective of this research are: 1) to provide a new approach for the base price prediction for Airbnb; 2) to give a more accurate and faster prediction model for Airbnb base price prediction.

II. Related Works

In previous studies, price forecasting has been carried out. There are several ELM methods that have been applied to price forecasting, namely gold price forecasting [14]; in this paper, a learning algorithm for a single hidden layered feedforward neural network called Extreme Learning Machine (ELM) is utilized, which has good learning capacity. Also, this examination dissects the five models, explicitly feedforward backpropagation networks, feedforward networks without feedback, radial basis function, ELMAN networks, and ELM learning model. The outcomes demonstrate that ELM learning performs in a way that is better than different techniques.

Based on [15], this research focus on the issue of how to plan an approach that can improve the forecast exactness just as accelerate expectation measure for stock market prediction. Initially, so as to get the most critical highlights of the market news records, this article is proposed a new feature selection algorithm called NRDC, just as another component weighting calculation (N-TF-IDF) to help increase the expected precision. Exploratory outcomes demonstrate that the N-N-K-ELM model can accomplish better execution on the thought of both forecast exactness and expectation speed as a rule.

In [5], ELM is used as the predictor for real-time. This article surveys the ELM's applications in power planning and a short time later develops an ELM-based indicator for real-time frequency stability assessment (FSA) of power systems. The contributions of the indicator are power system operational parameters, and the yield is recurrence soundness edge that measures the security level of the power system subject to a chance. By disconnected training with a frequency stability database, the indicator can be online was applied for real-time FSA. Profiting by the rapid speed of ELM, the predictor can be online are refreshed for upgraded robustness and reliability.

ELM is also used for electricity market prices by [6]. In this article, a fast electricity market price forecast is proposed dependent on an as of late developed learning technique for single hidden layer feedforward neural networks, the extreme learning machine (ELM), to defeat these disadvantages. The new methodology additionally has improved value stretches gauge exactness by incorporating a bootstrapping method for vulnerability assessments. The outcomes show the extraordinary capability of this proposed approach for online precise price forecasting at the spot market costs assessment.

In this research, ELM will be used for predicting Airbnb property base price with one preferred position is predominant quicker learning pace and better speculation execution with a theory that will improve the accuracy of the model.

III. Theoretical Basis

A. Extreme Learning Machine

Extreme Learning Machine is a single hidden layer feedforward neural network (SLFNs) and is a sort of straightforward and powerful learning algorithm [12]. It merely needs to set the hidden layer nodes and infinitely differentiable actuation work before preparing by clarifying the minimum norm least-squares of a linear equation to the ideal arrangement. Arbitrarily picked the hidden biases and input weights, and the yield loads are determined systematically with a given number of hidden neurons. The entire cycle of computation finishes once at a time without iteration.

The output function of ELM for generalized SLFNs as shown in (1).

$$f_L(x) = \sum_{i=1}^{L} \beta_i h_i(x) = h(x)\,\beta \quad (1)$$

Where $\beta = [\beta_1, \dots, \beta_L]^T$ is the output weight vector between the hidden layer of L nodes to the $m \geq 1$ output nodes, and $h(x) = [h_1(x) h_L(x)]$ is a nonlinear feature mapping. The output (row) vector of the hidden layer with respect to the input x. $h_i(x)$ is the output of the i^{th} hidden node output. The output functions of hidden nodes may not be unique. Different output functions may be used in other hidden neurons. In particular, in a real application $h_i(x)$ can be formulated in (2).

$$h_i(x) = G(w_I, b_I, x) \quad w_i \in R^d, b_i \in R \quad (2)$$

Where G(w,b,x) (with hidden node parameters (w,b)) is a nonlinear piecewise continuous function satisfying ELM universal approximation capability theorems [11], [16].

A standard SLFNs with $L(N_0 \geq L)$ hidden layer nodes and the activation function g(x) are mathematically modeled as (3).

$$\sum_{i=1}^{L} \beta_i g_i(x_j) = \sum_{i=1}^{L} \beta_i g_i\,(w_i, x_i + b_i) = o_j \quad j = 1,2 \dots, N \quad (3)$$

where $w_i = [w_{i1}, w_{i2}, \dots, w_{in}]^T$ is the weight vectors connecting the i^{th} hidden node and input nodes, $\beta_i = [\beta_{i1}, \beta_{i2}, \dots, \beta_{in}]^T$ is the weight vectors connecting the i^{th} hidden node and output nodes, b_I is the bias of the i^{th} hidden node, $w_i . x_j$ represents the inner product of w_i and x_j, the network structure, as shown in Fig. 1.

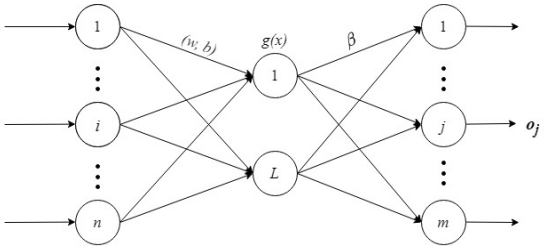


Fig. 1. A General ELM Network Structure.

To make SLFNs with L hidden layer nodes and the activation function $g(x)$ approximate the N samples with zero error mean that $\sum_{i=1}^{L}\|o_j - t_j\| = 0$, namely, existing β_I, w_I and b_I make formula (4) established.

$$\sum_{i=1}^{L}\beta_i g\left(w_i . x_j + b_i\right) = t_j, j = 1,2 \dots, N \quad (4)$$

The equation (4) can be written compactly as (5)

$$H\beta = T \quad (5)$$

where,

$$H(w_1, \dots, w_L, b_1, \dots, b_L, x_1, \dots, x_N)$$

$$\begin{bmatrix} g(w_1 . x_1 + b_1) & \cdots & g(w_L . x_1 + b_1) \\ \vdots & \dots & \vdots \\ g(w_1 . x_N + b_1) & \cdots & g(w_L . x_N + b_1) \end{bmatrix}_{N\times L},$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_1^T \end{bmatrix}_{L\times m}, \text{ and, } T = \begin{bmatrix} t_1^T \\ \vdots \\ t_n^T \end{bmatrix}_{N\times m} \quad (6)$$

From (6) H is the hidden layer output matrix of the neural network, the ith column of H represents the output matrix about $x_I, \dots, x_n$ of the ith hidden layer node.

When the activation function infinitely differentiable, the input connection weights w_I moreover, hidden layer bias b_I can randomly set at the beginning of the training, and they will be fixed in the training process; the output connection weights β are obtained by solving the least-squares solution of linear (7), the result as (8).

$$\min\|H\beta - T\| \quad (7)$$

$$\beta = H^{\dagger}T \quad (8)$$

where $H^{\dagger}$ is the Moore-Penrose generalized inverse of the hidden layer output matrix H . The mathematical transformation determines the output weights. This ensures that the long training phrase when network boundaries are iteratively changed with some reasonable learning boundaries (like iterations and learning rate) is not needed.

ELM learning algorithm steps can be summarized as follows:

Step 1: Given a training set $(x_I, t_i) = (i = 1,2, \dots, N)$, the activation function is $g(x)$, number of hidden layer nodes is L, setting the input weights w_I and hidden layer bias b_I randomly.

Step 2: Calculate the output matrix H of the hidden layers.

Step 3: Calculate the output weights β by (8).

Because the ELM algorithm does not require iterative input weights and bias in training adjustment, it reduces the complexity of the training, and the training speed improved obviously.

B. Performance Evaluation

An evaluation needs to be performed to measure performance and to provide feedback that can serve to improve the model. In this research, root mean squared error (RMSE), mean square error (MSE), and mean absolute percentage error (MAPE) measurements were performed [17].

MSE is a measure of prediction accuracy by squaring each error for each observation in a data set and then obtaining the average number of squares. MSE gives greater weight to the error compared to a small error because the error is raised before adding up. MSE can be calculated by (9).

$$MSE = \frac{1}{n} \times \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (9)$$

RMSE is the square root value of the average square error and formulated as in (10).

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (10)$$

MAPE is the average percentage of the sum of the differences between prediction results with actual data. The formula for the mean absolute percentage error can be written as follows in (11).

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \quad (11)$$

Given y_i is actual value, $\hat{y}_i$ is predicted value, i is the ith data, n is the number of data.

A study from [18] shown that using MAPE as a measure of quality for regression models is feasible both on a practical point of view and on a theoretical one.

IV. Methods

A. Data Description

The dataset utilized for this examination originates from InsideAirbnb.com. The dataset was downloaded on 9 April 2019 and contained data on all London Airbnb listings that were live on the webpage on that date, which is 79.671 Airbnb listing. The data itself has 106 features and saved in `.csv` format.

In this study, the base price prediction is predicted from 21 input features that will be chosen after the data preprocessing of the Airbnb listing. In the sample dataset, the input features that contains a price value are the advertised price of its Airbnb listing, security deposit fee, cleaning fee, and extra people fee.

B. Preprocessing

The dataset needs to be changed or prepared according to the needs. The original dataset has 106 features, including quite a few text columns of all the different description fields.

Some features or columns will be dropped because of: not indicated to be useful for predicting price, and there are many null or NaN entries. Some entries connected with the fee and having missing values were replaced with the median to avoid fractions by 0.

C. Network Architecture

To solve the Airbnb base pricing issue, the ELM model was designed. The proposed methods can see in Fig. 2. Twenty-one input was picked to ELM from the dataset,

though the base price is assigned the yield of ELM. The performance of ELM relies upon be utilized the type of activation function and the number of hidden neurons. A sigmoid function was chosen as it is not too delicate to the user-determined parameters and determined by a testing parameter to find the best model of ELM.

Twenty-one features are chosen after the preprocessing as input due to their relationship to basic fee types of a property price are shown in Table I.

The test will be divided into two parts. The first test was conducted to learn the amount of training data and test data against the evaluation parameter value. The second test is to determine how the number of hidden neurons affects the evaluation parameter value and execution time; for this purpose, 100 steps regularly increase the number of hidden layer neurons from 100 to 1000. Each test will be conducted ten attempts and will be evaluated based on the average evaluation value and execution time.

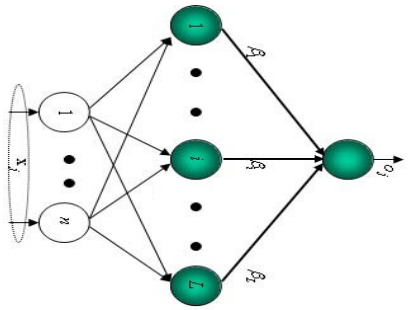


Fig. 2. The Architecture of Extreme Learning Machine.

TABLE I. FEATURES INPUT

No	Feature
1	number of people that a property can accommodate
2	the cleaning fees
3	number of days available for the next 90 days
4	the fee per extra person
5	total reviews that a property has
6	the number of bathrooms
7	the security deposit
8	the minimum night stays
9	the maximum night stays
10	property type
11	room type
12	location of the property
13	review ratings
14	the duration between the first and most recent review
15	the type of cancellation policy
16	whether the property is instantly bookable
17	the presence of a wide range of amenities
18	host response times in the past 30 days
19	host percentage of response rates
20	Super-host
21	total host listings

V. RESULT AND DISCUSSION

The ELM parameter testing aims to determine the best parameters for benchmarking and results in Airbnb base price predictions. The first tests were conducted to learn the amount of training data and test data against the evaluation parameter value; the second is to resolve how the count of hidden neurons affects the evaluation parameter value and execution time. Each parameter value will be tested as many as ten times the experiment.

The first test conducted is to determine the size ratio of data training and data testing. The size ratio of training and testing data is 70:30 and 80:20. At this stage, the count of hidden neurons will be used as a control variable in testing. The number of hidden neurons used is 100. The results from testing the parameter ratio of training and testing data are indicated in Table II.

From the test results at the training data ratio and test of the ELM model, it was apparent that the training and test data with a ratio of 70%:30% had better average performance compared to the training and test data ratio of 80%:20% for all accuracy parameters.

A. Number of Hidden Neuron Testing

The next test is determining the count of hidden neurons on a hidden layer. The count of hidden neurons to be tested are 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000 as indicated in [20]. The MSE accuracy from this step is shown in Table III.

The MSE result shows that accuracy tends to decrease as the number of hidden neurons increases, and it is more evident in the chart of average MSE versus hidden neurons size in Fig. 3.

The experiment results, as depicted in Fig. 3, implicates that 1000 hidden neurons perform better MSE accuracy than the other hidden neurons for the lowest, highest, and average MSE training and testing except for 500 hidden nodes on average MSE testing.

TABLE II. MSE, RMSE, AND MAPE ON DIFFERENT DATA SPLIT

	MSE Training		**MSE Testing**	
Data Split	80:20	70:30	80:20	70:30
Lowest	0.590	0.573	0.622	0.589
Highest	0.901	0.910	1.027	0.940
Average	0.770	0.713	0.802	0.743
	RMSE Training		**RMSE Testing**	
Data Split	80:20	70:30	80:20	70:30
Lowest	0.768	0.757	0.789	0.767
Highest	0.949	0.954	1.013	0.970
Average	0.875	0.842	0.893	0.860
	MAPE Training		**MAPE Testing**	
Data Split	80:20	70:30	80:20	70:30
Lowest (%)	14.027	13.374	14.113	13.417
Highest (%)	17.054	17.497	17.624	17.393
Average (%)	15.821	15.241	15.879	15.148

TABLE III. MSE ACCURACY ON A DIFFERENT NUMBER OF HIDDEN NEURONS

Neuron	MSE TRAINING			MSE TESTING		
	Low	*High*	*Avg.*	*Low*	*High*	*Avg*
100	0.314	0.385	0.350	0.338	0.425	0.374
200	0.108	0.135	0.124	0.129	0.303	0.194
300	0.081	0.097	0.086	0.096	0.302	0.170
400	0.066	0.072	0.068	0.086	0.230	0.134
500	0.055	0.061	0.058	0.072	0.151	0.096
600	0.051	0.055	0.053	0.070	0.251	0.135
700	0.046	0.048	0.047	0.064	0.238	0.134
800	0.043	0.046	0.044	0.061	0.299	0.106
900	0.040	0.042	0.041	0.064	0.247	0.137
1000	0.038	0.039	0.039	0.057	0.198	0.098

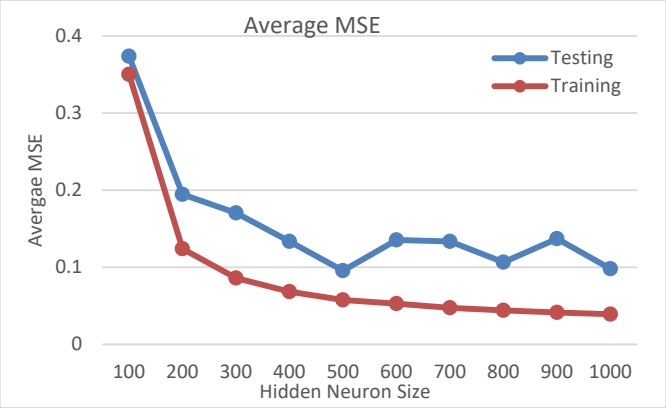


Fig. 3. Average MSE Testing Result.

The next accuracy testing is for RMSE, as shown in Table IV. It has the same behavior as MSE in which the more hidden nodes in the model, the smaller RMSE is.

As the average RMSE is transformed in to chart for the average RMSE, it is more obvious too that the behavior is detected as seen in Fig. 4.

TABLE IV. RMSE TESTING ON THE NUMBER OF HIDDEN NEURONS

Neuron	RMSE TRAINING			RMSE TESTING		
	Low	*High*	*Avg.*	*Low*	*High*	*Avg*
100	0.560	0.621	0.592	0.582	0.652	0.611
200	0.329	0.368	0.352	0.360	0.550	0.437
300	0.285	0.311	0.293	0.310	0.549	0.405
400	0.256	0.268	0.261	0.293	0.480	0.358
500	0.234	0.247	0.240	0.268	0.389	0.306
600	0.225	0.234	0.230	0.264	0.501	0.357
700	0.214	0.220	0.218	0.252	0.488	0.354
800	0.208	0.214	0.210	0.247	0.547	0.315
900	0.201	0.206	0.203	0.253	0.497	0.359
1000	0.196	0.198	0.198	0.239	0.445	0.304

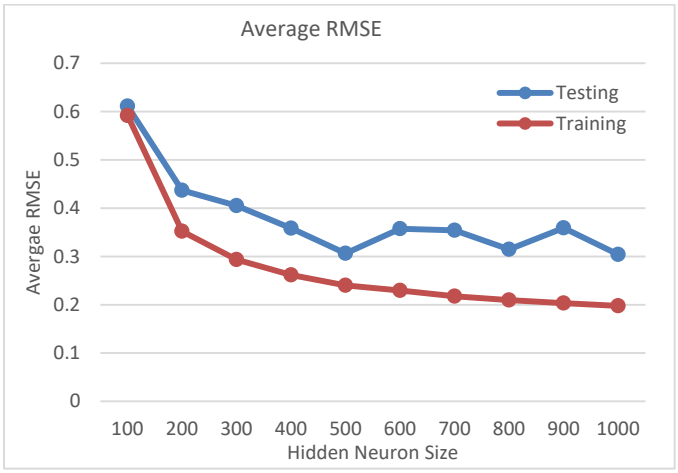


Fig. 4. Average RMSE Testing Result.

The last accuracy data is for the MAPE on various hidden neurons, as shown in Table V. The result is similar with two other accuracy parameters where the more neurons, the less accuracy parameter for the MAPE. It demonstrates the same behavior too for the nodes of 500 and 800.

Table V shown the average MAPE results for each testing parameter; like the previous evaluation method, testing the number of neurons on a hidden layer indicates that with many neurons, it will result in a small error value. If the average MAPE is charted, the same pattern is appearing for the MAPE, as shown in Fig. 5.

From Fig. 3, 4, and 5, it can be seen when the number of neurons increased, and the accuracy tends to increase too. It could be caused by the more neurons, the more link between the input and output layer, and leads to a better quality of learning process. The more link between input and output layer needs more computation time, as indicated in Table VI.

From Table VI, we can see that 1000 hidden neurons in terms of duration cost almost six times longer than 100 hidden neurons. The time increasing for the computation tend to be a linear graph, as shown in Fig. 6.

TABLE V. MAPE TESTING ON SOME HIDDEN NEURONS

Neuron	MAPE TRAINING			MAPE TESTING		
	Low	*High*	*Avg.*	*Low*	*High*	*Avg*
100	9.87	10.99	10.39	9.93	11.15	10.38
200	5.84	6.52	6.24	6.25	10.87	7.87
300	4.97	5.47	5.14	5.05	11.16	7.43
400	4.38	4.65	4.50	4.55	9.52	6.22
500	3.94	4.17	4.05	4.01	6.84	4.90
600	3.75	3.94	3.83	3.79	10.00	6.31
700	3.51	3.62	3.57	3.51	9.98	6.23
800	3.35	3.47	3.40	3.40	11.03	5.14
900	3.19	3.34	3.25	3.56	9.84	6.33
1000	3.06	3.15	3.11	3.07	8.57	4.88

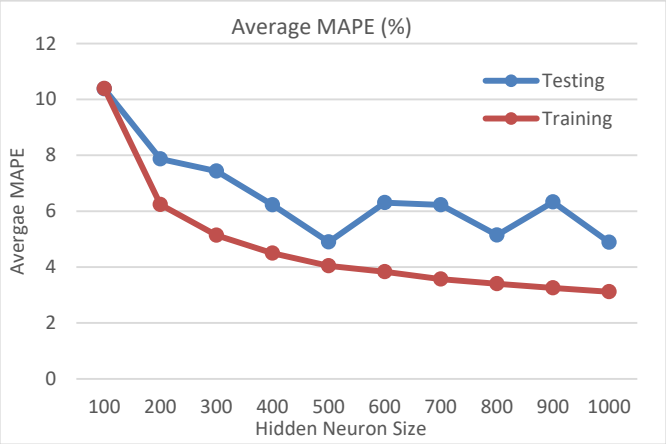


Fig. 5. Average MAPE Testing Result.

TABLE VI. ELM COMPUTATIONAL TIME

Neuron	TRAINING TIME (s)		
	Low	*High*	*Avg.*
100	0.931	1.062	1.032
200	1.833	2.154	1.967
300	2.248	2.692	2.366
400	2.623	2.833	2.723
500	3.015	3.206	3.097
600	3.447	3.991	3.566
700	3.929	4.111	4.029
800	4.300	4.565	4.472
900	4.884	5.260	5.033
1000	5.420	5.653	5.522

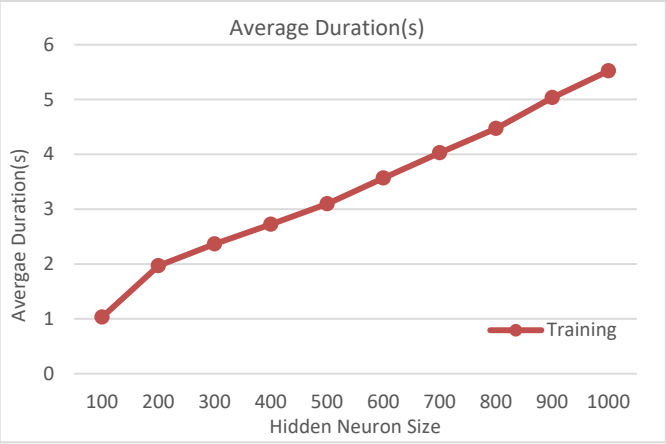


Fig. 6. The Average Computational Time for Test Data.

As observed in Fig. 6, the execution time for the model continue expanding as the number of hidden neurons ascends because the learning time of ELM is nearly spent on computing the Moore-Penrose generalized inverse $H^{\dagger}$ of the hidden layer output matrix H.

Based on the tests already conducted, the parameters to be used in determining Airbnb price prediction results with the ELM method are the ratio of training and test data of 70%: 30% and the number of hidden neurons = 1000.

B. Comparison with XGBoost

In the testing phase, the performance of the resulting model was evaluated using three selected methods, namely MAPE, MSE, and RMSE. As a comparison, other predictions are made using the XGBoost method [19]. The test results for the three methods can be seen in Table VII.

TABLE VII. ELM MODEL COMPARISON AGAINST XGBOOST

Method	ELM		XGBoost	
	Train Data	*Test Data*	*Train Data*	*Test Data*
MAPE (%)	3.06	3.07	6.70	6.75
MSE	0.038	0.057	0.181	0.198
RMSE	0.196	0.239	0.426	0.445

In the MAPE method, the model produces a value of 3.06% for the training data and 3.07% for the test data, while for the MSE method, the model has a value of 0.038 for the training data and 0.057 for the test data, and for the RMSE method, the model produces a value of 0.196 for the training data and 0.239 for test data. While for the XGBoost model, In the MAPE method, the model has a value of 6.70% for the training data and 6.75% for the test data, while for the MSE method, the model produces a value of 0.181 for the training data and 0.198 for the test data, and for the RMSE method, the model produces a value of 0.4260 for the training data and 0.4454 for test data. This is very clear that the ELM model outperforms the XGBoost model for the three accuracy parameters.

As observed in Table VIII, In terms of execution time, the ELM model takes 5.4 seconds for training data and 6.341 seconds for test data. On the other hand, the XGBoost model takes 12.8950 seconds for training data and 12.8950 seconds for test data.

TABLE VIII. THE AVERAGE DURATION TIME OF ELM AND XGBOOST

Algorithm	Training time (s)	Testing time (s)
ELM	5.419	6.341
XGBOOST	12.895	12.895

VI. CONCLUSION

In conclusion, we have developed an ELM prediction model approach on Airbnb base price and tested it by London Airbnb Listing in April 2019. The model is trained using 21 features, 70%:30% data split, and a maximum of 1000 neurons. The experiment results show that the model is having a good accuracy with the best average MSE value of 0.096, RMSE value of 0.304, and MAPE value of 4.88% for the test data. These accuracies are mostly achieved by 1000 neurons. From the experiments, the model show as the count of neurons raises, the link between the input and output layers would consequently increase. This leads to a better quality of learning. These accuracy parameters are outperforming the XGBoost algorithm and having a much faster learning time with better accuracy.

For further research, the number of neurons for training can be expanded to more than 1000 neurons with more powerful hardware, so the convergence point with the number of neurons where the accuracy reached the optimum value could be found. The base price prediction can be expanded to a daily basis prediction with more features like scheduled events, holidays, and many other features.

ACKNOWLEDGMENT

The author would like to express our gratitude to the Direktorat Penelitian Universitas Gadjah Mada, Indonesia, for providing the research grant.

REFERENCES

[1] M. Chattopadhyay and S. K. Mitra, “Do Airbnb host listing attributes influence room pricing homogenously?,” Int. J. Hosp. Manag., vol. 81, no. September 2018, pp. 54–64, 2019, doi: 10.1016/j.ijhm.2019.03.008.

[2] P. Ye et al., “Customized regression model for Airbnb dynamic pricing,” Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 932–940, 2018, doi: 10.1145/3219819.3219830.

[3] Z. Zhang, R. J. C. Chen, L. D. Han, and L. Yang, “Key factors affecting the price of Airbnb listings: A geographically weighted approach,” Sustain., vol. 9, no. 9, pp. 1–13, 2017, doi: 10.3390/su9091635.

[4] T. Mohd, S. Masrom, and N. Johari, “Machine learning housing price prediction in Petaling Jaya, Selangor, Malaysia,” Int. J. Recent Technol. Eng., vol. 8, no. 2 Special Issue 11, pp. 542–546, 2019, doi: 10.35940/ijrte.B1084.0982S1119.

[5] Y. Xu, Y. Dai, Z. Y. Dong, R. Zhang, and K. Meng, “Extreme learning machine-based predictor for real-time frequency stability assessment of electric power systems,” Neural Comput. Appl., vol. 22, no. 3–4, pp. 501–508, 2013, doi: 10.1007/s00521-011-0803-3.

[6] X. Chen, Z. Y. Dong, K. Meng, Y. Xu, K. P. Wong, and H. W. Ngan, “Electricity price forecasting with extreme learning machine and bootstrapping,” IEEE Trans. Power Syst., vol. 27, no. 4, pp. 2055–2062, 2012, doi: 10.1109/TPWRS.2012.2190627.

[7] F. L. Chen and T. Y. Ou, “Sales forecasting system based on Gray extreme learning machine with Taguchi method in retail industry,” Expert Syst. Appl., vol. 38, no. 3, pp. 1336–1345, 2011, doi: 10.1016/j.eswa.2010.07.014.

[8] Y. Xu, Z. Y. Dong, Z. Xu, K. Meng, and K. P. Wong, “An intelligent dynamic security assessment framework for power systems with wind power,” IEEE Trans. Ind. Informatics, vol. 8, no. 4, pp. 995–1003, 2012, doi: 10.1109/TII.2012.2206396.

[9] A. Balbay, Y. Kaya, and O. Sahin, “Drying of black cumin (Nigella sativa) in a microwave assisted drying system and modeling using extreme learning machine,” Energy, vol. 44, no. 1, pp. 352–357, 2012, doi: 10.1016/j.energy.2012.06.022.

[10] G. Bin Huang, Q. Y. Zhu, and C. K. Siew, “Extreme learning machine: Theory and applications,” Neurocomputing, vol. 70, no. 1–3, pp. 489–501, 2006, doi: 10.1016/j.neucom.2005.12.126.

[11] G. Bin Huang, H. Zhou, X. Ding, and R. Zhang, “Extreme learning machine for regression and multiclass classification,” IEEE Trans. Syst. Man, Cybern. Part B Cybern., vol. 42, no. 2, pp. 513–529, 2012, doi: 10.1109/TSMCB.2011.2168604.

[12] G. Bin Huang, X. Ding, and H. Zhou, “Optimization method based extreme learning machine for classification,” Neurocomputing, vol. 74, no. 1–3, pp. 155–163, 2010, doi: 10.1016/j.neucom.2010.02.019.

[13] G. Bin Huang, Q. Y. Zhu, and C. K. Siew, “Extreme learning machine: A new learning scheme of feedforward neural networks,” IEEE Int. Conf. Neural Networks - Conf. Proc., vol. 2, pp. 985–990, 2004, doi: 10.1109/IJCNN.2004.1380068.

[14] S. K. Chandar, M. Sumathi, and S. N. Sivanadam, “Forecasting gold prices based on extreme learning machine,” Int. J. Comput. Commun. Control, vol. 11, no. 3, pp. 372–380, 2016, doi: 10.15837/ijccc.2016.3.2009.

[15] F. Wang, Y. Zhang, H. Xiao, L. Kuang, and Y. Lai, “Enhancing Stock Price Prediction with a Hybrid Approach Based Extreme Learning Machine,” Proc. - 15th IEEE Int. Conf. Data Min. Work. ICDMW 2015, pp. 1568–1575, 2016, doi: 10.1109/ICDMW.2015.74.

[16] G. Bin Huang and L. Chen, “Enhanced random search based incremental extreme learning machine,” Neurocomputing, vol. 71, no. 16–18, pp. 3460–3468, 2008, doi: 10.1016/j.neucom.2007.10.008.

[17] J. G. De Gooijer and R. J. Hyndman, “25 Years of Time Series Forecasting,” Int. J. Forecast., vol. 22, no. 3, pp. 443–473, 2006, doi: 10.1016/j.ijforecast.2006.01.001.

[18] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, “Mean Absolute Percentage Error for regression models,” Neurocomputing, vol. 192, pp. 38–48, 2016, doi: 10.1016/j.neucom.2015.12.114.

[19] L. Lewis, “Predicting Airbnb prices with machine learning and deep learning,” Medium - Toward Data Science, 2019, [Online]. Available: https://towardsdatascience.com/predicting-Airbnb-prices-with-machine-learning-and-deep-learning-f46d44afb8a6.

[20] F.H. Priambodo, A. Sihabuddin, Extreme Learning Machine Prediction Model on Airbnb Base Price, Thesis, Universitas Gadjah Mada, Yogyakarta, 2020.

ITTP-PG: A Novel Grouping Technique to Enhance VoIP Service Bandwidth Utilization

Mayy Al-Tahrawi[1], Mosleh Abulhaj[2], Yousef Alraba'nah[3], Sumaya N. Al-Khatib[4]
Department of Computer Science, Al-Ahliyya Amman University, Amman, Jordan[1]
Department of Networks and Information Security, Al-Ahliyya Amman University, Amman, Jordan[2, 4]
Department of Software Engineering, Al-Ahliyya Amman University, Amman, Jordan[3]

***Abstract*—Recently, the field of telecommunications started to migrate to Voice over Internet Protocol (VoIP) service. VoIP service applications produce packets with short payload sizes to reduce packetization delay. That is, increasing the preamble size and expends the network link bandwidth. Packet grouping is a technique to enhance the employment of network link bandwidth. Numerous grouping techniques are suggested to enhance link bandwidth employment when using RTP/UDP protocols. Unlike previous research, this article suggests a packet grouping technique that works over the Internet Telephony Transport Protocol (ITTP), not RTP/UDP. This technique is called ITTP Packet Grouping (ITTP-PG). The ITTP-PG technique groups VoIP packets, which exist in the same route, in a single ITTP/IP preamble instead of an ITTP/IP preamble to each packet. Consequently, preamble size is diminished and network link bandwidth is saved. ITTP-PG also adds 3-byte runt-preamble to each packet to distinguish the grouped packets. The suggested ITTP-PG technique is simulated and compared with the conventional ITTP protocol (without grouping) using three elements, namely, the number of concurrent VoIP calls, preamble overhead, and bandwidth usage. Based on all these elements, the ITTP-PG technique outperforms the conventional ITTP protocol. For example, the result shows that bandwidth usage improved by up to 45.9% in the tested cases.**

Keywords*—Voice over Internet Protocol (VoIP); Internet Telephony Transport Protocol (ITTP); packet grouping; network bandwidth

I. INTRODUCTION

The Internet has grown explosively in the fast few decades. This explosive growth has been accompanied by a tremendous number of new technologies such as Voice over IP (VoIP) [1,2]. There are quite a large number of applications that are used for making VoIP conversations, including Skype, FaceTime, and Google Hangout. The number of registered Skype users in 2017 was around 1.33 million [3,4]. Despite that, VoIP faces two main dilemmas that slows down its propagation among organizations and people. First, the low quality of the VoIP conversation, particularly compared to the typical telecommunication system. This is because VoIP infrastructure is IP-based, thus, it shared and does not provide a dedicated channel for the VoIP conversations. However, the traditional telecommunication systems provide a dedicated channel for its VoIP conversation [5,6]. Second, the failure in making the best use of network bandwidth. This is because a considerable size of VoIP packet protocols is attached to the small packet payload that is produced by the VoIP codecs [7,8].

Codec (a portmanteau of coder-decoder) is a tool (hardware or software) that converts the analog voice data to digital data. The digital voice data, then, constitute the voice frame (VoIP packet payload). The Codec digitizes analog voice data after analog voice is captured for a certain period. The longer the period the bigger the voice frame size and the more the delay. The shorter the period the smaller the voice frame size and the less the delay. Considering this, and because VoIP is highly delay-sensitive, the Codec reduces the analog voice digitization period to avoid introducing an inadmissible delay. Therefore, the Codec produces small voice frames size, typically, 10 to 30 bytes based on the used codec. Table I shows some of the common VoIP Codecs [9,10,11]. As for VoIP protocols, there are two types: signaling protocols and media transfer protocols [12,13]. H.323 and Session Initiation Protocol (SIP) are the two common signaling protocols [13,14]. On the other hand, the 12-bytes Real-time Transport Protocol (RTP), 6-bytes Internet Telephony Transport Protocol (ITTP), and 4-bytes Inter-Asterisk eXchange (IAX) are the main examples of media transfer protocols [9,13,15]. Both RTP and IAX take the help of the 8-bytes User Datagram Protocol (UDP) to be able to convey the voice data, while ITTP is able to carry the voice data by itself [9,13,15]. As mentioned earlier, adding these protocols along with the 20-bytes IP protocol to the small VoIP packet payload leads to a considerable amount of the wasted bandwidth [7,8]. The wasted bandwidth is calculated by dividing the protocol size over packet size (payload +protocol). For example, the wasted bandwidth caused when using the ITTP protocol, which is our concern in this article, is up to 72.2 % with 10 bytes codec frame size [16,17]. Fig. 1 shows the typical ITTP packet format.

TABLE I. COMMON VOIP CODES

Code Name	Frame Size (B)	Bit Rate (kbps)
G.723.1	20	5.3
G.726	30	24
LPC	14	5.6
G.729	10	8
G.728	10	16

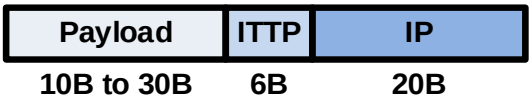


Fig. 1. ITTP Packet Format.

VoIP packets grouping is one of the key techniques to handle this problem and improve the efficiency of utilizing the network bandwidth by VoIP service [3, 18]. VoIP packets grouping techniques group several VoIP packets in one preamble, thus, saving the bandwidth. The saved bandwidth depends on the number of grouped packets in one preamble [3, 18]. This article will suggest a new VoIP packets grouping technique that groups a multiplicity of VoIP packets that travel over the same path to the same destination. Unlike the existing techniques, the suggested technique performs the grouping process at the ITTP layer, not the IP layer, which achieves better bandwidth saving.

The remainder of this article is arranged as follows: Section 2 highlights some of the key VoIP packet grouping techniques. Section 3 explains in detail the technique. It explains the location at which the suggested technique produces its best performance, the main components of the suggested method, and the internal process of the suggested method. Section 4 investigates the performance of the suggested method compared with the conventional ITTP protocol (without grouping). Finally, section 5 concludes the article.

II. Related Work

The VoIP service media carriers protocols, in particular ITTP, groups with IP, introduce extraordinary preamble size to the VoIP data and, thus, expends the VoIP network bandwidth. VoIP packet grouping is one of the primary techniques that are used to lessen the preamble size resulting from ITTP/IP. In this section, we will highlight some of the key VoIP packet grouping techniques.

In 2012 [19], Azevêdo Filho PH et al have suggested a packet grouping technique, called Holding Time Aggregation (HTA), that works over ad hoc wireless networks. The main objective of the HTA technique is to reduce the number of transmissions and transmission overhead, by adaptively grouping several VoIP packets in one preamble. HTA technique keeps the VoIP packets for a certain time to group them. The time chosen by the HTA technique changing adaptively based on the current network conditions, in order to keep the VoIP service quality with the acceptable score. The suggested HTA technique was estimated in comparison to several other techniques, in which it outperformed all the comparable techniques. Whereas, HTA proved that it is able to reduce the number of resources needed to carry the generated traffic, both in terms of packet transmissions and the volume of bytes carried over the network. The implementation of the HTA technique has proven that it is able to maintain jitter and total delay within tolerable limits of VoIP service. In addition, the HTA technique achieves a substantial reduction in the number of packet transmissions as well as the overall packet overhead. The savings in terms of packet transmissions can be as high as 80% in the evaluated scenarios.

Another VoIP grouping technique was suggested by Nomura, Yoshihide, et al in 2014 [18]. The suggested packet grouping technique works in IEEE802.11ac wireless network. The suggested technique grouped several packets directing to the same mobile terminal into a single IEEE802.11ac wireless frame. In order to achieve high efficacy, the suggested technique takes into account the multi-rate transmissions and VoIP service. In addition, it gives the priority of transmission to the mobile terminal that has buffered VoIP packets beyond the tolerable VoIP delay. Moreover, the space channel time is reduced by regulate the transmission times of wireless frames through adjusting the number of grouped packets considering their modulation and coding scheme (MCS) level. The suggested scheme is evaluated by computer simulation to demonstrate its effectiveness under the WLANs with multi-rate transmission providing VoIP services. The results of computer simulations show that the suggested scheme improves the system throughput, the space channel time ratio during MU-MIMO transmissions, and maximum delay time for the VoIP packets in the WLANs with multi-rate transmissions. Consequently, the suggested packet grouping scheme is effective for the downlink MU-MIMO channels under recent WLAN environments.

More recently, in 2019 [20], Abualhaj M. et al have suggested a grouping technique, named compression-aggregation over ITTP (CA-ITTP), that works over the ITTP protocol. As the name suggests, the CA-ITTP aims to save the bandwidth by i) compressing the VoIP packet payload and ii) grouping several ITTP protocol VoIP packets in one preamble. The VoIP packet payload compression is achieved by finding and transmitting the difference between the successive VoIP packets payload. The ITTP protocol VoIP packets grouping is achieved grouping several VoIP packets at the IP layer. In other words, several VoIP packet payload and ITTP preamble of different packets are grouped in one preamble. In order to reconstruct the original packet at the receiver side, a mini-preamble is attached to each VoIP packet payload and ITTP preamble at the sender side. The performance evaluation, of the suggested CA-ITTP technique in comparison to the traditional ITTP protocol, has shown considerable improvement of the bandwidth usage. Whereas, the number of concurrent calls that run over the same network bandwidth is almost doubled, when using the suggested CA-ITTP technique in comparison to the traditional ITTP protocol.

One of the most recent grouping techniques was suggested by Abualhaj M. et al in 2020 [21]. Similar to the previous technique (CA-ITTP), the suggested grouping technique works over the ITTP protocol. The name of the suggested technique is payload shrinking and packets coalesce (PS-PC). The main objective of PS-PC technique is to lessen the wasted bandwidth resulted from the VoIP service. In order to do that, the PS-PC technique groups several VoIP packets in one IP preamble while adding a new runt preamble to each grouped VoIP packet. The runt preamble helps to extract and construct the original VoIP packet. Besides grouping the packets, the PS-PC technique compresses the VoIP packet payload based on a new novel algorithm. Together, packet grouping and payload compression have achieved a very high bandwidth saving. The PS-PC technique has been implemented and compared to the traditional technique (with no grouping neither compressing). The empirical results have shown that the suggested PS-PC technique outperforms the traditional technique in terms of preamble overhead and consumed bandwidth. Whereas, the preamble overhead and the consumed bandwidth have reduced by up to 25% and 51%, respectively, in the tested cases.

As we can see, VoIP packets grouping achieves a noticeable improvement of bandwidth saving. In this article, we suggest a new grouping technique that works with ITTP protocol. Unlike the previous techniques, the grouping technique that works over ITTP, the suggested technique performs the grouping process at the ITTP layer, not the IP layer, which achieves better bandwidth saving. The suggested technique is called ITTP Packets Grouping (ITTP-PG). The following section will discuss the suggested ITTP-PG technique in detail.

III. Suggested ITTP-PG Technique

This section elaborates on the suggested ITTP-PG technique in detail. The ITTP-PG technique has the necessary components and performs the necessary steps to successfully fulfill better bandwidth improvements. The core of the suggested ITTP-PG technique is based on a key approach of improving the VoIP bandwidth utilization; which is reducing the preamble overhead [18,20]. As mentioned earlier, the suggested ITTP-PG technique intended to do so by group a plurality of VoIP packets in one preamble. Fig. 2 shows a VoIP network topology, in which the suggested ITTP-PG technique may be implemented and gives the maximal performance. The VoIP network topology in Fig. 2 assumes a company with several branches: branch A, branch B, branch C, branch D, etc. (e.g., a bank), with plenty of VoIP users in each branch. The VoIP user at each branch may make a call to any other branch at any time. The more the concurrent calls between any two branches (e.g., Branch A and Branch B) the better the performance of the suggested ITTP-PG technique, as we will discuss later in this section. The ITTP-PG technique is made up of two main entities. The first entity performs packets grouping (Pkt-G) and resides at the sender gateway. The second entity performs packets de-grouping (Pkt-DG) and resides at the receiver gateway. The following sections discuss the Pkt-G entity and Pkt-DG entity in detail.

A. ITTP-PG Technique: Pkt-G Entity

The Pkt-G entity resides at the sender side gateway. The main task of the Pkt-G entity is to group the VoIP packets from different client conversation sources and group them in one ITTP/IP preamble. The grouped packet must be intended to go to the same destination VoIP gateway. The process of packet grouping goes through several steps at the Pkt-G entity. The packets from different conversion sources are gathered in one buffer. Then, the packets are distributed to a different buffer based on their destination gateways. After that, the voice payload of the buffered packets is extracted and a runt-preamble is attached to each voice payload to form a run-packet. The runt-preamble will be discussed later in this section. Fig. 3 shows the format of the runt-packet. Subsequently, the resulting run-packets from each buffer, separately, are grouped in one ITT/IP preamble, which constitutes a giant-packet. Fig. 4 shows the format of the giant-packet. Finally, the giant-packets are transmitted to their destination gateways. Fig. 5 shows the Pkt-G entity at the sender gateway. Fig. 6 shows a flowchart of the internal process of the Pkt-G entity in the sender gateway.

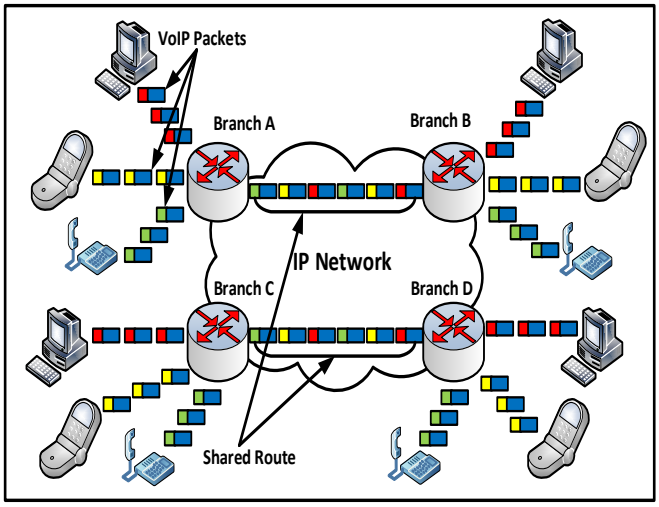


Fig. 2. ITTP-Mux Network Topology Scenario.

Fig. 3. Runt-Packet Format.

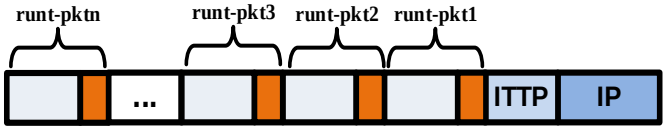


Fig. 4. Giant-Packet Format.

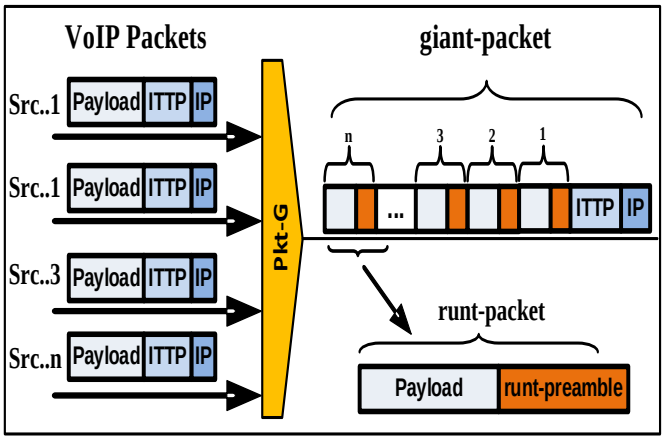


Fig. 5. ITTP-PG Technique: Pkt-G Entity.

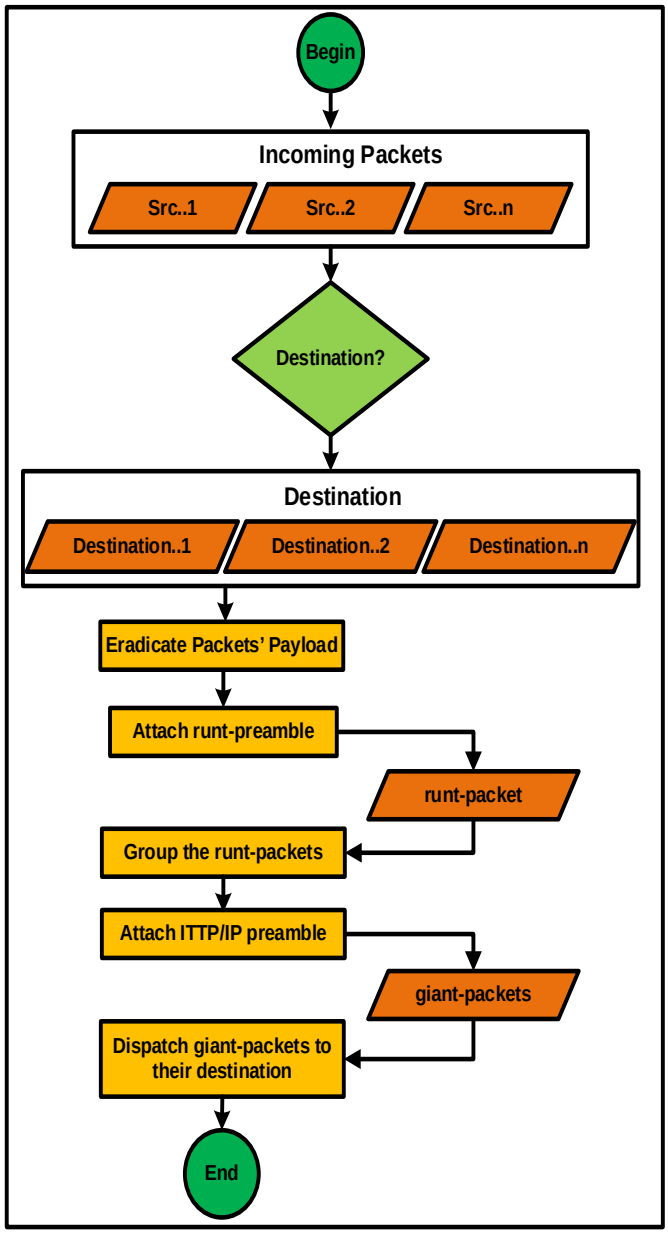


Fig. 6. Pkt-G Entity Operations.

B. ITTP-PG Technique: Pkt-DG Entity

The Pkt-DG entity resides at the receiver side gateway. The main task of the Pkt-DG entity is to de-group the giant-packets and restore the original VoIP packets. The process of packet de-grouping goes through several steps at the Pkt-DG entity. The incoming giant-packets are gathered in a single buffer by the Pkt-DG entity at the receiver gateway. Then, each giant-packet is de-grouped into runt-packets by inspecting the runt-preamble. After that, the voice frame is extracted by removing the runt-preamble from the runt-packet. Subsequently, the original packets are produced by adding ITTP/IP preamble to each voice frame, based on the information in the runt-preamble. Finally, the original packets are sent to their destinations. Fig. 7 shows the Pkt-DG entity in the receiver gateway. Fig. 8 shows a flowchart of the internal process of the Pkt-DG entity in the receiver gateway.

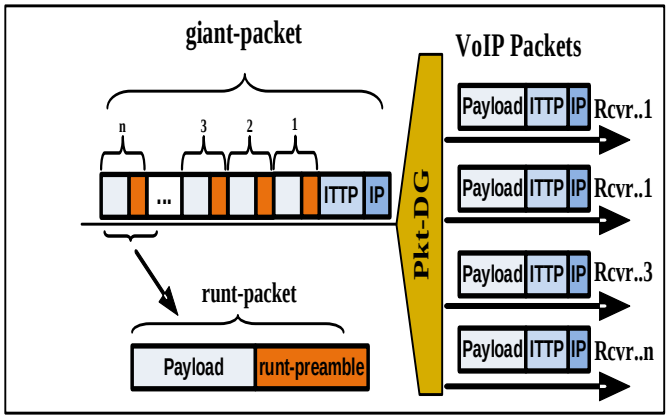


Fig. 7. ITTP-PG Technique: Pkt-DG Entity.

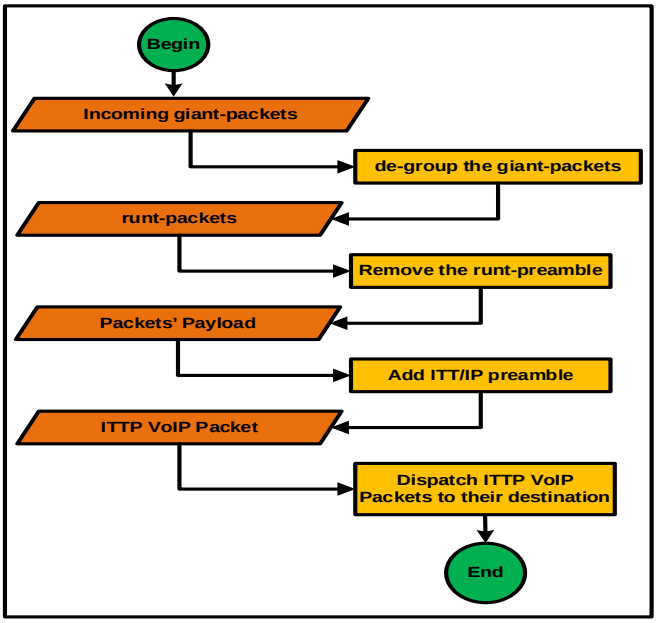


Fig. 8. Pkt-G Entity Operations.

C. ITTP-PG Technique: Runt-Preamble

The key goal of the suggested ITTP-PG technique is to eliminate the ITTP/IP preamble of each packet and group the voice frames of the packets directing to the same destination gateway into a single giant-packet of one ITTP/IP preamble. Before grouping these voice frames, a 3-byte runt-preamble is attached to each voice frame instead of the ITTP/IP preamble, which creates the run-packet. The Pkt-DG entity at the receiver side gateway uses this runt-preamble to separate the runt-packets within the giant-packet. The runt-preamble is composed of two fields namely a 1-byte conversation ID (CID) and 2-byte timestamp. Fig. 9 shows the runt-preamble format.

The timestamp is a key field in the ITTP protocol. The timestamp represents the number of milliseconds since the first data packet transmission of the call. The same value of the timestamp field in the ITTP protocol is saved in the timestamp field in the runt-preamble. The timestamp value is copied back to the original ITTP preamble at the receiver side by the Pkt-DG entity.

Fig. 9. Runt-Preamble Format.

The key use of the CID field is to find the destination address of the runt-packet, before being manipulated and grouped at the sender side. While building the giant-packet, the Pkt-G entity at the sender side gateway generates a unique CID for each runt-packet within a giant-packet. For each CID, a record is saved in a special table, called the address table, at the Pkt-G entity at the sender side. Each record in the address table contains the CID of the runt-packet and the original address (IP: Port Number) of the runt-packet before being grouped. A copy of the content of the address table is transmitted periodically to the Pkt-DG entity at the receiver side. Table II shows the address table. The size of the CID field can vary depend on the number of grouped runt-packets inside a giant-packet. Clearly, a 1-byte CID is enough to group 256 runt-packets inside one giant-packet. Obviously, replacing the 26-byte ITT/IP preamble with the 3-byte runt-preamble will reduce the preamble overhead problem and saves the network bandwidth.

TABLE II. ADDRESS TABLE

Pkt-G entity- Sender Side			**Pkt-DG entity- Receiver Side**	
CID	***IP address:port number***		***CID***	***IP address:port number***
92	172.168.0.8:5050		92	172.168.0.8:5050
20	172. 168.0.9:5051		20	172. 168.0.9:5051
7	172. 168.0.10:5052		7	172. 168.0.10:5052
15	172. 168.0.11:5055		15	172. 168.0.11:5055

IV. ITTP-PG TECHNIQUE EVALUATION

This section assesses the suggested ITTP-PG technique against the conventional ITTP protocol (without grouping) technique, which is called, for simplicity, C-ITTP. The effectiveness of the ITTP-PG technique against the C-ITTP technique is estimated in terms of bandwidth consumption, based on three primary elements. The first element is the number of concurrent VoIP calls (Call-N). Fig. 10 presents the Call-N of the ITTP-PG technique and the C-ITTP technique. The Call-N when using the ITTP-PG technique is greater than the Call-N when using C-ITTP. In addition, the difference in the Call-N increases when the available bandwidth increases.

The second element is the preamble overhead. The preamble overhead ratio is calculated by dividing the packet preamble size by the whole packet size. Fig. 11 presents the preamble overhead ratio of the ITTP-PG technique against the C-ITTP technique. The ITTP-PG technique achieved a significant reduction in preamble overhead against the C-ITTP technique.

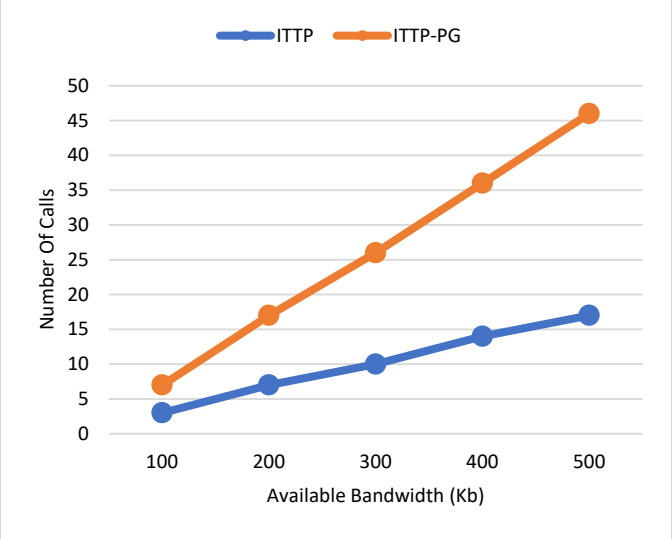


Fig. 10. Call-N of C-ITTP and ITTP-PG Techniques.

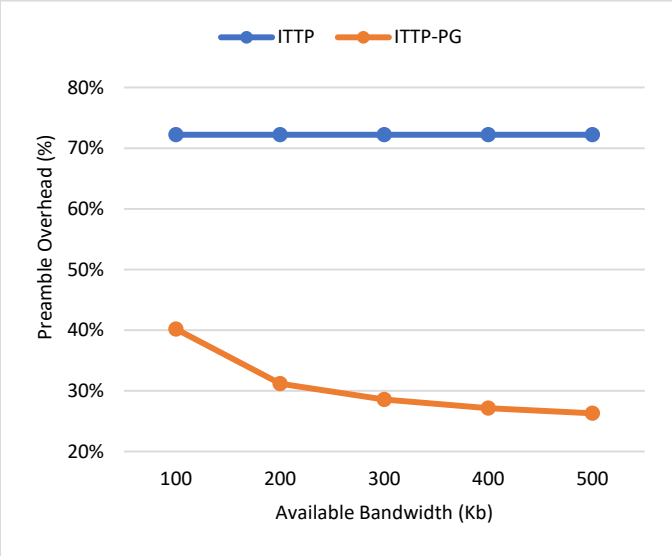


Fig. 11. Preamble Overhead Ratio.

The third element is the bandwidth usage. The bandwidth usage ratio is calculated by dividing the packet payload size by the whole packet size. Fig. 12 presents the bandwidth usage ratio of the ITTP-PG technique against the C-ITTP technique. The ITTP-PG technique achieved a significant enhancement in bandwidth usage against the C-ITTP technique. The enhancement of these three elements is because the C-ITTP technique attaches 26 bytes of a preambleto each payload while the ITTP-PG technique attaches only 3 bytes of a preamble to each payload and 20 bytes of IP preamble to the entire giant-packet.

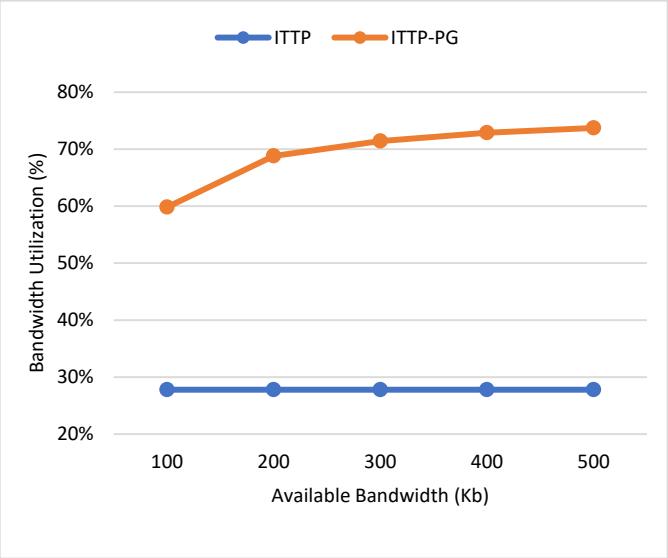


Fig. 12. Bandwidth usage Ratio.

V. Conclusions

VoIP appeared in the last decade as a new service in the telecommunications industry. VoIP service applications convey VoIP packets in short sizes, which increases preamble overhead of the packets and results in inefficient bandwidth employments. In this article, we suggested an efficient grouping technique called ITTP-PG. The ITTP-PG technique made up of the Pkt-G entity resides at the sender gateway and of Pkt-DG entity resides at the receiver gateway entities. The Pkt-G entity groups the VoIP packets in one ITTP/IP preamble. The Pkt-DG entity de-groups the giant-packets and restores the original VoIP packets. The performance of the ITTP-PG technique was assessed based on the number of concurrent VoIP calls, preamble overhead, and bandwidth usage. Based on these three elements, the ITTP-PG technique outperformed the C-ITTP technique. The three elements reflect bandwidth employments efficiency. In general, the simulation result showed that ITTP-PG improves bandwidth usage by up to 45.9%, in comparison to the C-ITTP technique. In future, the proposed method will be implemented in real environment scenarios and investigated with the other comparable methods.

References

[1] Gupta and A. Chaudhary, "A metaheuristic method to hide MP3 sound in JPEG image," Neural Computing and Applications, vol. 30(5), pp. 1611-1618, 2018.

[2] Q. Shambour, S. N. Alkhatib, M. M. Abualhaj, and Y. Alrabanah, "Effective voice frame shrinking method to enhance VoIP bandwidth exploitation," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11(7), pp. 313-319, 2020.

[3] M. M. Abualhaj, Q. Y. Shambour, and A. H. Hussein, "Effective packet multiplexing method to improve bandwidth utilisation," International Journal of Computer Applications in Technology, vol. 63(4), pp. 327-336, 2020.

[4] P. Christian https://blog.telegeography.com/voice-traffics-slump-continued-in-a-big-way-last-year, 2019.

[5] M. M. Abualhaj, S. N. Al-Khatib, M. Kolhar, A. Munther, and Y. Alraba'nah, "Effective voice frame pruning method to increase VoIP call capacity," TEM Journal, vol. 9(1), pp. 48-54, 2020.

[6] C.Olariu, J.Fitzpatrick, Y. Ghamri-Doudane, and L. Murphy, "A delay-aware packet prioritisation mechanism for voice over ip in wireless mesh networks," In 2016 IEEE wireless communications and networking conference, pp. 1-7, 2016.

[7] C.Vulkan, A. Rakos, Z. Vincze, and A. Drozdy, "Reducing overhead on voice traffic," U.S. Patent No. 8,824,304. Washington, DC: U.S. Patent and Trademark Office, 2014.

[8] P. Fortuna and M. Ricardo, "Header compressed VoIP in IEEE 802.11," IEEE Wireless communications, vol. 16(3), pp. 69-75, 2009.

[9] M. M. Abualhaj, M. M. Al-Tahrawi, and S. N. Al-Khatib, "A new method to improve voice over ip (VoIP) bandwidth utilization over internet telephony transport protocol (ittp)," Proceedings of the 2019 8th International Conference on Software and Information Engineering, pp. 192-195, 2019.

[10] E. Charfi, L. Chaari, and L. Kamoun, "Joint urgency delay scheduler and adaptive aggregation technique in IEEE 802.11 n networks," 2015 5th International Conference on Communications and Networking (COMNET), IEEE, pp. 1-6, 2015.

[11] H. Sathu and A. S. Mohib, "Performance comparison of VoIP codecs on multiple operating systems using IPv4 and IPv6," International Journal of e-Education, e-Business, e-Management and e-Learning, vol 2(2), pp. 122-125, 2012.

[12] S. Karapantazis and F.N. Pavlidou, "VoIP: a comprehensive survey on a promising technology," Computer Networks, vol. 53(12), pp. 2050-2090, 2009.

[13] M. M. Abualhaj, M. M. Al-Tahrawi, and S. N. Al-Khatib, "Performance evaluation of voip systems in cloud computing," Journal of Engineering Science and Technology, vol. 14(3), pp. 1398-1405, 2019.

[14] R. Safoine, S. Mounir, and A. Farchi, "Comparative study on DOS attacks detection techniques in SIP-based VoIP networks," 2018 6th International Conference on Multimedia Computing and Systems (ICMCS), IEEE, pp. 1-5, 2018.

[15] M. Spencer, B.Capouch, E.Guy, F.Miller, and K. Shumard, Iax: Inter-asterisk exchange version 2, Internet Request for Comments, 2010.

[16] N. Gupta, N. Kumar, and H. Kumar, "Comparative analysis of voice codecs over different environment scenarios in VoIP," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, pp. 540-544, 2018.

[17] A. K. Chilli, K. R. Prasanna Kumar, H. A. Murthy, and C. C. Sekhar, "Approaches to codec independent speaker identification in VoIP speech," 2018 Twenty Fourth National Conference on Communications (NCC), IEEE, pp. 1-5, 2018.

[18] Y. Nomura, K. Mori, K. Naito, and H. Kobayashi, "High efficient packet aggregation scheme for multi-rate and VoIP packet transmissions in next generation MU-MIMO WLANs," 2014 International Conference on Advanced Technologies for Communications (ATC 2014), IEEE, pp. 517-521, 2014.

[19] P. H. Azevêdo Filho, M. F. Caetano, and J. L. Bordim, "A packet aggregation mechanism for real time applications over wireless networks," International Journal of Networking and Computing, Vol 2(1), pp. 18-40, 2012.

[20] M. M. Abualhaj, "CA-ITTP: An efficient method to aggregate voip packets over ittp protocol," International Journal of Innovative Computing, Information and Control (IJICIC), vol. 15(3), pp. 1067-1077, 2019.

[21] M. M. Abualhaj, S. N. Al-Khatib, and Q. Y. Shambour, "PS-PC: An Effective Method to Improve VoIP Technology Bandwidth Utilization over ITTP Protocol," Cybernetics and Information Technologies, vol. 20(3), pp. 147-158, 2020.

Improving Intelligent Personality Prediction using Myers-Briggs Type Indicator and Random Forest Classifier

Nur Haziqah Zainal Abidin[1], Muhammad Akmal Remli[2], Noorlin Mohd Ali[3]
Danakorn Nincarean Eh Phon[4], Nooraini Yusoff[5], Hasyiya Karimah Adli[6], Abdelsalam H Busalim[7]
Faculty of Computing, Universiti Malaysia Pahang, 26600, Pekan, Pahang, Malaysia[1, 3, 4]
Institute for Artificial Intelligence and Big Data, Universiti Malaysia Kelantan, City Campus
Pengkalan Chepa 16100, Kota Bharu, Kelantan, Malaysia[2, 5, 6]
The Irish Institute of Digital Business, DCU Business School
Dublin City University, Dublin, Ireland[7]

***Abstract*—The term "personality" can be defined as the mixture of features and qualities that built an individual's distinctive characters, including thinking, feeling and behaviour. Nowadays, it is hard to select the right employees due to the vast pool of candidates. Traditionally, a company will arrange interview sessions with prospective candidates to know their personalities. However, this procedure sometimes demands extra time because the total number of interviewers is lesser than the total number of job seekers. Since technology has evolved rapidly, personality computing has become a popular research field that provides personalisation to users. Currently, researchers have utilised social media data for auto-predicting personality. However, it is complex to mine the social media data as they are noisy, come in various formats and lengths. This paper proposes a machine learning technique using Random Forest classifier to automatically predict people's personality based on Myers–Briggs Type Indicator® (MBTI). Researchers compared the performance of the proposed method in this study with other popular machine learning algorithms. Experimental evaluation demonstrates that Random Forest classifier performs better than the different three machine learning algorithms in terms of accuracy, thus capable in assisting employers in identifying personality types for selecting suitable candidates.**

***Keywords*—Machine learning; random forest; Myers–Briggs Type Indicator® (MBTI); personality prediction; random forest classifier; social media; Twitter user**

I. INTRODUCTION

Machine learning is a well-known technique that is broadly utilised by researchers for personality prediction. Due to the advantages of machine learning in learning historical data and making a prediction on future data, the researcher can also use it for learning personality patterns [1]. Such an application is also well-known in psychological science as an assessment tool to predict personality. Nowadays, businesses and recruiters are investing in personality prediction technologies that utilise the machine learning technique. By developing a machine learning algorithm, selecting the best candidates can be achieved, and an error occurred due to the manual analysis process can be reduced.

Motivational influences and human behaviour are the best predictors in personality that will predict an individual's work performance. People's experiences which are emotionally significant with situations, can also be influenced by personality. This approach reflects a person's character and can identify using the Myers-Briggs Type Indicator (MBTI). Based on [2], they defined the personality of a person as a set of attributes that describes a likelihood on the uniqueness of behaviour, feeling and thoughts of the person. These attributes of a person change through time and positions. In a simpler term, we can regard personality as a mixture of characteristics and standards that built an individual's unique character. There are many different personality models used to characterise personality such as the Big Five model (Five-factor model) [3], Myers-Briggs Type Indicator (MBTI) [4], and Theory of Personality Types Carl Jung [5]. Among these personality models, the Big Five and MBTI models are currently popular among researchers. Compared to other models, MBTI is more robust as it has broader applications in different disciplines, although it suffers some issues in terms of reliability and validity. In this study, we select the MBTI personality model due to its popularity and potential to be utilised in different fields.

People nowadays deliver their thoughts and emotions through social media platforms [6]. The posts can be in so many ways, such as using an image, URL link, and music. People's personality also can be examined using social media. The personality of people shown to be useful in predicting job satisfaction, professional and romantic relationship success. In the process of selecting the right candidates, companies nowadays tend to examine the candidates' social media profiles to know the personality of the candidates for a particular job [7]. They intend to reduce the time spent in the preliminary phases of recruitment which is typically known as social media mining. In this paper, we used Twitter as it is one of the most popular social media platforms used nowadays.

It is not easy for employers to select the best candidates for their companies [8][9]. Furthermore, the traditional procedure usually requires employers to spent time conducting interviews with all shortlisted candidates. With the rapid

development of the internet, some researchers have been developing personality prediction system based on the candidates' social media postings to identify candidates' personality for employers [7] accurately.

Over the past few years, many studies use various machine learning algorithm for predicting personality types. One of the earliest studies on Predicting Personality System from Facebook users was developed in 2017 by Tandera [6]. The goal of this study was to build a prediction system that can automatically predict the users' personality based on their activities on Facebook [6]. They also analysed the accuracy of traditional machine learning and deep learning algorithm on predicting personality by implementing Big Five Personality models. Also, in 2018, Giulio Carducci conducted a study on Computing Personality Traits from Tweets Using Word Embedding and Supervised Learning [10]. [10]. The researcher used a supervised learning approach to compute personality traits from an individual's historical tweets. They developed three machine learning algorithms, namely Support Vector Machine (SVM), LASSO and Logistic Regression to predict Big Five Personality model. Mohammad Hossein Amirhosseini conducted a study on Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator on 2020 [11]. The study developed a new machine learning method for automating the process of meta programme detection and personality type prediction based on the MBTI personality type indicator [11]. The natural language processing toolkit (NLTK) and XGBoost, which is based on Gradient Boosting library in Python is used for implementing machine learning algorithms.

In this paper, an intelligent personality prediction system is proposed to predict the personality of candidates based on Twitter data. The proposed method uses machine learning to mine user characteristics and learn patterns from large amounts of personal behavioural data. This system can automatically evaluate candidates' personality traits by processing various attributes and eliminate time-consuming process required in the conventional approach. The rest of the paper is organised as follows: Section 2 presents related work; Section 3 provides material and methods; Section 4 highlights the result of the experiments, and Section 5 concludes the paper.

II. Related Works

Many different machine learning algorithms have used by researchers in the study of personality prediction. Almost all research in this field involved several stages, including data gathering, pre-processing, extracting features and perform classification to determine the accuracy of the model. This section highlights the Myers–Briggs Type Indicator® (MBTI) and related works from previous researchers using machine learning algorithms.

A. Myers–Briggs Type Indicator® (MBTI)

Study on personality has always been a topic of interest for psychologists and sociologists, and one such experiment was performed by the psychiatrist Carl Jung on "Myers-Briggs type indicator". According to [14] in the 1920s, Isabel Myers-Briggs and Katherine Briggs designed the Myers-Briggs type indicator test based on Carl Jung's psychological types. There are 16 personality types on the Myers-Briggs Type Indicator® instrument, which is called a "type table" as shown in Fig. 1 [15]. As an example, someone labelled as INTP in the MBTI system prefers introversion, intuition, thinking and perceiving personality. Based on the label, we can classify the person's desire or behaviour, and more knowledge can be learned by the machine.

The 16 personality types are combined to indicate the personality preferences in four dimensions. Each dimension represents two personalities. The four dimensions are Extroversion–Introversion (E–I), Sensation–Intuition (S–N), Thinking–Feeling (T–F), and Judgment–Perception (J–P) as shown in Fig. 2 [16].

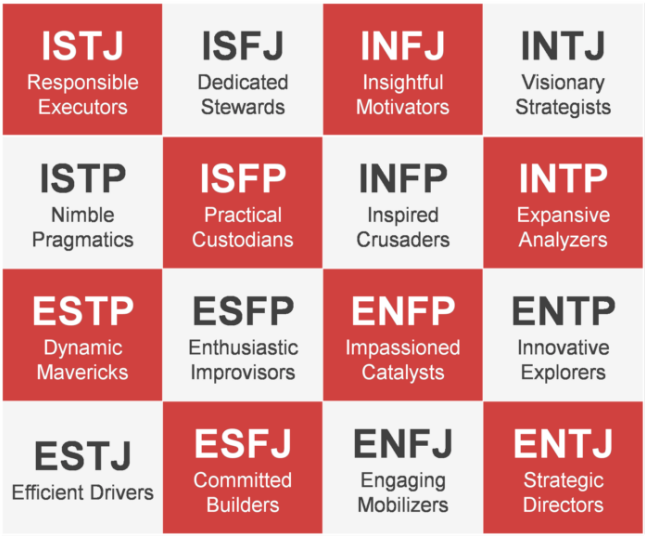


Fig. 1. The Myers Briggs Type Indicator (MBTI).

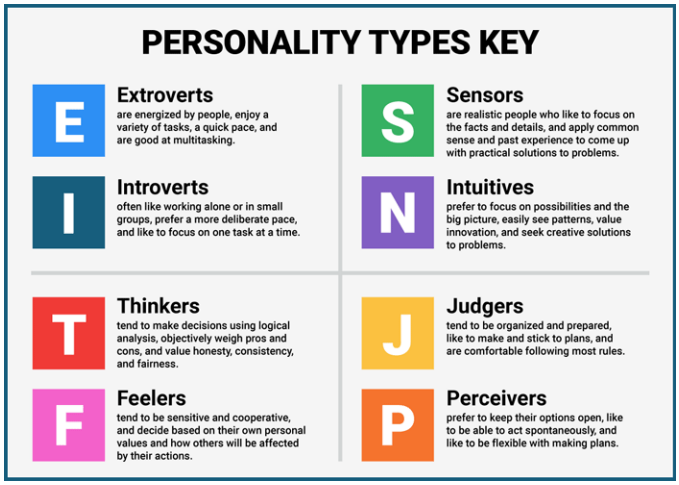


Fig. 2. Key Personality Types.

There are several features related to the various personality types that can be extracted from text or related data. We can use user's posts in social media such as video, image, or other links to analyse their MBTI types using Term Frequency-Inverse Document Frequency (TF-IDF). We can use TF-IDF as a tool to detect and measure the most popular words posted by a person. Beside URL, other potential features can be extracted from text data, including hashtags, emoticons, number of words, ellipses, action words and many more. These extra features also have significant characteristics that could relate to the various personality types. For example, when users of social media are categorised under one of the MBTI, their linguistic contents such as number of words or ellipses will generate extra personality features for the person.

B. Machine Learning (ML)

Machine learning (ML) is a subset of artificial intelligence (AI) that gives frameworks the capacity to naturally take in and improve for a fact without being unequivocally modified [12]. There are three machine learning algorithms which are supervised learning, unsupervised learning and reinforcement learning. The most popular and generally embraced methods are supervised learning and unsupervised learning. Supervised learning is an algorithm that consists of input data (also called training data) and target (or outcome) variable. The input contains a set of features that determine the desired output for the prediction model [13]. Some examples of supervised learning algorithms are Decision Tree, Linear Regression and Logistic Regression. In ML, classification is used to predict the outcome of a given sample when the output variable is in the form of categories. Example of classification algorithms is Naïve Bayes Classifier, Support Vector Machine (SVM) and K-Nearest Neighbour (KNN).

On the other hand, unsupervised learning is an algorithm used for collecting population. This algorithm can describe hidden structures by exploring the unlabeled data. Example of such algorithms is K-Means, Mean Shift and K models. Meanwhile, reinforcement learning is algorithms that continuously train data via trial and error method to make specific decisions. This learning method applies to some cases with trial and error search and delayed reward [13]. In order to decide on the best decision, this method will try to apprehend the best possible knowledge by analysing sample data that had been trained before. Example of reinforcement learning algorithms includes Markov Decision Process and Q Learning.

C. Personality Prediction System from Facebook user

For many years, Facebook has been using Personality Prediction Systems that can predict a user's personality automatically from their Facebook functions [6]. Facebook uses the Big Five Personality model that accurately predicts a user's personality based on someone's personality traits. Several traits can be discovered using this model such as extraversion, conscientiousness, neuroticism, agreeableness and openness. In this study, the researchers used two collections of datasets to predict the users' personality. The first dataset is samples data from the myPersonality project, and the second dataset is data that was generated manually. In the pre-processing stage, the texts written in the English language are corrected before it goes through to the next stage. Pre-processing steps consist of removing URLs, symbols, names, spaces, lowering case, stemming, and removing stop words. For data in Bahasa Melayu language, slang words or non-standard words are manually replaced in a different pre-processing stage before we translated the texts to English.

For the classification process in this study, various series of tests were conducted using deep learning and traditional machine learning algorithms for predicting the personality type of candidates for a particular job position to achieve optimum accuracy. Traditional machine learning algorithms used include Naive Bayes, Support Vector Machine (SVM), Logistic Regression, Gradient Boosting, and Linear Discriminant Analysis (LDA). Meanwhile, deep learning implementations used four architectures, namely Multi-Layer Perceptron (MLP), Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), and 1-Dimensional Convolutional Neural Network (CNN 1D). Results of experiments on traditional machine learning algorithms proved that in myPersonality dataset, the LDA algorithm has the most significant degree of average accuracy. Other than that, the SVM algorithm has the highest average accuracy in a manually gathered dataset (although the difference with other algorithms is not significant).

Meanwhile, the results of experiments on deep learning algorithms proved that MLP architecture has the highest average accuracy in myPersonality dataset and LSTM+CNN 1D architectures have the highest accuracy in a manual gathered dataset. In conclusion, we can improve the accuracy of datasets by using deep learning algorithms, even for traits with relatively low accuracy. This happens because, in this study, only a small number of dataset is used.

D. Personality Traits from Tweets using Word Embedding and Supervised Learning

In this study, we used Twitter as a source to derive personality traits. This social media platform is a rich source of textual data, and users' behaviour, a platform people use to reflect many aspects of life, including personality. People widely share their feelings, moods, and opinions that provide a rich and informative collection of personal data that could be used for a variety of purposes [10]. Other than that, there is a recent work that constructed a questionnaire which is called Big Five Inventory (BFI) personality test for the personality traits. It consists of 44 short phrases with a five-level Likert scale, and can accurately measure the five personality traits plus six underlying facets for each trait. Then, 26 panellists were asked to share their Twitter handles and to answer the questionnaire. The pre-processing stage includes URL removal, mention removal and hashtag removal that consists of textual features created by users. Aside from that, we also removed retweets without additional content. Then, they separately fed each tweet vector to the trained model to obtain a prediction, and average all the values to compute the final personality trait score.

To derive the best performing predictive model, researchers explored different ML algorithms and performances. The ML algorithms are evaluated based on the training set through minimising the mean squared error as their loss function. They also compared the learning model

(SVM) with two baseline algorithms (Linear Regression and LASSO). These baseline algorithms are used in state-of-the-art approaches for personality prediction. The result showed that SVM classifier was able to predict the personality of Twitter users with a certain degree of accuracy, and achieve lower mean squared error. Linear Regression and LASSO models that were trained with lack of discriminative power and tend to predict personality values that are close to the average score in the myPersonality Gold Standard data.

E. Machine Learning Approach to Personality Type Prediction based on MBTI

Myers–Briggs Type Indicator® (MBTI) combines 16 different personality types in four dimensions. These basic dimensions describe the preferences of an individual. The four dimensions which are also known as basic meta-programmers are Extroversion–Introversion (E–I), Sensation–Intuition (S–N), Thinking–Feeling (T–F), and Judgment–Perception (J–P). There are two types of personality for each dimension. This study predicted the personality type of a person based on the MBTI [11]. In the pre-processing stage, they collected data from an Internet forum and removed the MBTI types by using NLTK. Then, we transformed bent forms of words into their root words; a process is known as text lemmatised. Then, we categorised 16 classes of personality types into four binary classes (dimensions). Each of these binary classes represents an aspect of personality according to the MBTI personality model.

After the pre-processing stage, we created the Gradient Boosting Model. In this stage, we split the data into training and testing datasets after the MBTI type indicators were trained individually. We used training data to fit the model while testing data for predictions. Then, they used another existing method which is a recurrent neural network to determine the accuracy of the prediction. Based on the comparison, XGBoost that is based on Gradient Boosting classifier showed better accuracy than the recurrent neural network.

Before we build the new approach, we need to consider the existing systems that have been implemented to ensure that our new approach is better and constructed correctly. We made the comparison based on the personality model and method implemented in the existing systems. Table I shows a comparison of existing approaches.

TABLE I. PREVIOUS STUDIES ON PERSONALITY PREDICTION USING MACHINE LEARNING

Studies	Personality Model	Method
Tandera et al., 2017 [6]	Big Five Personality Model	Traditional machine learning, Deep learning
Carducci et al., 2018 [10]	Big Five Personality Model	SVM Classifier, Linear Regression, LASSO
Amirhosseini and Kazemian, 2020 [11]	Myers–Briggs Type Indicator® (MBTI)	NLTK, XGBoost

Based on the comparison of existing systems, we can improve the new approach to achieve more accurate data and better personality results. Also, increasing size of the dataset could potentially give a more precise prediction. In this research, we used tweets from Twitter social media extracted from Kaggle repository as our dataset.

F. Random Forest Classifier

Random Forest Classifier also was known as ensemble algorithm is a supervised learning algorithm [18] that combines the same or different kind of more than one algorithm for classifying object [17]. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. A random forest combines hundreds or thousands of decision trees, trains each one on a slightly different set of observations, splitting nodes in each tree considering a limited number of features and merging them to get a more accurate and stable prediction [19]. The average prediction of each tree is the final predictions of the random forest [20]. Advantage and disadvantage of Random Forest Classifier include:

Advantages

- It is a robust method that consists of many decision trees, making it highly accurate.
- There is no overfitting problem. It will cancel out biases by taking the average of all predictions.
- The algorithm can be used in both classification and regression problems.
- This algorithm handles missing value in two ways: using median value and computing proximity-weighted average of missing value.

Disadvantages

- This algorithm requires multiple decision trees resulting to slow prediction generation process. This is because the same given input needs to be predicted and voted for all the trees in the forest, making the process time-consuming.
- The prediction model is challenging to interpret because it is hard to make a decision based on the path in the tree compared to the whole decision tree.

III. MATERIALS AND METHODS

This section discusses the method used for developing Intelligent Personality Prediction using Machine Learning. This methodology is used to help in structuring the model development process.

A. Model Development

1) Dataset: We collected the data from Kaggle repository (https://www.kaggle.com/datasnaek/mbti-type). In this study, the dataset contains over 8675 rows of data with two columns, as shown in Fig. 3. In each row, the data held a person's:

- Type: The person's four letters MBTI code/type

- Posts: Each of the last 50 things the person posted on Twitter. Each entry is separated by "|||" (3 pipe character).)).

We collected the dataset in 2017 from users of an online forum,personalitycafe.com (https://www.personalitycafe.com /forums/myers-briggs-forum.49/). We conducted the data collection in two phases. In the first phase, the users answered a set of questionnaire that sorts them based on their MBTI type. In the second phase, users were allowed to chat publicly with other users in the forum. The chatting sessions allowed more personality type data to be generated based on MBTI type.

2) Exploratory data analysis: We conducted exploratory data analysis was to get visual representation for further investigation through a violin plot printing. The number of words per comment was examined to obtain the intuitive idea of sentence structure for each personality, as shown in Fig. 4.

After that, seven additional features were created since there are currently only two features in the dataset, namely Type and Posts. The additional features are as below:

- words per comment,
- ellipsis per comment,
- links per comment,
- music per comment,
- question marks per comment,
- images per comment,
- exclamation marks per comment.

For every feature, the average number of words, punctuation, etc., are calculated. After we added these features, we analysed the Pearson correlation between words per comment and ellipses per comment for overall set of data to see how the raw data looks like and to see how the features distinguish between the four MBTI types as shown in Fig. 5. In this step, we used 'Seaborn' which is a Python data visualisation library and 'Matplotlib' which is a Python 2D plotting library for data visualisation and correlation of the MBTI personality types.

From Fig. 5, we can see that there is a high correlation between words per comment and ellipses per comment. 69% of the words are correlated with an ellipsis. To observe which personality type has the highest correlation, we charted joint plot and pair plot on the correlation variables for the different types of personality in comparison to the words per comments and ellipses per comment as shown in Table II. Fig. 6, 7, 8, and 9 shows the relationship between ellipsis per comment and words per comment for ISTP, ISTJ, ISFP, and ISFJ personality type. Meanwhile, Fig. 10, 11, 12 and 13 represents the relationship between ellipsis per comment and words per comment for INTP, INTJ, INFP, and INFJ personality type. And lastly, Fig. 14, 15, 16, and 17 shows the relationship between ellipsis per comment and words per comment for ENTJ, ENTP, ENFP, and ENFJ personality type.

```
   type                                               posts
0  INFJ  'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1  ENTP  'I'm finding the lack of me in these posts ver...
2  INTP  'Good one  _____   https://www.youtube.com/wat...
3  INTJ  'Dear INTP,   I enjoyed our conversation the o...
4  ENTJ  'You're fired.|||That's another silly misconce...
5  INTJ  '18/37 @.@|||Science  is not perfect. No scien...
6  INFJ  'No, I can't draw on my own nails (haha). Thos...
7  INTJ  'I tend to build up a collection of things on ...
8  INFJ  I'm not sure, that's a good question. The dist...
9  INTP  'https://www.youtube.com/watch?v=w8-egj0y8Qs||...
****************************************
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8675 entries, 0 to 8674
Data columns (total 2 columns):
type     8675 non-null object
posts    8675 non-null object
dtypes: object(2)
memory usage: 135.7+ KB
None
```

Fig. 3. MBTI Personality Type Dataset.

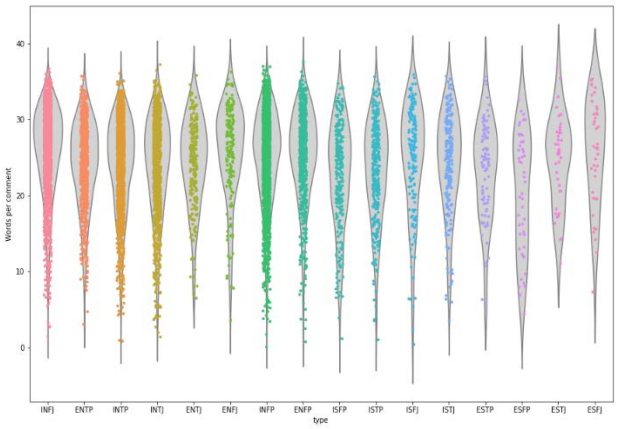


Fig. 4. Words Per Comment for each Personality Type.

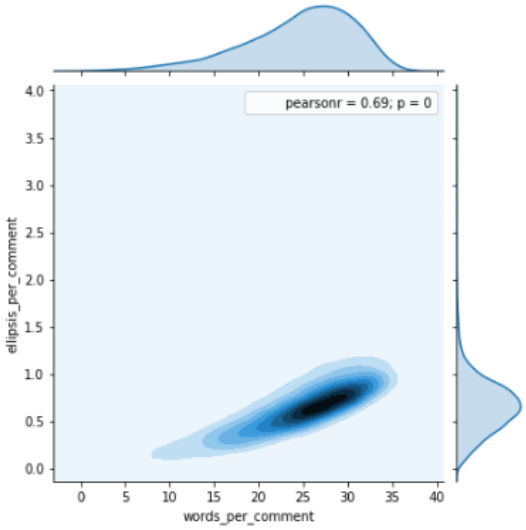


Fig. 5. Pearson Correlation.

Each of the personality comes with the results of Pearson correlation (pearsonr = 0.73). For Fig. 8 and 12, 72% of the words are correlated with an ellipsis. Other than that, for Fig. 10, 13, and 16, 64% of the words are correlated with an ellipsis. Meanwhile, for Fig. 11 and 14, 74% of the words are correlated with an ellipsis.

TABLE II. PEARSON CORRELATION FOR WORDS PER COMMENT VS ELLIPSES PER COMMENT FOR EACH MBTI PERSONALITY TYPE

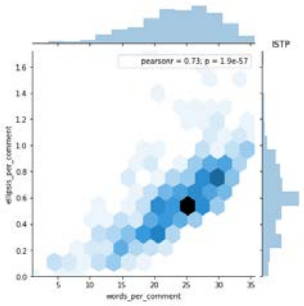


Fig. 6. ISTP Personality Type.

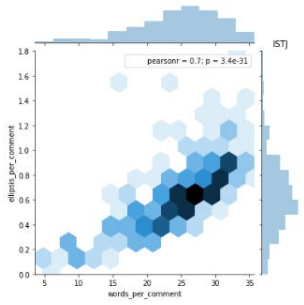


Fig. 7. ISTJ Personality Type.

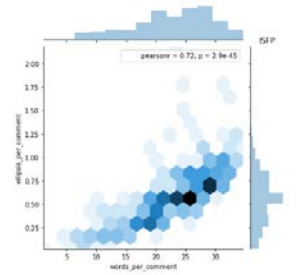


Fig. 8. ISFP Personality Type.

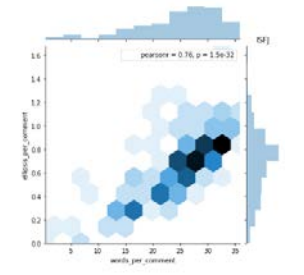


Fig. 9. ISFJ Personality Type.

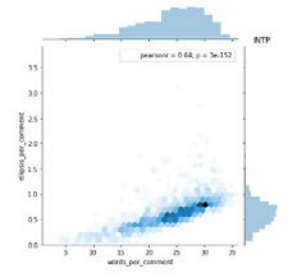


Fig. 10. INTP Personality Type.

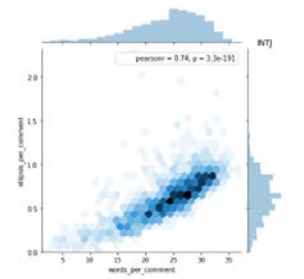


Fig. 11. INTJ Personality Type.

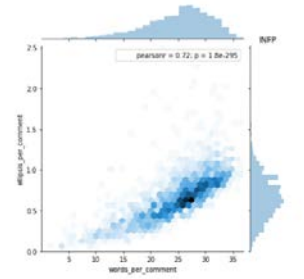


Fig. 12. INFP Personality Type.

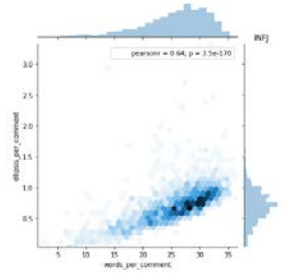


Fig. 13. INFJ Personality Type.

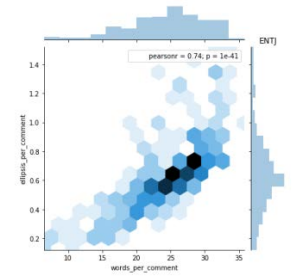


Fig. 14. ENTJ Personality Type.

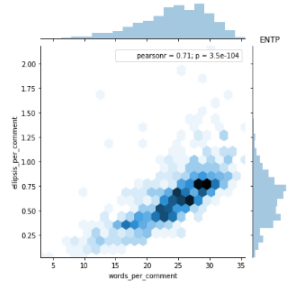


Fig. 15. ENTP Personality Type.

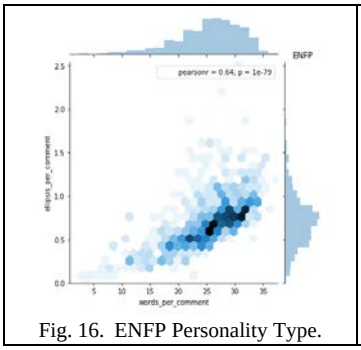


Fig. 16. ENFP Personality Type.

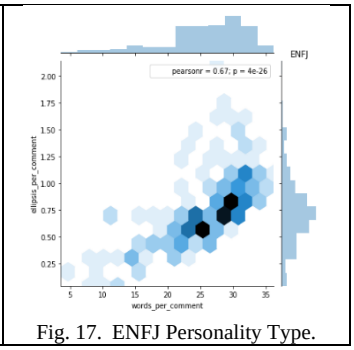


Fig. 17. ENFJ Personality Type.

From Fig. 18, the top three highest correlation values for the ellipses per comment and words per comment are:

- INFJ – The advocate - Introversion Intuition Feeling Judging
- INTP - The Thinker - Introversion Intuition Thinking Perceiving
- ENFP - The Inspirer - Extroverted Intuition Feeling Perceiving

From this exploratory data phase, each MBTI type has a different correlation between ellipses per comment and words per comments. The correlation determines how closely each feature is affected by another feature. INFJ, INTP and ENFP recorded the highest correlation, which is an excellent sign to train the data and build machine learning models.

3) Data pre-processing: To get further insight on the dataset, we created four new columns that divided the respondents based on the four dimensions of MBTI namely Extroversion–Introversion (E–I), Sensation–Intuition (S–N), Thinking–Feeling (T–F), and Judgment–Perception (J–P). The process is to improve the accuracy of the results.

Furthermore, we also used word2vec technique in this pre-processing step. Word2vec is an algorithm to construct vector representations of words, also known as word embedding. In this paper, we converted textual data into numeric signals. For example:

- I = 0, E = 1
- N = 0, S = 1
- T = 0, F = 1
- J = 0, P = 1

4) Dataset splitting: To test the model's accuracy, we split the dataset into two parts which were training dataset and testing dataset. We used 90% of data for training, and 10% for testing and keeping random state five using sci-kit learn's internal module train_test_split (). The testing dataset is a set of unseen data that was used only to access the performance of a fully specified classifier.

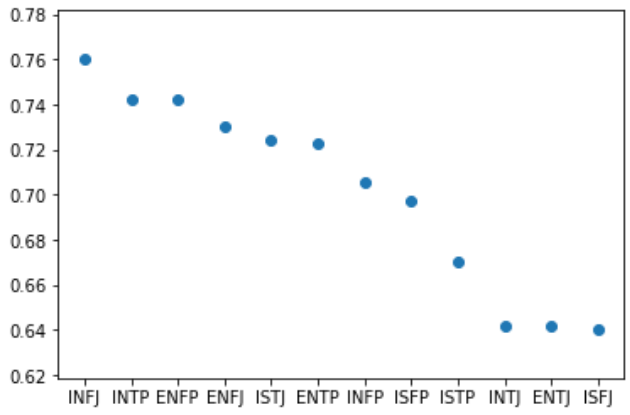


Fig. 18. Pair Plot of Pearson Correlation For words Per Comment vs Ellipses Per Comment.

5) Model building: To build the Random Forest, Logistic Regression, KNN Neighbor and Support Vector Machine (SVM) models, we used Numpy and sklearn. We use train_test_split function from sklearn library to split the data into training and testing datasets while the MBTI type indicators were trained individually. In total, we used 90% of the data for the training set (data fitting), and 10% for testing (making a prediction). We first remove all columns irrelevant to our features. From there, we can see that the Random Forest algorithm was able to classify all respondents (100%) to the right types. Then, we do a deeper dive into our model to get a better perspective on our prediction by performing four machine learning algorithms with Extroversion–Introversion (E–I) column, Sensation–Intuition (S–N) column, Thinking–Feeling (T–F) column, and Judgment–Perception (J–P) column.

6) Comparing the accuracy of machine learning models: In this step, the accuracy of the Random Forest and three other models namely Linear Regression, KNN neighbor and Support Vector Machine (SVM) were evaluated using the testing dataset.

7) Evaluating results: Evaluation of the results helps in finding the best model that represents the data. The result of this evaluation is presented in Section 4.

IV. Results and Discussion

This section discusses the results of the experiment conducted. We did several experiments to obtain the most significant model for predicting MBTI personality types. Firstly, we determine the arrangement of the MBTI personality types by calculating words per comment in the dataset. We added several features in this experiment since the original dataset only has two features. We analysed these features by calculating the average of each feature, namely average words per comment and average ellipses per comment. After that, we measured the Pearson correlation coefficient to know the strength between variables and relationships. Since there is a large correlation (69%) between words per comment and ellipses per comment, we chose this variable to train the machine learning model.

From Pearson correlation conducted, it is evident that INFJ, INTP, and ENFP personality types have the highest correlation between words per comment and ellipses per comment. Next, data pre-processing using word2vec technique was done to make the dataset more organised and easy to understand. Lastly, we use train_test_split function from sklearn library to split the data into training and testing datasets while the MBTI type indicators were trained individually. We used training data to fit the model and testing data for prediction. The last step is we develop four machine learning models, and we obtained the accuracy of each machine learning model for every MBTI personality type.

Table III shows that the Random Forest model has better accuracy (100%) in all four dimensions of MBTI personality types compared to other machine learning models. Accuracy of the Random Forest model is considerably higher than the Support Vector Machine (SVM) model for Intuition (I)–Sensation (S) and Introversion (I)–Extroversion (E) categories, while for Sensation–Intuition (S–N) category, the accuracy is a little bit better. Accuracy of SVM for Judgment (J)–Perception (P) is considerably worse than the Random Forest model. Thus, the overall performance of the Random Forest model is better than the three other machine learning models for this dataset.

TABLE III. Results

Model	E vs I	S vs N	F vs T	J vs P	Overall
Random Forest	100%	100%	100%	100%	100%
Logistic Regression	77.11%	86.03%	63.35%	60.37%	23.35%
KNN neighbor	83.66%	88.11%	77.64%	77.74%	40.62%
Support Vector Machine (SVM)	77.16%	86.03%	56.54%	47.16%	16.94%

V. Conclusions and Future Works

In conclusion, this research could predict personality by using social media data, and the best model of a machine learning algorithm, which are the Random Forest machine learning algorithm. With that, this will significantly benefit companies because they can analyse their candidates' social media accounts before they choose the right employees.

A. Limitations

This research only studied people with particular social media, namely Twitter. The are other social media platforms that could give significant data, thus improve the prediction model. In addition, this research only focused on the prediction of strengths and weaknesses in terms of personality. It is essential to consider a technical position on the team when we are creating teams to fight crime, develop unique software, or play sports. Aside from this, we also need to explore people's soft skills. It is also essential to consider other factors, namely, mindset and personality. In short, this is just the first step in creating a personality type model based on MBTI personality assessment from social media comment data.

This research trains the model based on a large number of tweets, and it is not easy to collect such a massive database for this process. By improvising this research, future research can use a small number of tweets for both training and testing to examine the performance of the method. Finally, this research only used English data. To improve this, we recommend future research to study on multiple social media platforms or different cultures. They can use various data sources to get more insights and exciting finding, using machine learning approaches.

B. Future works

In the future, we plan to collect and build more datasets to get a more accurate result. We also plan to use XGBoost algorithm and deep learning algorithm, their architectures, and other processes to improve this prediction system. XGBoost algorithm which optimised distributed gradient boosting machines is scalable and well-known for its excellent performance in terms of computational speed. This algorithm can push the limits of computing power for boosted trees algorithms. Other than that, deep learning is also a suitable candidate to address this challenge as it can generate new features from a limited series of features located in training datasets. Due to this, the method requires less time to analyse big data.

ACKNOWLEDGEMENT

This work was funded by Universiti Malaysia Pahang [RDU190396] and Universiti Malaysia Kelantan via UMK Fund [R/FUND/A0100/01850A/001/2020/00816].

Conflicts of Interest: The authors declare no conflicts of interest.

REFERENCES

[1] X. Teng and Y. Gong, “Research on Application of Machine Learning in Data Mining,” IOP Conf. Ser. Mater. Sci. Eng., vol. 392, no. 6, 2018.

[2] F. Ahmed, P. Campbell, A. Jaffar, S. Alkobaisi, and J. Campbell, “Learning & Personality Types: A Case Study of a Software Design Course,” J. Inf. Technol. Educ. Innov. Pract., vol. 9, no. January, pp. 237–252, 2010.

[3] B. de Raad and B. Mlačić, “Big Five Factor Model, Theory and Structure,” Int. Encycl. Soc. Behav. Sci. Second Ed., no. December, pp. 559–566, 2015.

[4] T. L. C. Yoong, N. R. Ngatirin, and Z. Zainol, “Personality prediction based on social media using decision tree algorithm,” Pertanika J. Sci. Technol., vol. 25, no. S4, pp. 237–248, 2017.

[5] N. R. Ngatirin, Z. Zainol, and T. L. C. Yoong, “A comparative study of different classifiers for automatic personality prediction,” Proc. - 6th IEEE Int. Conf. Control Syst. Comput. Eng. ICCSCE 2016, pp. 435–440, 2017.

[6] T. Tandera, Hendro, D. Suhartono, R. Wongso, and Y. L. Prasetio, “Personality Prediction System from Facebook Users,” Procedia Comput. Sci., vol. 116, pp. 604–611, 2017.

[7] W. Rc, Y. Munas, K. Cs, F. Ta, and Vithana N, “Personality Based E-Recruitment System,” Int. J. Innov. Res. Comput. Commun. Eng., vol. 5, 2017.

[8] S. Sharma, “Predicting Employability from User Personality using Ensemble Modelling,” no. July, p. 37, 2015.

[9] B. Y. Pratama and R. Sarno, “Personality classification based on Twitter text using Naive Bayes, KNN and SVM,” Proc. 2015 Int. Conf. Data Softw. Eng. ICODSE 2015, no. November, pp. 170–174, 2016.

[10] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio, “TwitPersonality: Computing personality traits from tweets using word embeddings and supervised learning,” Inf., vol. 9, no. 5, pp. 1–20, 2018.

[11] M. H. Amirhosseini and H. Kazemian, “Machine Learning Approach to Personality Type Prediction Based on the Myers – Briggs Type Indicator ®,” 2020.

[12] N. Dhanda, S. S. Datta, and M. Dhanda, “Machine Learning Algorithms,” no. June 2016, pp. 210–233, 2019.

[13] G. M., T. A., S. C., and R. P., “Soybean Under Water Deficit: Physiological and Yield Responses,” A Compr. Surv. Int. Soybean Res. - Genet. Physiol. Agron. Nitrogen Relationships, 2013.

[14] T. Varvel, S. G. Adams, and S. J. Pridie, “A study of the effect of the myers-briggs type indicator on team effectiveness,” ASEE Annu. Conf. Proc., pp. 9525–9533, 2003.

[15] Z. Poursafar, N. Rama Devi, and L. L. R. Rodrigues, “Evaluation of Myers-Briggs Personality Traits in Offices and Its Effects on Productivity of Employees: an Empirical Study,” Res. Artic. Int J Cur Res Rev, vol. 7, no. 21, pp. 53–58, 2015.

[16] S. D. Mallari, “Myers-Briggs Type Indicator (MBTI) Personality Profiling and General Weighted Average (GWA) of Nursing Students.,” Online Submiss., no. October, pp. 1–11, 2017.

[17] “Chapter 5: Random Forest Classifier - Machine Learning 101 - Medium.” [Online]. Available: https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1. [Accessed: 02-May-2020].

[18] “Random Forests Classifiers in Python - DataCamp.” [Online]. Available: https://www.datacamp.com/community/tutorials/random-forests-classifier-python#features. [Accessed: 02-May-2020].

[19] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.

[20] “An Implementation and Explanation of the Random Forest in Python.” [Online]. Available: https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76. [Accessed: 02-May-2020].

The Development of Parameter Estimation Method for Chinese Hamster Ovary Model using Black Widow Optimization Algorithm

Nurul Aimi Munirah[1], Muhammad Akmal Remli[2], Noorlin Mohd Ali[3], Hui Wen Nies[4]
Mohd Saberi Mohamad[5], Khairul Nizar Syazwan Wan Salihin Wong[6]
Faculty of Computing, Universiti Malaysia Pahang, 26600, Pekan, Pahang, Malaysia[1, 3]
Institute for Artificial Intelligence and Big Data, Universiti Malaysia Kelantan
City Campus, Pengkalan Chepa, Kota Bharu 16100, Kelantan, Malaysia[2, 5, 6]
Malaysia Japan International Institute of Technology (MJIIT)
Universiti Teknologi Malaysia, Kuala Lumpur 54100, Malaysia[4]

***Abstract*—Chinese Hamster Ovary (CHO) cells are very famous in biological and medical research, especially in the protein production industry. It is because the characteristic of the cells with low chromosome numbers make it suitable for genetic study. However, all the data tends to be noisy and not fit. That is why many parameter estimation methods have been developed since their first introduction to determine the best value for a particular parameter. Metaheuristic parameter estimation is an algorithm framework that is processed using some technique to generate a pattern or graph. It will help the researcher get the fitted graph model, correct data, and estimate the value based on the data's behaviour. This process started with implementing the parameter estimation that can be generated by using the combination of mathematical models and all the data obtained from the researcher's experiments. This way, biomedical research's cell culture can benefit from all this metaheuristic parameter estimation used. A kinetic model can estimate the data obtained from the Chinese Hamster Ovary (CHO) cells. Therefore, this paper proposed a Black Widow Optimisation (BWO) algorithm inspired by the bizarre mating behaviour of a spider as the method to use to solve the problem. The proposed algorithm has been compared with the other three famous algorithms, which are Particle Swarm Optimization (PSO), Differential Evolutionary (DE), and Bees Optimization Algorithm (BOA). The results showed that the proposed algorithm could get better value in terms of the best cost despite taking a long time to use.**

Keywords*—Chinese Hamster Ovary; Black Widow optimization; metaheuristic; parameter estimation; genetic study

I. Introduction

A. Background

Biotechnology, in the most specific word, is a combination of biology with technology. It is a process of modifying biological systems that are found in an organism to make them adaptable in others field. Within this process, bioinformatics will need many research and development process to take place in a laboratory. Industrial biotechnology is also known as white biotechnology because it is a technology that used renewable sources and living cells microorganisms such as bacteria, yeasts, and cells. The researcher uses various techniques to create and improve the collected enzymes [1]. They use the DNA obtained from the experiments to find the enzyme's catalytic capabilities. This technology also makes less waste during manufacturing and production. It is also widely used in industrial sectors like chemicals, pharmaceuticals, and food industries. Fig. 1 shows the development of industrial biotechnology [2]. The most commonly used biotechnology is in the protein production industry under genetic engineering. The process of protein production will need a researcher to manipulate the genes from an organism such as a hamster and make them compatible to use, mainly producing a pharmaceutical product [3].

The kinetic model is one of the tools used in biotechnology. It can represent the biochemistry of cells that is more complex and utterly different from other models. Those models are necessary used to understand and analyse before they can be used industrially. A metaheuristic is an algorithm framework that has been upgraded to a high-level procedure. It is designed to find and generate useful heuristics that may solve the independent problem or an optimisation problem [4]. It will also provide the researcher with smart guidelines or strategies to develop a suitable algorithm. Since Grover introduced the terms in 1986, there are so many types of metaheuristics have been set [4].

Parameter estimation is also knowns as coefficients and sample statistics. It is a technique or process using sample data collected from various fields, such as engineering and biology. Parameter estimation is a descriptive measure for a population. It is also in charge of the response changes related to changing predictors even for one – unit change while other predictors stay constant. It is estimated by using the least square estimations. However, the values are unknown to measure the whole population as it is infeasible. Therefore, the main goal of statistical analysis is to obtain the number of errors parallel with the parameters used. Parameter estimation is usually used to estimate parameters for a selected distribution that is processed using sample data or a number of populations from any potential interest. Also, parameter estimation can determine the best value of parameters by

going through data assimilation or other similar techniques [5]. Parameter estimation also refers to selecting the best deals from data assimilation for a particular parameter in a numerical model format. With parameter estimation in kinetic models, it is necessary to implement an optimisation method for reducing the distance between estimation models and sample data. Two different ways can be used, which are local and global optimisation. Each method rates based on their minimal best value for each experiment. The local optimisation method has many drawbacks as it can easily be stuck in local minima, that is why the metaheuristic algorithm has the most efficient global optimisation introduced. Hence, most objective functions have several local minima [6].

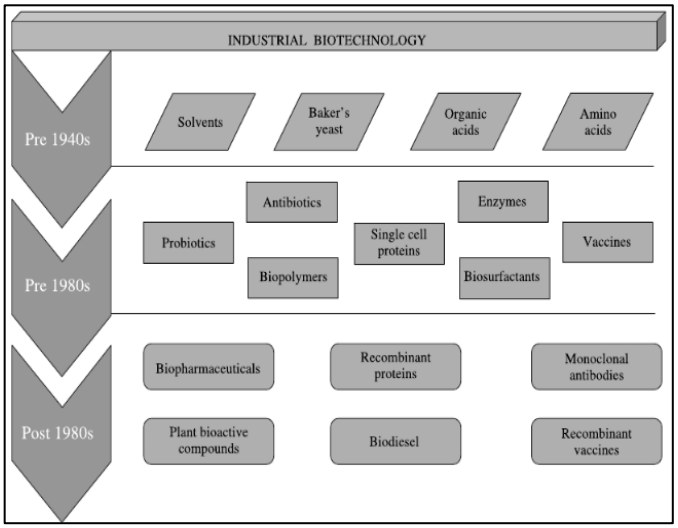


Fig. 1. Development of Industrial Biotechnology.

The metaheuristic parameter estimation for Chinese Hamster Ovary (CHO) research uses a type of metaheuristic to estimate the results of the parameter involved in the CHO cells. This research will develop a kinetic model that is suitable for the sample data to get the required graphs [7]. Nowadays, the researcher uses the benefits of technologies and becomes more appreciated towards software developed by the software developer. They learn how to use the software, and the results are outstanding. Besides, they use MATLAB to generate the graphs to show the right prediction towards the research.

This research focuses on developing the parameter estimation method to test CHO sample data using the Black Widow Optimisation (BWO) algorithm. The measurement used is Root – Mean – Square – Error to train the existing sample data. This research also emphasises on comparing the performance of the proposed algorithm and existing metaheuristic optimisation algorithms.

B. Problem Formulation

Genetic engineering is a complex industry because it will involve a complex number of interacting reactions and regulations [8]. That is one of many main reasons researchers nowadays become more focused on using mathematical models to solve the problems. The estimation used can reduce the scale of error and noise in terms of the distance between the prediction of the model and the data used in the experiments. Many equations represent parameter estimation used to estimate the parameters for each selected sample data such as Probability Plotting, Maximum Likelihood Estimation, and Bayesian Estimation Method. However, this paper will focus on Mean – Square – Error (MSE) as this is denoted how close must regression line be to a set of points. This squaring method is essential to deduct any negative signs. This calculation shows that when the squared error values are getting smaller, the closer the best fit of the line is found. Below is the equation [9]:

$$MSE = [\frac{1}{n}\sum_{i=1}(y_i - \bar{y}_i)^2 \quad (1)$$

Where Σ denoted as the summation of all numbers from (i=1) to n, next, take the y – coordinate and subtract with $\bar{y}_i$ – coordinate value before calculating the square of the results. Last, take the sum of $(y - \bar{y}_i)2$ and start to divide it by n. That is how to get the mean, and it is crucial to minimise the value of the mean so that the best lines can be achieved [9].

Therefore, the new forms of algorithm framework and techniques that can be used to the sample of collected data have been using by modern researchers nowadays. The researcher will get accurate prediction results based on their existing data, especially when implementing the right parameter estimation suitable for their research scopes. It can give them an early preview of how the research will work in the future. The graph that will be generated will ease the researcher to make a prediction even in seconds. Therefore, the nearer the line of the graph with the data dotted by the researcher means the minimisation of time has happened, so the better the research's results will be. Not only that, but the metaheuristic used will also help all the biomedical researchers with the kinetic model generated by using those parameter estimation approaches [10-12]. There were many kinds of optimisation algorithms in this world. Still, all of these algorithms have different competency levels as there are algorithms that are not stable in convergence, stuck at local optima, and difficult to tune the parameters.

Thus, a study on Black Widow Optimization (BWO) Algorithm has been developed to test the CHO sample data, and it is inspired by the mating behaviour of black widow spiders. This algorithm has been proved its ability to maintain the balance between the exploitation process and exploration process, which can provide fast convergence and prevent local optima problems.

II. Metaheuristic Optimisation Algorithm

A metaheuristic is a high-level procedure designed to search, generate, and select a technique developed to solve the problems more quickly and search for a solution that classical methods failed to do. Metaheuristic algorithms can be divided into two types, which are single solution and population-based. Single solution metaheuristic focuses on modifying and improving a solution with a single candidate, while the population-based will maintain and improve multiple solutions for candidates. It will enhance a particular search method to get a better solution.

Metaheuristic algorithms is also an efficient global optimisation method that, for its search process, it can be divided into two other types which are exploitation and explorations. Exploitation is a search process that depends on information that is already obtained from the problem. It is used to develop a better solution and suitable to be used in the local search method. Although it has a very high convergence rate, it is elementary to stuck at local minima [1]. While for explorations, it will explore the search space more broadly, that is why it can find global solutions even though it is far from the initial point. Hence, the process will be much slower in terms of convergence rate and high in computational cost.

The existing metaheuristic algorithms are summarised their advantages and disadvantages in Table I. Hence, these algorithms will be compared according to their advantages and disadvantages throughout the experiment, in order to achieve more best values.

TABLE I. ADVANTAGES AND DISADVANTAGES OF EXISTING METAHEURISTIC ALGORITHMS

Algorithms	Advantages	Disadvantages
Particle swarm optimisation (PSO)	• Can be simple to implement • Have a few parameters to adjust • Able to run the parallel computation. • Can be robust.	• Can be challenging to define initial design parameters. • All solutions converge prematurely
Differential Evaluations (DE)	• Better Explorations • Better diversification. • Easy to use	• Not stable in convergence. • Need tuning in parameters
Bees Optimisation Algorithm (BOA)	• Easy to implement • Able to perform both searches (local and global)	• Need a manual setting of parameters. • Has only a few tuning parameters.

A. Particle Swarm Optimization (PSO)

Particle swarm optimisation (PSO) is one of the well-known metaheuristics ever developed. It is a method to iteratively optimise the problem by improving the solution with a quality measure that has been given. PSO also represents the movement of organisms such as in a bird flock or fish school. Thus, it is meant to simulate those organisms' social behaviour in a way to search for food [13].

A bunch of particles will be used to represent PSO, and it is called a "swarm." Each of the particles is moving based on the influence of its best-known positions. All of those represented particles are freely move around and allowed to search space [14]:

1) Each and own particle will hold its own previous velocity. (Inertia).

2) The distance from its individual particles position. (Cognitive Force).

3) The distance from swarms' position. (Social Force).

Therefore, PSO is a method that can make an assumption or not making any as the problem will be optimised, and the searching can involve a larger space of its candidate solutions. However, PSO will not guarantee in finding the optimal solution.

B. Differential Evaluation (DE)

Differential Evaluation (DE) is an optimisation algorithm that has been introduced in 1996 by Storn and Price. It is a technique used to optimise a problem by improving its candidate solutions depending on the quality measurements [15]. DE will maintain a population and create new solutions by combining them with its formulae. The best core of solutions will be kept. Therefore, the optimisation problem is a black box that is providing when needed. The use of the DE makes the search adaptable during an evolutionary process [15].

At the initial stage, the parents are far from each other, making the perturbations larger. However, when the process becomes matured, all of the population will converge throughout small regions and make perturbations short [16].

DE is worked when it has the candidate solutions that are also called agents in a population. Those agents will move freely around the search space before combining itself with existing agents' positions if the new agents bring an improvement to be kept and accepted into the populations. In contrast, for not accepted agents, they will be discarded. All those processes will be repeated but not guaranteed.

C. Bees Optimisation Algorithm (BOA)

Bees Optimisation Algorithm (BOA) is the social behaviour of honeybees to search for food. This algorithm was developed in 2005 [17]. It measures the distance between the defined solutions. BOA will go through local and global search, but BOA will not use any probability approach. It will use a fitness evaluation to proceed with the search process.

BOA, in other words, also has the local and global search that will utilise the exploitation and exploration process [18]. The fitness value will then be checked and sorted descending or going through a process called maximisation problems. The local method will cover the fittest location as its best locations and classify it into two subgroups known as elite and non-elites' sites. The local process will start by recruiting the bees in the elite and non-elite places. The global search will then randomly run the search process before the location is sorted to their fitness value, and global optimum appeared [17].

III. BLACK WIDOW OPTIMISATION (BWO)

In this paper, a new set of algorithms has been proposed called Black Widow Optimization (BWO) algorithm, which implements from nature. This algorithm gets the idea from the observations of the black widow spider's unique mating behaviour. Australian Black Widow, also known as redback spider (Latrodectus Hasselti), originated from South Australia or adjacent Western Australian deserts [19]. Now, this spider has also been migrated throughout Australia, Southeast Asia, and New Zealand. A significant difference between a female spider and a male spider is that the adult female spider would be bigger than the male spider. That is why the female spider can reach 10 millimetres, and the male spider is only 3 – 4 millimetres long [20]. The adult female spider has a red stripe on the abdomen's upper side at her spherical shape body.

The mating behaviour's uniqueness can be seen when the female black widow would consume the male black widow, and this behaviour would be called Sexual Cannibalism [21]. Sexual Cannibalism will happen when a female eats a male either before, during, or immediately after mating. For the black widow's case, the adult female spider would mark a particular spot at her net with pheromone hormone. This pheromone would act as a smell that can attract male spiders to come to their trap. After that, the female spider will bite the male spider with her two fangs and inject a complex venom to its prey that definitely would kill them. The female spider will then wrap the male spider with silk before she sucked out all the liquid inside the male spider. This process will help the female spider lay 4 to 10 eggs sacs containing around 250 eggs [22]. The process does not stop here, as the process of Cannibalism will proceed among spiderlings. It is known as Sibling Cannibalism [19]. The spiderlings will consume each other while they live together in their maternal web before being carried away by the wind. This behaviour obviously will affect the population level; however, it is good to raise parental fitness as the survivors have an enhanced body condition.

Lastly, for the unfertilised spiderlings, it is also recorded that they will eat their mother slowly, but this does not happen frequently. This behaviour is called as Matriphagy behaviour. Therefore, several factors caused all of these cannibalistic behaviours:

- Being competitive among each other.
- Lack of food source.

BWO algorithm is used to estimate the parameter of Chinese Hamster Ovary (CHO) cells. This optimisation will use a black widow's behaviour to develop the algorithms to find the best solutions that minimise the nonlinear least-square value that determines the differences between model prediction and experimental data. From the basic concept of the black widow's mating behaviour, a pseudo-code has been developed to give a simple explanation [20], shown in Table II. The process will generate the initial population of the spider and evaluate each spider's fitness. Next, we will decide either we want to end the condition or proceed with the situation. If we move with the problem, it will start with randomly selecting the parents before procreation happens as they will try to reproduce the new generation of spiders in pairs. After that, the process of Cannibalism will take place before the mutation process. Lastly, the population of the spiders will be updated.

A. Initial Population

The optimisation problem can be solved with appropriate values of problem variable structure, and for the BWO algorithm, it is known as "widow." Each of the black widow spiders will show the problem variables, and for this paper, the structure will be represented as an array [20].

For the dimensional optimisation problem $[Nvar]$ the solution for the problems will be represented by the widow as an array $[1\ x\ Nvar]$. This array would be defined as follows[1]:

$$Widow = [x_1, x_2, \dots, Nvar] \quad (2)$$

All of the variable values $[x_1, x_2, \dots, Nvar]$ will act as a floating-point number. Therefore, the evaluation of fitness function f will be helpful in order to the fitness value of the widow.

$$Fitness\ = f\ (widow) = f\ [x_1, x_2, \dots, Nvar] \quad (3)$$

In order to start the optimisation algorithm, the initial populations of spiders are very important in terms of generating the candidate widow matrix $[Npop\ x\ Nvar]$ before performing the procreating step via mating by randomly selected pairs of parents.

TABLE II. PSEUDO - CODE OF BWO ALGORITHM

Input Maximum number of iterations, Rate of procreating, Rate of Cannibalism, Rate of mutation Output Near-optimal solution for the objective function
// Initialization 1 The initial population of black widow spiders Each pop is a D-dimensional array of chromosomes for a D-dimensional problem ***// Loop until the terminal condition*** 2 Based on the procreating rate, calculate the number of reproductions "nr" 3 Select the best nr solutions in pop and save them in pop1 ***// Procreating and Cannibalism*** 4 For i=1 to nr do 5 Randomly select two solutions as parents from pop1 6 Generate D children using Equation 1 7 Destroy father 8 Based on the cannibalism rate, destroy some of the children (new achieved solutions) 9 Save the remain solutions into pop2 10 End For ***// Mutation*** 11 Based on the mutation rate, calculate the number of mutation children "nm" 12 For i=1 to nm do 13 Select a solution from pop1 14 Randomly mutate one chromosome of the solution and generate a new solution 15 Save the new one into pop3 16 End For ***// Updating*** 17 Update pop = pop2 + pop3 18 Return the best solution 19 Return the best solution from pop

B. Procreate

The spider will start to mate here as to generate new generations of the spiders. Researches show that each mating process will produce around 1000 eggs, but only the strongest and fit baby spiders will be survived[1]. Now with the presence of random numbers of widow array, an array named alpha will be created so the offspring that have been produced will be using α.

$x_1$and x_2= parents & $y_1$and y_2 = offspring

$$y_1 = \alpha \times x_1 + (1 - \alpha) \times x_2 \quad (4)$$

$$y_1 = \alpha \times x_2 + (1 - \alpha) \times x_1 \quad (5)$$

This process should not be duplicated but must be repeated for $\left[\frac{Nvar}{2}\right]$ times. Then, the spiders, both mother and children will be added to the array and be sorted according to their fitness value.

C. Cannibalism

There were three kinds of Cannibalism involved in this algorithm[1]:

1) Sexual Cannibalism

- The female black widow will eat her husband during or after mating process.
- The fitness value can determine their gender.

2) Sibling Cannibalism

- The competitions between spiderlings occurred as the stronger spiderlings would eat the weaker spiderlings.
- The survivors among siblings would be set as Cannibalism Rating (CR).

3) Matriphagy

- It rarely happened but still being observed.
- The unfertilised spiderling will eat their mother slowly.
- The quality of the spiderlings, whether strong or weak, can be determined using fitness value.

D. Mutation

The mutation process will begin by selecting Mutepop as the number of individuals randomly. Two elements in the array would be randomly exchanged by each chosen solution. The mutation rate that has been calculated called Mutepop. Fig. 2 shows the flow of BWO Mutation.

E. Convergence

There are three stop conditions that can be considered, and I choose (i) as I define the iterations and can see later that BWO will have better results compared to other algorithms:

1) The predefined number of iterations.

2) Observe the unchanged fitness value for several iterations.

3) Reached the level of accuracy that has been specified.

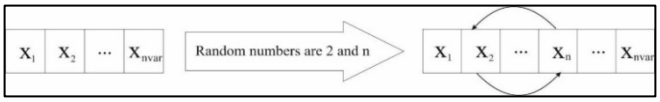


Fig. 2. BWO Mutation.

F. Parameter Setting

Some parameters need to be taken care of to get better results. These parameters need to be adjusted well to improve the algorithm's success rate. The right parameter values are very important to control the exploitation balance and stages of explorations.

1) Procreating Rate (PP)

- The percentage of procreating.
- Determine the number of individuals involved to procreate.
- Control the offspring productions and provide more chances to explore the search space.

2) Cannibalism Rate (CR)

- Acts as a controller to the cannibalism operator that would deduct unfit individuals from populations.
- High performance of the exploitation stage can be ensured by adjusting the parameter with proper values.

3) Mutation Rate (PM)

- The individuals participate during mutations in percentage.
- Control transferring between search agents from the global and local stage.
- Propelling towards best solutions.

IV. Experimental Setup

A. Dataset

In this research, Chinese Hamster Ovary (CHO) cells have been used as the datasets as it is widely used in the biological engineering process. CHO has already been used in research since 1919. It is because of CHO cells' characteristics compared to other cells regarding their rapid growth in suspension culture and high protein production [5]. Therefore, the CHO cells are useful for radiation cytogenetics and tissue culture. Table III shows the details of the CHO model used in this research.

CHO cells are also used in protein production for fermentation processes [16]. These cells will not be fed for 12.5 days, equal to 300 hours. This process is called the simulation batch process using resting cells. As mentioned above, the cells are usually used in the fermentation process, which has already become industrial to make products useful for the human being. Therefore, the fermentation process involves some metabolic processes that will act on its enzymes by making chemical changes to its organic substrate. This industrial fermentation has been applied to the food industry and the general industry to produce commonly used chemical concentrations such as citric acid, ethanol, and

lactate. All of these cells' concentration will be affected by the rate of fermentation. The fermentation process will need a fermenter medium to supply all of those nutrients needed by the organisms or cells. Examples of sources involved (including carbon and amino acid) are glucose and leucine or methionine, respectively. Therefore, the fermenter process will measure 13 metabolites, such as glucose, lactate, product protein, leucine, methionine, cytosol, aspartate, malate, pyruvate, oxaloacetate, ATP, and ADP. All of these samples will be taken daily during the whole fermentation process.

In CHO, 35 metabolites are comprised and divided into three compartments: fermenter, cytosol, and mitochondria. Thirty-two reactions are generated in CHO that included protein product formation (the process of connecting the protein using peptide bonds and the process of forming the polypeptide from amino acids substance), Embden – Meyerhof Parnas Pathway (the conversion of glucose using the metabolic pathway), TCA cycle, amino acid metabolism reduction, production of lactate, and electron transport chains. Therefore, the ordinary differential equation (ODE) model produced will comprise 117 parameters and ensure that the purpose of doing optimisation by generating pseudo experimental data and mimicking typical cells' behaviour is achieved.

TABLE III. CHO MODEL USED IN THIS RESEARCH

Model ID	Cell	Description Level	Parameter	Data Type
Chinese Hamster Ovary	CHO	Metabolic	117	Simulated

B. Comparisons of Methods Experiment

The best values of all the algorithms used, including BWO, PSO, DE, and BOA, were evaluated. The CPU times were also compared to show which algorithm takes longer times. All of them were tested with CHO sample data. Thus, this work was performed using Matlab R2015b.

V. RESULTS AND DISCUSSION

In this section, the Black Widow Optimization (BWO) algorithm has been tested with a constant number of iterations set to 100 iterations. All of these experiments are time-consuming and need excellent computational systems for it to work well. The results obtained from the experiments will be compared with other existing metaheuristics, which are Particle Swarm Optimization (PSO), Differential Evaluation (DE), and Bee Optimization Algorithm (BOA). Table IV presents all of the parameters involved in experiments for 20 runs.

An initial population of the spider has been generated and then evaluates each spider's fitness. Continue with the decision making, either the condition is ended or proceeded. The crossover process or procreating then happens, so the second population of spiders is generated. After that, the process of mutation generates the third population will take place. These crossover and mutation are significant because they can start to reproduce the new generation of the spiders by uniting all the populations that have been generated. Then, beginning to sort the spider population is based on its fitness value to determine the best cost of the BWO algorithm with 117 dimensions (nVar). Lastly, all the extra individuals are deleted before the population of the spiders is updated. Hence, the results are always different from each time of the experiments run because of the problem's stochastic nature. The results are reported in Table V. The table shows the experiments' results between BWO, PSO, DE, and BOA for 117 dimensions. According to the table, BWO, which always has the lowest best value, can be considered the best algorithm compared to the others despite its not having the shortest CPU Times. As a result, BWO can be claimed can converge better. Therefore, there were several advantages and disadvantages of BWO [20], shown in Table VI.

A. Comparison In Graph (Best Cost by 100 Iteration)

Fig. 3, 4, 5 and 6 below shows the graph of each experimental algorithm from BWO, PSO, DE and BOA for its best cost by 100 iterations.

B. Comparison In Graph (Best Cost by CPU Times)

Fig. 7, 8, 9, and 10 below shows the graph of each experimental algorithm from BWO, PSO, DE and BOA for its best cost by CPU Times.

TABLE IV. PARAMETERS AND VALUES INVOLVED IN EXPERIMENTS

Algorithms	Parameters	Values
BWO	Percent of Crossover = pc (Procreating Rate)	0.60
	Cannibalism Rate = pCannibalism	0.44
	Percent of Mutation = pMutation (Mutation Rate)	0.40
PSO	Inertia Weight = w	2.00
	Best personal Experience = c1	2.40
	Best Global = c2	2.20
	Intertia Weight Damping Ratio = wdamp	0.98
DE	Mutation Factor = beta	0.50
	Crossover Constant = pCR	0.20
BOA	Neighborhood Radius Damp Rate = rdamp	0.95

TABLE V. RESULTS

Algorithms	Iteration	Best Cost	CPU Times
BWO	100	22488.5989	6288.8125
PSO	100	155817.0963	1714.3438
DE	100	2992715.1241	2577.6563
BOA	100	4703432.5252	20781.25

TABLE VI. ADVANTAGES AND DISADVANTAGES OF BWO

Advantages	Disadvantages
Provide better results in exploitation and exploration stages.	New optimisation algorithm.
Deliver fast convergence speed.	Does not fully exposed yet.
Able to check the large area to get the best solutions.	

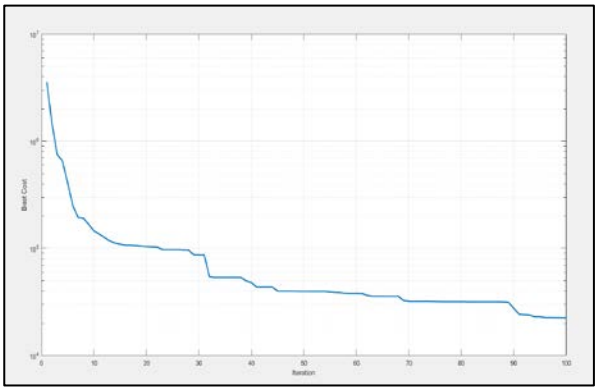


Fig. 3. The Best Cost of BWO by 100 Iteration.

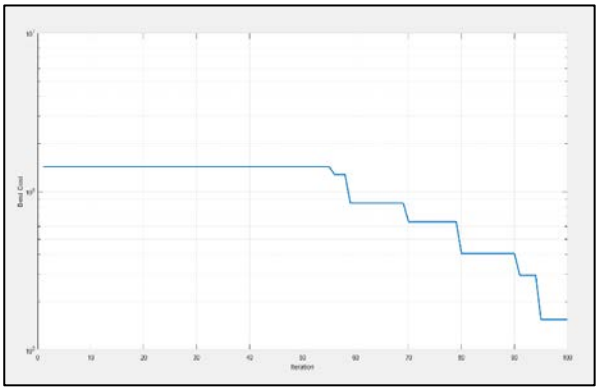


Fig. 4. The Best Cost of PSO by 100 Iterations.

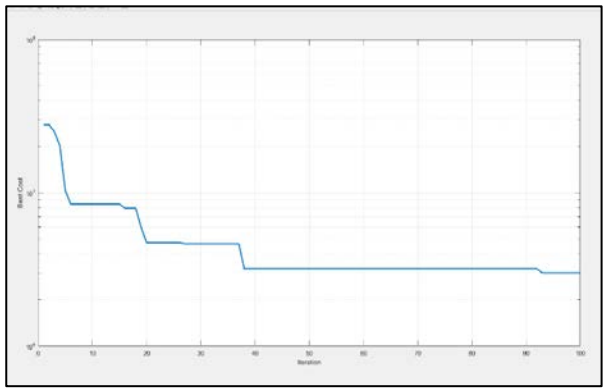


Fig. 5. The Best Cost of DE by 100 Iterations.

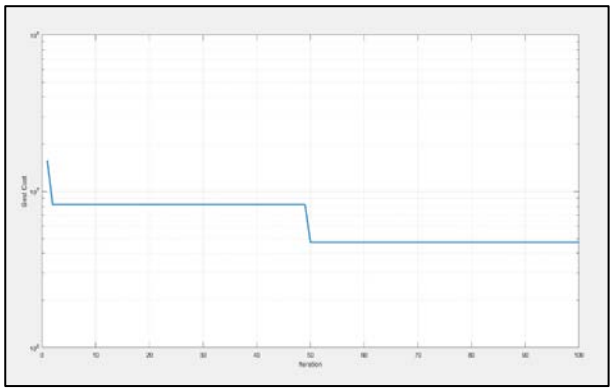


Fig. 6. The Best Cost of BOA by 100 Iterations.

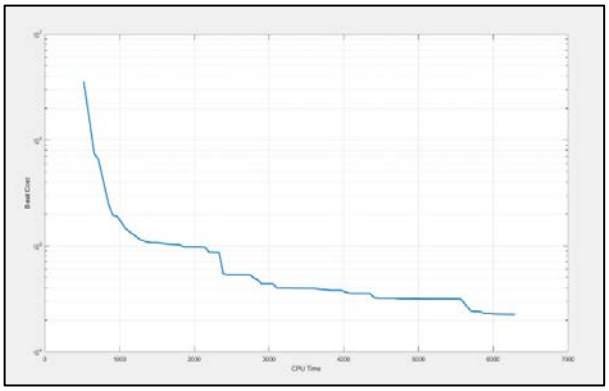


Fig. 7. The Best Cost of BWO by CPU Times.

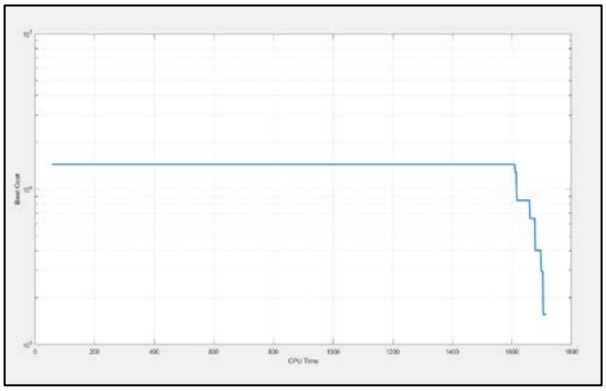


Fig. 8. The Best Cost of PSO by CPU Times.

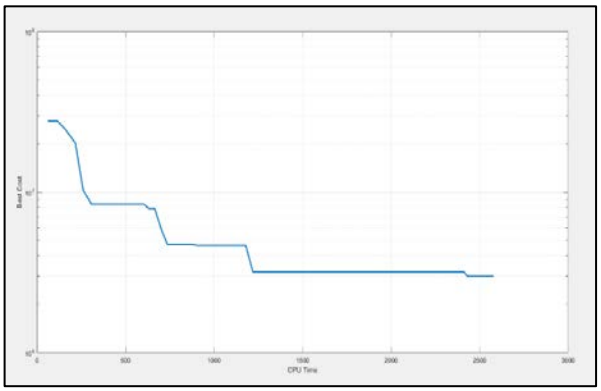


Fig. 9. The Best Cost of DE by CPU Times.

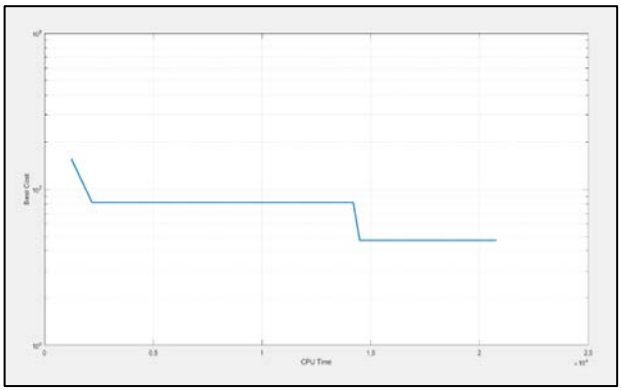


Fig. 10. The Best Cost of BOA by CPU Times.

VI. Conclusions

In conclusion, this paper discussed the new algorithms called Black Widow Optimization (BWO), inspired by a venomous spider's unique and extreme mating behaviour. This algorithm has been studied with three other algorithms, which are Particle Swarm Optimization (PSO), Differential Evolutionary (DE), and Bees Optimization Algorithm (BOA). It has shown high performance to find the most nearer optimal values that are accurate and fast in converging. According to the results, BWO can be a useful contribution to solving the larger optimisation scale problem. The proposed algorithm would be limited in terms of the metabolic field and can contribute to real-world optimisation problems and be applied to various areas. Besides, randomly selecting the parents of the spider in BWO has ensured the exploration process. At the same time, offspring produced in the procreate step already emphasises it and help prevent traps in local optima. The solutions that have been omitted during the cannibalism stage also help BWO algorithms work faster to achieve the best solutions. Lastly, the BWO algorithm can be considered to be a smart algorithm to solve optimisation problems.

This research has a few limitations. Although BWO shows obtained the best value in terms of the best cost, this algorithm sometimes takes to finish the experiments compared to other algorithms in this study: PSO, DE, and BOA. Second, this research also needs a good quality of PC or laptop processor to ensure that the software users do not lag and stop working suddenly.

For future study, BWO could be tested with more other metaheuristics and different kinds of datasets. Also, trying the BWO with a larger iteration is to ensure fitness.

Acknowledgement

We would like to thank Malaysia Ministry of Higher Education for their support via Fundamental Research Grant Scheme [FRGS/1/2018/ICT02/UMP/02/8] and Fundamental Research Grant Scheme - Malaysia Research Star Award [R/FRGS/A0800/01655A/003/2020/00720].

References

[1] J. Almquist, M. Cvijovic, V. Hatzimanikatis, J. Nielsen, and M. Jirstrand, “Kinetic models in industrial biotechnology - Improving cell factory performance,” Metabolic Engineering, vol. 24. Academic Press Inc., pp. 38–60, 01-Jul-2014.

[2] “10 Everyday uses of Biotechnology | CPI.” [Online]. Available: https://www.uk-cpi.com/blog/10-everyday-uses-of-biotechnology. [Accessed: 20-Jul-2020].

[3] “(PDF) Industrial Biotechnology: An Overview.” [Online]. Available: https://www.researchgate.net/publication/311576484_Industrial_Biotechnology_An_Overview. [Accessed: 19-Jul-2020].

[4] G. G. Wang, S. Deb, and Z. Cui, "Monarch butterfly optimisation," Neural Comput. Appl., vol. 31, no. 7, pp. 1995–2014, 2019.

[5] M. A. Remli, S. Deris, M. S. Mohamad, S. Omatu, and J. M. Corchado, “An enhanced scatter search with combined opposition-based learning for parameter estimation in large-scale kinetic models of biochemical systems,” Eng. Appl. Artif. Intell., vol. 62, pp. 164–180, Jun. 2017.

[6] “Parameter Estimation.” [Online]. Available: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Confidence_Intervals/BS704_Confidence_Intervals2.html. [Accessed: 20-Jul-2020].

[7] J. Y. Kim, Y. G. Kim, and G. M. Lee, “CHO cells in biotechnology for production of recombinant proteins: Current state and further potential,” Applied Microbiology and Biotechnology, vol. 93, no. 3. pp. 917–930, Feb-2012.

[8] “genetic engineering | Definition, Process, & Uses | Britannica.” [Online]. Available: https://www.britannica.com/science/genetic-engineering. [Accessed: 20-Jul-2020].

[9] “Machine learning: an introduction to mean squared error and regression lines.” [Online]. Available: https://www.freecodecamp.org/news/machine-learning-mean-squared-error-regression-line-c7dde9a26b93/. [Accessed: 20-Jul-2020].

[10] A. F. Villaverde et al., “BioPreDyn-bench: A suite of benchmark problems for dynamic modelling in systems biology,” BMC Syst. Biol., vol. 9, no. 1, pp. 1–15, 2015.

[11] Remli MA, Mohamad MS, Deris S, Sinnott R, Napis S. An Improved Scatter Search Algorithm for Parameter Estimation in Large-Scale Kinetic Models of Biochemical Systems. Current Proteomics. 2019 Oct 1;16(5):427-38.

[12] Kunna MA, Kadir TA, Remli MA, Ali NB, Moorthy K, Muhammad N. An Enhanced Segment Particle Swarm Optimization Algorithm for Kinetic Parameters Estimation of the Main Metabolic Model of Escherichia Coli. Processes. 2020 Aug;8(8):963.

[13] X. S. Yang and M. Karamanoglu, “Swarm Intelligence and Bio-Inspired Computation: An Overview,” in Swarm Intelligence and Bio-Inspired Computation, Elsevier Inc., 2013, pp. 3–23.

[14] Z. Cui and X. Gao, “Theory and applications of swarm intelligence,” Neural Computing and Applications, vol. 21, no. 2. pp. 205–206, Mar-2012.

[15] R. Storn and K. Price, "Differential Evolution - A Simple and Efficient Heuristic for Global Optimisation over Continuous Spaces," J. Glob. Optim., vol. 11, no. 4, pp. 341–359, 1997.

[16] S. Mirjalili, S. M. Mirjalili, and A. Lewis, “Grey Wolf Optimizer,” Adv. Eng. Softw., vol. 69, pp. 46–61, 2014.

[17] D. T. Pham and M. Castellani, “A comparative study of the Bees Algorithm as a tool for function optimisation,” Cogent Eng., vol. 2, no. 1, Dec. 2015.

[18] B. Yuce, M. S. Packianather, E. Mastrocinque, D. T. Pham, and A. Lambiase, “Honey bees inspired optimisation method: The bees algorithm," Insects, vol. 4, no. 4, pp. 646–662, Nov. 2013.

[19] F. F. S. Daly, R. E. Hill, G. M. Bogdan, and R. C. Dart, "Neutralisation of Latrodectus mactans and L. hesperus venom by redback spider (L. hasseltii) antivenom,” J. Toxicol. - Clin. Toxicol., vol. 39, no. 2, pp. 119–123, Jan. 2001.

[20] V. Hayyolalam and A. A. Pourhaji Kazem, "Black Widow Optimization Algorithm: A novel meta-heuristic approach for solving engineering optimisation problems," Eng. Appl. Artif. Intell., vol. 87, p. 103249, Jan. 2020.

[21] J. E. Garb, A. González, and R. G. Gillespie, “The black widow spider genus Latrodectus (Araneae: Theridiidae): Phylogeny, biogeography, and invasion history,” Mol. Phylogenet. Evol., vol. 31, no. 3, pp. 1127–1142, 2004.

[22] O. R. Pires, W. Fontes, and M. S. Castro, “Recent Insights in Latrodectus (‘Black Widow’ Spider) Envenomation: Toxins and Their Mechanisms of Action,” in Spider Venoms, Springer Netherlands, 2016, pp. 333–344.

Self-Organizing Map based Wallboards to Interpret Sudden Call Hikes in Contact Centers

Samaranayaka J. R. A. C. P[1], Prasad Wimalaratne[2]
University of Colombo School of Computing
Colombo, Sri Lanka

***Abstract*—In a contact center, it is required to foresee and excavate any disturbance to the daily experiencing call pattern. Abnormal call pattern may be a result of a sudden change in the organization's external world. Expecting a methodological analysis prior to meet customers' demand may introduce a delay for queuing customers. It is required a fast and promising method to predict and reasoning any unwilling event. It is not possible to draw conclusions by considering one dimension such as total call count. Total call count may increase in same way due to a failure in any service. Research mainly focusses on reasoning multidimensional events based on historical records. In contrast to traditional wallboards, our approach is capable of clustering and predicting disturbances to the normal call patterns based on historical knowledge by considering many dimensions such as queue statistics of many service queues. Our approach showed improved results over traditional wallboards equipped with 2D or 3D graphs.**

***Keywords*—Multidimensional data; visualization; contact centers; self-organizing map; clustering**

I. Introduction

Contact center is a very sensitive customer touch point within an organization [1]. In a different explanation, any sudden incident in the external world of an organization will be reflected immediately via contact centers. Contact center is a combination of both technology and human who are waiting to provide information required by the callers. Neither customer care officer nor a customer will be able to conclude a sudden change in the outside world of an organization. Simply, incident can be exampled as an electricity failure of a region for national electricity provider or common telephone cable damage due to a heavy rain for national telephone provider. In both cases, customers may unhappy about the service interruption and they will try to complain about it by resulting a huge demand upon the relevant contact centers.

It will create a need for a software mediator which could extract information from both caller and customer care office, analyze against historical knowledge and present to the supervisors who can decide any deviation from normal operational procedures. Such an operational deviation is mandatory in a sudden event because contact centers were not designed to absorb abnormal conditions to associated capital investments. Financially benefited strategy is maintaining a separate approach for sudden incidents. To have a competitive advantage over the competitors who are in the same field, it is required a good insight about when to switch strategies without diluting customer satisfaction. Our research focusses on introducing wallboards with capability to visualize multidimensional data along with clustering to support real time decision making with the support of historical knowledge.

II. Related Work

Due to increasingly generating and accumulating large amount of unstructured data, it is often tending to excavating knowledge behind the data. While analyzers interested about user-friendly interactive systems for knowledge extraction, managers like people who are working in a contact center or help desk are more interested about wallboard [2] type information systems to alert timely any deviation with related to their predictions or organizational directions. This para will motivate the trending research interest about visualizing dynamic data. Galkin and others have examined a pipeline approach to visualize and analyze dynamic data to facilitate anomaly detection, clustering, trend analysis and variation analysis [3]. Their research explained that analyzing complex multidimensional data will introduce a deep insight to data science and related industries. This can be accomplished by using the influence of trending technologies such as artificial intelligence, machine learning and neural networks [4], [5]. In the past, several research interests have been showed in the area of time dependent multidimensional data visualization [6], [7] and one of the common observation in the above mentioned area is seeking the assistance of a domain expert to interpret the result which was projected to a 2D or 3D coordinate system. In Steven's approach, they are using a progressive technique which rendering the incoming data progressively. In other words, framework will initially construct a geometry to represent a past summery and updating a dynamic scalar field on top of that geometry [8]. In their research, they have tried to address a challenge which is handling and representing dynamic scalar in an efficient and user friendly manner. Mashima and other have presented a map based metaphor for visualizing dynamic data [9], [10]. Point (peg) in a map represents a multidimensional data while distances from the other points on the map are proportional to the similarity between points. In the other words, higher the distance mean less similar points. Researchers suggested that animations can be used as user interactions to successfully present dynamic scalars in a view. Wallboard kind of applications can be equipped with periodic updates or animations to successfully project multidimensional data with dynamic behavior to a 2D screen.

Throughout the history data visualization was a trending topic due to need of knowing "what is it hiding". People tried to interpret hidden knowledge in a human understandable way

[11]. A better visual representation is easily understandable and holds features like effectiveness, accuracy, robustness, easy to use, etc. [12], [13]. Historical effort for visualizing multidimensional data can be exampled with techniques like Scatter Plot matrix. With increase of velocity, volume and veracity of data collections, it opened new research areas like virtual reality for data visualization, augmented reality for data visualization, effective use of interaction methods for data visualization, etc. [14]- [16]. Wallboard is a display placed in a public area to convey any interested information. It can be equipped with effective interaction techniques.

Another important branch of visualization is dimensionality reduction techniques such as Principal Component Analysis (PCA) [17], Radial Coordinate visualization (RadViz) [18], self-organizing maps (SOM) [19], Ridge and Lasso Regression [20], [21], Singular Value Decomposition (SVD) [22]. Above two paragraphs and techniques were motivated to build a hybrid visualization model which is a collection of both multidimensional visualization models' characteristics and dimensionality reduction techniques' characteristics. Among the dimensionality reduction techniques, self-organizing maps showed better performance for clustering over other clustering techniques [23].

Interactivity can be embedded to a concept by using technologies like JavaScript, Ajax, HTML, Hadoop [24], NodeJS, etc. These technologies collectively introduce continuous connectivity in-between frontend and backend for enabling real time or near real time updates in the front end.

Since visualization is for humans, it is needed to consider human factors in visualization [25], [26]. Ability to operate in a low resources environment is a greater achievement with increasingly generating various types of data. This fact led research to design and implement a distributed architecture for information processing.

III. WALLBOARD

Below will demonstrate commonly available wallboard types with both open source and commercial contact center products. Asterisk can be referred under the free and open source contact centers. They may not provide identical wallboards as below figures (Fig. 1 and Fig. 2). But, it will maintain the same concept.

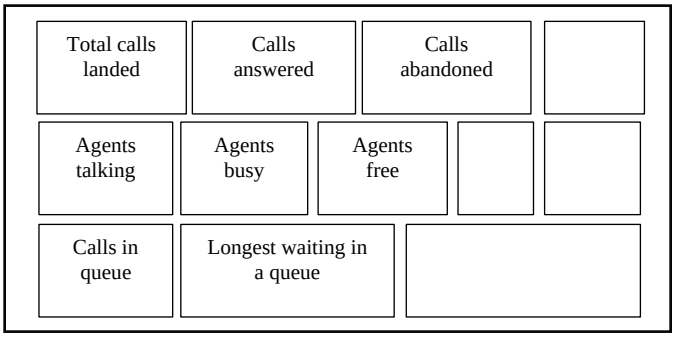


Fig. 1. Standard Wallboard of a Contact Center.

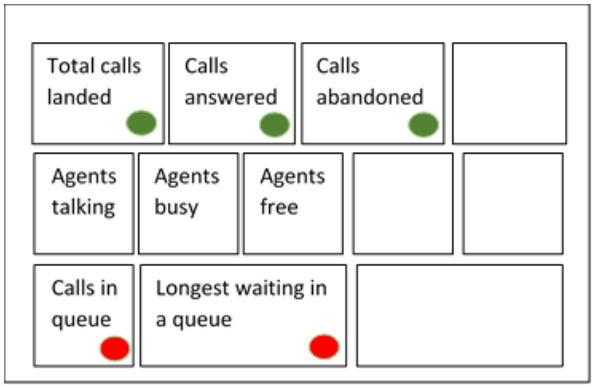


Fig. 2. Standard Wallboard of a Contact Center with Alarms based on Past Data (it will Indicate Whether a shown Value Exceeds a Threshould or Not).

Because of market competition, both open source and commercial contact center products tend to offer more informative dashboards or information visualization modules with their products. Presently, pattern recognition is a lacking and demanding area with information visualization which is related to contact center solutions.

IV. METHOD

A. Customer Information flow of a General Contact Center

Generally, contact center will consist with below major logical modules to cater demanding customer requirements.

- Greeting message: This is the first message customer can here. Example: Good morning.
- Skill: Skill is an attribute which is assigned to an agent. Agent Skills can be regard as the ability of an agent to handle a specific call which requires one of those skills. In relationship with contact center, skill can be thought as a specific customer need/requirement or perhaps a business need of contact center. Contact center will define skills based on the needs of customers and contact center.
- Interactive Voice Response (IVR): It provides a list of steps that process calls in a user-defined manner. The steps in an IVR can send calls to skill, play announcements and music, disconnect calls, give a busy signal to the calls, or route calls to other destinations.
- Queue: Queue is a holding area for calls which are waiting to be answered in order. Different calls in a queue may have different priority levels, in which case, calls with a higher priority will be answered first.
- Agents terminal: This is the place were agents can pick and answer the customer's call. Based on the organization's requirement calls may be landed automatically to the relevant agent's terminal.

B. Placement of the Analyzer

As above diagram (Fig. 3) shows, information flows through the above explained logical modules. Skill was embedded to agent who is using agent terminal and it is not separately drawn in the diagram. Selection of the information

flow which is needed to process was done based on below facts (Table I).

- Relevance to the presented analysis
- Richness of the information
- Experience of contact center management
- Experience of number management teams (Soft switch, etc.).

Based on the comments of number management team, due to central signal handling of phone calls, it introduces errors when grouping numbers into a geographical region by using number levels. For an example, number level 11111xxxxx – 11119xxxxx is belong to town A and number level 11121xxxxx – 11129xxxxx is belong to town B. Town A and B belongs to region X, since 111xxxxxxx assigned to region X. Because of the central management of signals (central switch), a number belongs to town A can be assigned to region Y. Because of this configuration, it dilutes the capability of localizing a regional failure due to a common reason which occurred in a particular region such as heavy rain.

Based on the experience of contact center management, contact center will be flooded from calls due to a failure in a region. In a such situation, we cannot experience a patient behavior from callers. All of the callers may have to wait longer minutes to reach agents. Most interesting observation is that caller is not capable to predict the source of failure and they will report what they feel. Because of that caller may select the required service based on their perspective. In this research, most significant information to analyze is queue statistic over the other available information when experiencing a sudden call hike. Based on the above explanation, we have selected information flow between queue and agent terminal to analyze.

C. Work Flow of the Proposed Software

Work flow (Fig. 4) was implemented with our solution to visualize results. As described by following figure, it will periodically calculate SOM and visualize its results. On top of the SOM's visualized content, it will place the present reading's 3D graphics. This is a periodic process with an appropriate delay to update visualized content or graphics.

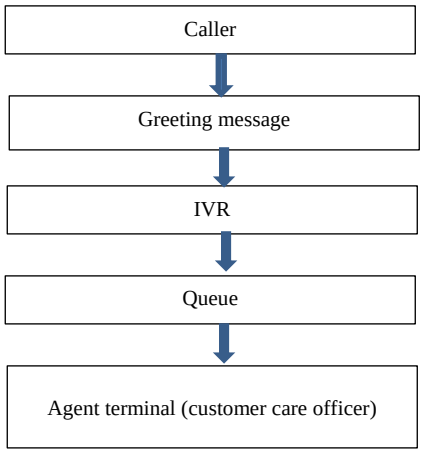


Fig. 3. Customer Information flow of a General Contact Center.

TABLE I. INFORMATION FLOWS BETWEEN COMPONENTS OF A GENERAL CONTACT CENTER

Information flow	Available information (richness)	Relevance to the presented analysis
Caller to greeting message	CLI (caller's telephone number)	Yes
Greeting message to IVR	CLI (caller's telephone number)	Yes
IVR to Queue	CLI, Customers required service	Yes
Queue to Agent	CLI, Customers required service, Queue statistics (waiting time, etc.)	Yes

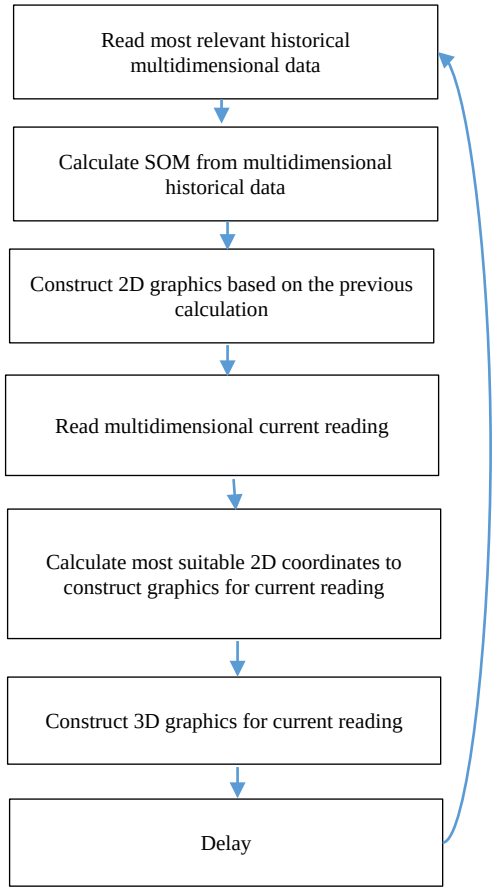


Fig. 4. Proposed Approach's Dynamic Data updating Work Flow.

D. Architecture of the Proposed Software

Below architecture (Fig. 5) consisted with two main modules such as Visualization layer and "SOM algorithm implementation" module. They are separately described in below subsections in detail. "DB interface" was designed to insert data originated via contact center. Below explained architecture was implemented by using multiple technologies collaboratively such as C#, PHP, MySQL, JavaScript, Three.js, WebGL, Java, etc.

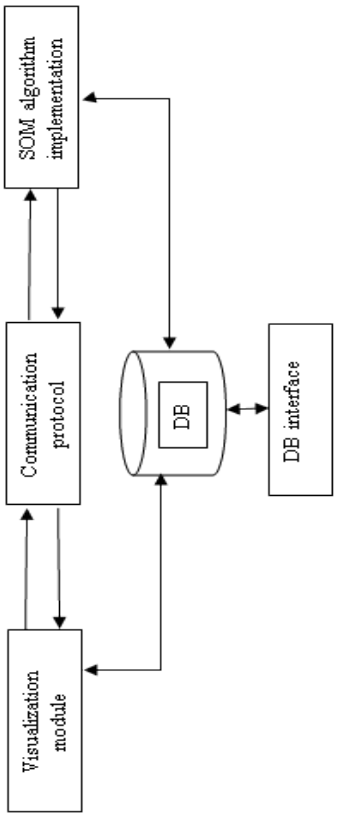


Fig. 5. Architecture of the Proposed Software.

V. Theory and Modules

A. "SOM Algorithm Implementation" Module

Kohonen Self-Organizing Map follows a two-layer approach which consists with input layer (top layer of above diagram (Fig. 6)) and output layer (bottom layer of above diagram (Fig. 6)). Both layers' neurons have same number of dimensions like $(w_{ij1}, w_{ij2}, w_{ij3}, \ , w_{ijn})$ and $(x_1, x_2, x_3, \ , x_n)$. Output layer's neurons' weights will be updated by using a neighborhood function and input layer's weights. As per the algorithm, input layer's neurons will be selected randomly and iteratively to update output layer's neurons via neighborhood function. Once the input layer's neuron was selected, matching output layer's neuron will be selected as per the below function.

$$D_{i,j} = |X - W_{i,j}| = \sqrt{(x_1 - w_{ij1}) + .. + (x_n - w_{ijn})} \quad (1)$$

Here, i and j: output layer indexes (raw number and column number)

Closest node,

$$D(k_1, k_2) = \min D_{i,j} \quad (2)$$

k_1 and k_2 : winning node's indexes

Secondly, below equations will explain the neighborhood function which will be used to update neighbors and its calculations.

$$h(\rho, t) = \exp(\frac{\rho^2}{2\sigma^2(t)}) \quad (3)$$

$$\rho = \sqrt{(k_1 - i)^2 + (k_2 - j)^2} \quad (4)$$

Here i and j: output layer indexes

k_1 and k_2 : matching node's indexes

$$\sigma(t) = \sigma_0 \exp(\frac{-t}{T_{max}}) \quad (5)$$

t: 0 .. T_{max}

σ_0: experimental constant.

B. Visualization Module

Cartesian coordinate system or spherical coordinate system is capable of handling maximum three variables at a time. We have combined two coordinate systems in our novel framework to enable plotting more than three variables simultaneously. Advantages of this technique over other available 3D visualization approaches are as follows.

- Consistency of all the plotted parameters/dimensions
- Ability to compare and contrast its own dimensions

Above explanation will elaborate in detail in the results and analysis section by using screenshots of the novel framework. In other words, this is the module which visualizes the output of "SOM algorithm implementation" module. This is an independent software and visualization module can be used to upload data by using data files, if it is not integrated with any application to pump data.

C. Communication between Modules

Communication between visualization layer (web application) and "SOM algorithm implementation" module was facilitated by this layer. As below diagram (Fig. 7) explained, this layer is a composition of TCP socket server and TCP socket client which were constructed as independent executables with separate configuration files. These executables will exchange metadata and results between above mentioned two modules. Transmitting preprocessed data or SOM algorithm's output is excluded from communication layer protocol and they will be placed on a shared location or can be transmitted by using FTP based on requirement.

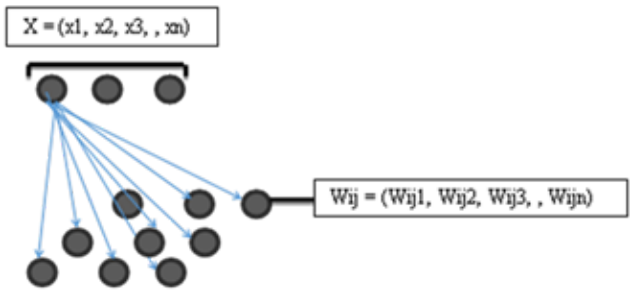


Fig. 6. Diagram of Kohonen Self-Organizing Map.

Fig. 7. Placement of TCP Socket Server and TCP Socket Client in Modules.

VI. Integration between Contact Center and Proposed Software

As per the below diagram (Fig. 8), contact center's Queue module can be integrated with "DB interface" to pump queue statistics data to database for the purpose of visualization and decision making. Upon successfully reaching an agent collected queue statistics will be sent to database via "DB interface". Then, visualization module can be loaded by using a preferred web browser to view output. It can be used as a wallboard to continuously monitor and analyze call patterns.

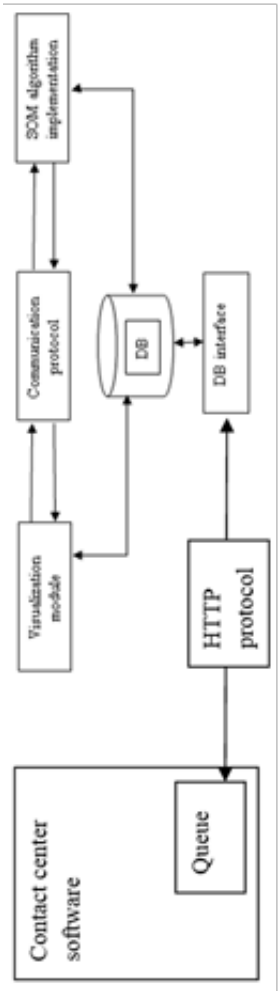


Fig. 8. Diagram of Integration between Contact Center and Proposed Software.

VII. Dataset and Dimensions

This section will explain the dataset which can be introduced as queue statistics and its structure in detail. Mainly it has two tables (Table II and Table III). Answered call load pattern table was filled with successfully answered calls by contact center agents while abandoned call load pattern table was filled with calls which is not answered by a contact center agent. Answered call load pattern table has columns like,

- (0-5) which means number of callers who had to wait in a queue between 0 to 5 seconds before answered by a contact center agent in a given month.
- (6-10) which means number of callers who had to wait in a queue between 6 to 10 seconds before answered by a contact center agent in a given month.
- (11-15) which means number of callers who had to wait in a queue between 11 to 15 seconds before answered by a contact center agent in a given month.
- (>60) which means number of callers who had to wait in a queue more than 60 seconds before answered by a contact center agent in a given month.

Other than monthly cumulative call counts, it is possible to calculate cumulative figures based on 15-minute duration, hourly, daily and monthly. Our proposed software can accommodate all of the above mentioned cumulative figures. We have used daily and monthly cumulative figures to verify and calculate our approach's accuracy.

There may be many reasons to experience a sudden call hike in a contact center such as,

- Credit control action. Operator will discontent customers who did not pay the bill as a bulk. In such a situation, customer may try to reach contact center by asking information related to his/her bill.
- Regional service failure. In such a situation, large number of affected customers will try to complain about service interruption via contact center.
- Natural event such as heavy rain. This kind of events may lead to service interruptions, because of damage to outside infrastructure and delayed maintenance.

We have selected to analyze rainy seasons with related to call hikes. This selection was solely made to evaluate our approach's accuracy. But, our approach can be used with any of the above mentioned reasons. Rain is a seasonal event and it was not started and ended in a same day of a month in two different years. Monthly cumulative call counts will provide better accuracy which is more than 95%.

Same dataset can be used with different durations to calculate cumulative figures to detect any interested reason which is contributing to a sudden call hike.

In contrast with answered call load pattern table, abandoned call load pattern table has columns like,

- (0-5) which means number of callers who had to wait in a queue between 0 to 5 seconds before call is dropped in a given month.
- (6-10) which means number of callers who had to wait in a queue between 6 to 10 seconds before call is dropped in a given month.
- (11-15) which means number of callers who had to wait in a queue between 11 to 15 seconds before call is dropped in a given month.
- (>60) which means number of callers who had to wait in a queue more than 60 seconds before call is dropped in a given month.

TABLE II. ANSWERED CALL LOAD PATTERN

Month	0-5	6-10	11-15		51-55	56-60	>60
01/2015	302708	13354	9994		6995	7161	158597
01/2016	405035	9517	5352		3269	3170	44328
01/2017	456929	10853	8063		4384	4131	80207
01/2018	473343	10302	6471		3185	3064	38553

TABLE III. ABANDONED CALL LOAD PATTERN

Month	0-5	6-10	11-15		51-55	56-60	>60
01/2015	6952	3683	2520		2429	9001	88680
01/2016	3793	1652	849		830	3545	18691
01/2017	5507	2344	1666		1956	5716	42317
01/2018	3839	1277	718		582	3643	14399

Same as explained with answered calls, it is possible to calculate cumulative figures based on 15-minute duration, hourly, daily and monthly for abandoned calls. Our proposed software can accommodate all of the above mentioned cumulative figures.

Dimensions of our multidimensional data instances will be columns of both answered call load pattern table and abandoned call load pattern table or columns of abandoned call load pattern table based on the scenario.

- Scenario 1: if the impact of the event is short duration (several minutes) and contact center employed extra agents to tolerate sudden hikes, both answered call load pattern table and abandoned call load pattern table can be considered as the input for our approach.
- Scenario 2: if the impact of the event is long duration (several hours or days) and contact center has no extra agents to tolerate sudden hikes, abandoned call load pattern table can be considered as the input for our approach.

Rain is a seasonal event and it will introduce a long term impact to a contact center. In our analysis, it was confirmed that there are no additional agents to flatten sudden hikes. In such a scenario, there is no considerable variation with the values of answered table while there is a considerable variation with the values of abandon table. This will be explained in detail in results and analysis section. Considering more dimensions for the research will introduce a requirement for high processing power with related to both calculations and graphics. In this research, we have selected abandoned table's columns as dimensions for our multidimensional data instances.

By reducing the duration for previously explained cumulative value calculations, dashboards can be converted to real time sudden call hike indicators. Although our proposed approach is independent from the dataset (our novel approach is generalized and we have used this dataset to verify and calculate our approach's accuracy which is more than 95%), domain knowledge is mandatory to select most appropriate dimensions and durations for an analysis.

VIII. RESULTS AND ANALYSIS

In this section, we are describing the results drawn via our novel approach which was described along with the dataset throughout several sections. We have selected queue statistics from historical data DB randomly for rainy months and normal months. Randomly selected data (multidimensional data arrays which was explained in previous section) were categorized by using SOM algorithm and visualized as below (Fig. 9) (blue and green clusters for rainy and normal months).

Workflow of the proposed software was explained in a previous section. According to the workflow, next step is placing the current value which is a multidimensional data instance on top of the visualized SOM. Three white spheres imply three tests for the proposed application and it was verified that drawn conclusions against actual records. A sample of tests which were conducted to verify shown results in the visualization layer were presented in below table (Table IV). It was achieved more than 95% accuracy level for detecting rainy conditions which may influence sudden call hikes. Although this testing was conducted to calculate accuracy and verify functionality of the proposed software, this approach was generalized to use with any different dataset and scenario.

White spheres consist with yellow bars which represent magnitudes of its own dimensions on its perimeter. Our concept's visualization layer was innovated to visualize multidimensional data with the help of SOM based categorization. Table V shows a comparison between our novel visualization concept vs groups of existing multidimensional visualization techniques.

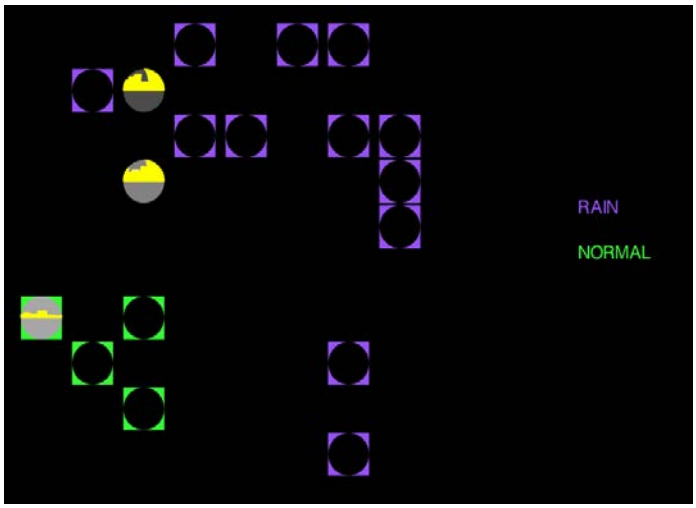


Fig. 9. Output of the Novel Application.

TABLE IV. SAMPLE OF TESTS CONDUCTED WITH VISUALIZATION LAYER

Number	Rainy or normal according to visualized output	Rainy or normal according to past records
1	Normal	Normal
2	Rainy	Rainy
3	Rainy	Rainy

TABLE V. COMPARISON OF DISTINGUISHABLE FEATURES OF VISUALIZATION LAYER

Feature	Our concept	Reducing dimensionality	Multi-dimensional data visualization (2D or 3D)
Flexible number of dimensions (any number of dimensions can be visualized without changing the overall appearance)	Yes	Yes (select most important 2 or 3 dimensions based on a selection algorithm)	Yes (example: scatter plot matrix, parallel coordinates plot, etc.)
Ability to compare behavior of its own dimensions (dimensions of a single data instance)	Yes (all dimensions will be visualized in same way)	No (not possible)	No (example: scatter plot matrix)
Ability to group by different dimensions over visualized dimensions (group by dim1 and dim2, but visualize dim1 and dim3)	Yes (SOM algorithm for grouping and different layer for visualizing)	No (not possible)	No (no separate layers for clustering and visualization)
Interactivity with the model	Yes	No	Yes

As explained in the later part of the section DATASET AND DIMENSIONS, there is no considerable variation with the values of answered table while there is a considerable variation with the values of abandon table. As per the below figure (Fig. 10), there is a high fluctuation in abandoned call counts over answered call counts. Our selection of abandoned table's columns as dimensions was justified. This behavior was experienced mainly because of unavailability of extra contact center staff to cater a sudden demand.

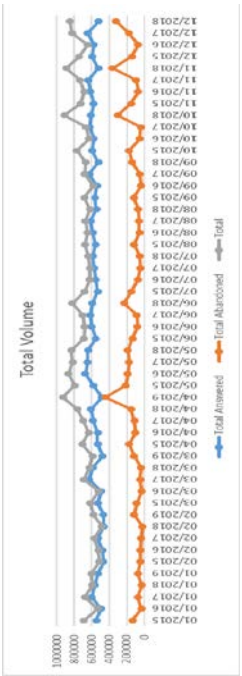


Fig. 10. Trend of Total Calls, Total Answered Calls and Total Abandoned Calls vs Time.

By combining two or more visualization and dimensionality reduction techniques, it is possible to have more features than sticking into a single visualization or dimensionality reduction technique. This can be concluded as the main finding of this research.

IX. CONCLUSION

Our approach showed improved results when predicting cause against traditional wallboards equipped with statistical analysis. SOM based approach shows better results when analyzing multidimensional data over traditional wallboards which shows charts with two or three axis. Results were highly depending on selection of dimensions correctly. In contrast with traditional wallboards, one wrong dimension selection may dilute the quality of overall prediction. Promising process for selecting dimensions will be a future work for this research.

ACKNOWLEDGMENT

I want to thank University of Colombo School of Computing, Colombo, Sri Lanka for guiding me to successfully complete my research project.

REFERENCES

[1] C. Samaranayaka and S. Premaratne, "Enhancing call center operations through sementic voice analysis," p. 10, 2018.

[2] "24/7 CALL Center Solution: Business Purpose Call Center System with Asterisk PABX," vol. 6, no. 9, p. 19, 2016.

[3] Galkin, Popov, Pilyugin, and Grigorieva, "The Visualization Method Pipeline for the Application to Dynamic Data Analysis," 2019.

[4] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.

[5] N. Kour and D. N. K. Gondhi, "Recent trends & Innovations in Artificial Intelligence based Applications," p. 6.

[6] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, "Big Data computing and clouds: Trends and future directions," Journal of Parallel and Distributed Computing, vol. 79–80, pp. 3–15, May 2015, doi: 10.1016/j.jpdc.2014.08.003.

[7] E. Olshannikova, A. Ometov, Y. Koucheryavy, and T. Olsson, "Visualizing Big Data with augmented and virtual reality: challenges and research agenda," Journal of Big Data, vol. 2, no. 1, Dec. 2015, doi: 10.1186/s40537-015-0031-2.

[8] S. P. Callahan, "Adaptive Visualization of Dynamic Unstructured Meshes," p. 2.

[9] D. Mashima, S. G. Kobourov, and Y. Hu, "Visualizing Dynamic Data with Maps," p. 8.

[10] E. R. Gansner, Y. Hu, and S. G. Kobourov, "GMap: Drawing Graphs as Maps," arXiv:0907.2585 [cs], Jul. 2009, Accessed: Jul. 19, 2020. [Online]. Available: http://arxiv.org/abs/0907.2585.

[11] M. Khan and S. S. Khan, "Data and information visualization methods, and interactive mechanisms: A survey," Int. J. Comput. Appl., vol. 34, no. 1, pp. 1–14, 2011.

[12] C. Eze, J. R. C. Nurse, and J. Happa, "Using Visualizations to Enhance Users' Understanding of App Activities on Android Devices," p. 19.

[13] D. P. Tegarden, "Business Information Visualization," vol. 1, p. 38.

[14] P. Hoffman and G. Grinstein, Visualizations for high dimensional data mining-table visualizations. Citeseer, 1997.

[15] G. Grinstein, M. Trutschl, and U. Cvek, "High-dimensional visualizations," in Proceedings of the Visual Data Mining Workshop, KDD, 2001, vol. 2, p. 120.

[16] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, "Big Data computing and clouds: Trends and future directions," J. Parallel Distrib. Comput., vol. 79–80, pp. 3–15, May 2015.

[17] N. A. Qureshi et al., “Application of Principal Component Analysis (PCA) to Medical Data,” Indian J. Sci. Technol., vol. 10, no. 20, pp. 1–9, Feb. 2017.

[18] L. N. Kova and O. S. T. Pánková, “Multidimensional clusters in RadViz,” p. 6.

[19] M. Schmuker et al., “SOMMER: self-organising maps for education and research,” J. Mol. Model., vol. 13, no. 1, pp. 225–228, Nov. 2006.

[20] G. K. Uyanık and N. Güler, “A Study on Multiple Linear Regression Analysis,” Procedia - Soc. Behav. Sci., vol. 106, pp. 234–240, Dec. 2013.

[21] F. Fechete, “Analysis of the Economic Performance of a Organization using Multiple Regression,” p. 6, 2014.

[22] E. Biglieri and K. Yao, “Some properties of singular value decomposition and their applications to digital signal processing,” Signal Process., vol. 18, no. 3, pp. 277–289, Nov. 1989.

[23] Dr. G. Singh and A. Kaur, “Comparative Analysis of K-Means and Kohonen-SOM data mining algorithms based on student behaviors in sharing information on facebook,” Int. J. Eng. Comput. Sci., Apr. 2017.

[24] H. S. Bhosale and D. P. Gadekar, “A Review Paper on Big Data and Hadoop,” Int. J. Sci. Res. Publ., vol. 4, no. 10, pp. 1–7, 2014.

[25] M. Tory and T. Moller, “Human factors in visualization research,” IEEE Trans. Vis. Comput. Graph., vol. 10, no. 1, pp. 72–84, Jan. 2004.

[26] S. Card, “Information Visualization,” p. 37.

SDCT: Multi-Dialects Corpus Classification for Saudi Tweets

Afnan Bayazed[1], Ola Torabah[2], Redha AlSulami[3], Dimah Alahmadi[4], Amal Babour[5], Kawther Saeedi[6]
Information Systems Department, King Abdulaziz University
Jeddah, Saudi Arabia

***Abstract*—There is an increasing demand for analyzing the contents of social media. However, the process of sentiment analysis in Arabic language especially Arabic dialects can be very complex and challenging. This paper presents details of collecting and constructing a classified corpus of 4180 multi-dialectal Saudi tweets (SDCT). The tweets were annotated manually by five native speakers in two stages. The first stage annotated the tweets as Hijazi, Najdi, and Eastern based on some Saudi regions. The second stage annotated the sentiment as positive, negative, and natural. The annotation process was evaluated using Kappa Score. The validation process used cross validation technique through eight baseline experiments for training different classifier models. The results present that the 10-folds validation provides greater accuracy than 5-folds across the eight experiments and the classification of the Eastern dialects achieved the best accuracy compared to the other dialects with an accuracy of 91.48%.**

Keywords*—Arabic dialects; dialects classification; language classification; natural language processing; Saudi dialects; sentiment analysis; Twitter

I. Introduction

Today, there are roughly 6500 spoken languages around the world, and each language involves different multiple dialects [1]. Arabic language is one of the most used languages in the world. Arabic is the official language of 22 countries, and it is spoken by over 400 million people. It is considered the fourth language used the most on the Internet [2]. There are three varieties of Arabic language which are Classical Arabic (CA), Modern Standard Arabic (MSA) and Arabic dialects (AD). The CA is a form of Arabic language used in literary texts and the Quran (Islam's Holy Book). The MSA is the essential Arabic form that is used commonly in formal conversations, media, education, newspapers, magazines, and formal TV programs. The AD is used in informal communication, and it is divided by geographical region [3]. The AD geographical regions are Egyptian, North Africa, Levantine, Iraqi, and Gulf [1]. However, the Gulf region consists of six countries: Saudi Arabia, United Arab Emirates, Qatar, Kuwait, Bahrain, and Oman, where each country has its own dialect. As for Saudi Arabia, also each different region has its own dialect. In Saudi Arabia, the dialects are Hijazi in the western region, Najdi in the Middle region, Southern dialect in the Southern region, Northern dialect in the Northern region, and Eastern dialect in eastern region. The AD has huge differences between them that can be considered different languages; therefore, Arabic language and its dialects required further intensive study and analysis. Most of Arabic Natural Language processing (NLP) applications are dedicated to the MSA like sentiment analysis, machine translation, speech recognition, and speech synthesis. Moreover, the Arabic NLP tools such as part-of-speech (POS) tagging, morphological analysis, and disambiguation are designed specifically for MSA, and for that, it gave a less accurate result for AD.

The Arabic NLP resources are focused on the MSA that has covered all orthographic varieties and have a rich morphology, and a strong syntactic system. As for the AD, the Arabic NLP resources do not cover it as well as the MSA. Furthermore, the AD is spoken languages with no writing system. Creating resources for the Arabic dialects is challenging in the Arabic NLP but it is necessary [4-6].

Particularly with the proliferation amounts of textual data on social media websites and microblogs, such as Twitter and Facebook, there is a huge resource for the Arabic dialects. Social media is an important communication tool for people to write about their daily life, share information, add reviews or opinions, explore the latest news and search for real-time news events. Arabic users tend to communicate with each other using the unstructured and ungrammatical slang Arabic language. Twitter is one of the world's most popular platforms for internet users. Twitter users send about 500 million tweets per day, where each tweet contains 280 characters [7]. The Arab people have been influenced by the recent evolution in technology. The total number of Arabic users on Twitter are more than 11 million, with 27.4 million tweets per day. The most active users are from Saudi Arabia with about 30% of all the tweets [8].

Al-Twairesh et al. in [9] claims that the lack of Arab corpora is one of the challenges facing a sentiment analysis of Arab. Accordingly, this research aims to utilize the huge Arabic textual data and prepare it as language resources for the Saudi dialects. This paper's contributions can be summarized as follows:

- Build Saudi Dialects Corpus from Twitter called SDCT and make it available as an open source for the research community.
- Classify SDCT depending on different Saudi dialects (Hijazi, Najdi and Eastern).
- Provide sentiment labelling of each dialect mostly to Positive, Negative and Neutral.

The paper is organized as follows: The previous related work is described in Section II. Section III explains the

methodology for creating the corpus and its annotation, includes a preview of the twitter corpus collection and Saudi tweet analyzes, and discusses the experimental findings. Section IV presents the challenges of this research. Finally, the conclusion and future recommendations are shown in section V.

II. Related Works

For providing a comprehensive overview, we survey the related works available in Arabic corpora and dataset. Several studies have proposed number of approaches in the Arabic dialects classification. Also, there are enormous studies conducted in sentiment analysis for Arabic dialects.

For the purpose of creating frameworks for sentiment analysis, Duwairi et al. in [10] developed a framework for sentiment analysis on Arabic Tweets in text reviews. They used a translated version of English lexicon called SentiStrength and extended it with synonyms list for every word in the lexicon as a seed list. The polarity for each word in the seed list is expressed as -1 for negative sentiment and 1 for positive sentiment. They used a set of 4400 Arabic tweets, where each tweet was tokenized into terms and the sentiment of the tweet was determined by summing the scores of all the terms in the tweets where the sentiment of the tweet was considered positive if its summation is greater than 0, negative if its summation is less than zero, and neural if the summation equals zero. For the performance of the proposed framework, they applied two experiment without and with stemming the tweets. The results showed that the framework achieved good results and improved the precision, recall, accuracy and reduced error rate with stemmed tweets. Duwairi et al. in [11] proposed a framework for Arabic text sentiment analysis based on a created crowdsourcing API to manually annotate a training dataset of 25000+ tweets as positive, negative, or neutral. To test the performance of the framework, they used Rapidminer built-in classifiers named Naïve Bayes (NB), k-nearest classifier (KNN), and Support Vector Machines (SVM) on a stratified sample of 1000 tweets from the training dataset. For each classifier, the applied two experiments without and with stopwords/ stemming. The result showed that the best accuracy was achieved by SVM when no stopwords and no stemming were used.

For creating corpus of Arabic sentiment analysis, Refaee and Rieser in [12] constructed a corpus supporting Subjectivity and sentiment analysis (SSA) and collected 8,868 Arabic twitter feeds from multiple Arab dialects. Then, manual annotation processes have been performed by two annotators to polar, positive, negative, neutral and mixed. Furthermore, Nabil et al. in [13] presented an Arabic social sentiment analysis dataset (ASTD) consisting of 10K tweets that were manually annotated using Amazon Mechanical Turk (AMT) to objective, subjective positive, subjective negative, and subjective mixed. Assiri et al. in [14] created the first Saudi annotated corpus collected randomly from Twitter trending hashtags promotion in Saudi Arabia. In accordance with the preprocessing and manual annotation, they collected around 4700 tweets. The dataset has manually annotated based on the sentiment text polarity as positive, negative, and neutral using an application user interface. Furthermore, there is a similar effort in [15] where Altwairesh et al. presented and produced a comprehensive corpus of 18K tweets by using a specific annotation system.

For creating corpus of Arabic dialects and language, Alshutayri and Atwell in [16] built a corpus of 13.8M collected from Twitter, newspapers, and Facebook. The corpus was annotated into five different dialects: Egyptian, Gulf, Levantine, Iraqi, and North African. For the annotation process, they developed an online game via a website where players can involve in the annotation by classifying their dialects. Likewise, Mubarak and Darwish in [17] created a large corpus of Arabic dialects collected from the Twitter platform. The size of the corpus is 175M. The corpus was annotated into six different dialects: Saudi Arabian, Egyptian, Algerian, Iraqi, Lebanese and Syrian. For the annotation process, native speakers of each dialect have involved in determining if a tweet belongs to their dialects or not. Alshutayri and Atwell in [1] created a corpus of 210,915K tweets containing five Arabic dialects: Gulf, Iraqi, Egyptian, Levantine, and North African. For the annotation process, they used Waikato Environment for Knowledge Analysis (WEKA). Altamimi et al. in [18] created the corpus containing 122K tweets for Arabic dialects collected from twitter. Tweets were annotated manually into five labels: Gulf; Egyptian; Levantine; Maghrebi; and Iraqi; in addition to Modern Standard Arabic (MSA) and Classical Arabic (CA). Maghfour et al. in [19] centered their study on classified the Facebook comments as expressed in MSA or in Moroccan Dialect besides the Sentiment analysis (SA) classification in comments. Hence, they performed two different schemas. The first one is a classical schema that considered all Arabic dialects and languages as homogeneous thus, then they applied sentiment analysis on the collected dataset at once. In the second schemas, they proposed to classify the Arabic language into two sub-dataset MSA and Dialect Arabic (DA) beforehand sentiment analysis. Therefore, they applied different preprocessing and Arabic dialect stemmers on each sub-dataset. In supervised classification, they employed the most two reported sentiment classification algorithms, Naive Bayes (NB) and Support Vector Machine (SVM). In the testing phase, the four combinations of weighting schemes n-gram and extraction schemes have been utilized. This study has recorded a high score in the classical schema with the NB algorithm. The Similar effort presented by Medhaffar et al. [20] where they developed the Tunisian dialect dataset that composited 17K comments collected from Facebook. They applied three classification algorithms SVM, NB, and multi-layer perceptron MLP. Their models have shown better accuracy than other models that trained on MSA.

Accordingly, there are enormous studies that provide a public Arabic dialects' lexicon and corpus to address complexity and difference in Arabic language and its dialects. Furthermore, some of the researches have specialized to study sentiment analysis in Saudi dialects text aligns with the increasing demand to analyze social media content in the Saudi market. However, the contributions related to Saudi dialects are still insufficient and limited. In addition, there is no corpus that classified the different Saudi dialects according to the regions. Besides no sufficient research that study sentiment

analysis in each specific classified Saudi dialects. Therefore, we seek to center our work in creating a public corpus of sentiment analysis and classification of Saudi dialects.

III. Experiment

This section illustrates our approach for building a SDCT corpus that is dedicated to Saudi dialects. First, we collected the data and targeted Twitter as the main source of data collection. Then we conducted the preprocessing phase that involved three main tasks, which are data cleaning, normalization, and lemmatization. After the data collection and preprocessing, the data were manually annotated, hence the annotation was evaluated using Kappa Score [24]. In addition, we extracted the features to be used in the training set. Then the classification was conducted by the classifiers and, finally, we validated the classifiers via the cross-validation technique. Fig. 1 illustrates the proposed approach for building the SDCT corpus.

A. Data Collection

This research aims to build a SDCT Corpus for Saudi dialects collected from the social network application Twitter. Initially, we planned to cover all five of the dialects in Saudi Arabia: Hijazi, Najdi, Southern, Northern, and Eastern. However, according to our study of tweets in Saudi Arabia we observe that the southern and northern dialects are not widely used, and the number of tweets in these dialects are scarce compared to the Najdi, Hijazi and Eastern dialects. Hence, we limited our experiment and focused only on the three main dialects: Hijazi, Najdi, and Eastern, which are mostly used on Twitter in Saudi Arabia. In the data collection process, we used Twitter API for developers, which is provided by Twitter to allow access to their social media content, and 'Tweepy', a Python library for retrieving tweets. Retrieved tweets were stored in the CSV format and, therefore, could be accessed using an Excel spreadsheet. Furthermore, we retrieved approximately 8923 tweets and reached a total of 4181 after the cleaning process, as illustrated in Table I. In addition, we mainly relied on the terms used particularly in each specific dialect and used both the time zone of the dialect region and trending hashtags, as illustrated in Table II. The collection process was accomplished in around two weeks, from March 15th, 2020 to March 28th, 2020. In addition, the corpus is available upon any request from the authors for research and testing.

B. Data Preprocessing

The preprocessing phase is one of the important steps in text mining. It prepares the raw text for the next phase by removing unwanted or annoying data and reduces the dimensionality size of text data as well as normalizing the text.

Therefore, in SDCT corpus, we divided the data preprocessing into three phases, which are the data cleaning phase, normalization phase, and finally, the lemmatization phase. Firstly, we manually removed Ads tweets in the data cleaning phase, the tweets that are not related to any of the dialects that we identified such as Lavanteen or EGY dialects, and unhelpful short tweets, such as تمامwhich means "OK", and كيف الحالwhich means "How are you". Then we automatically removed by coding noise data, such as links (http://, https://), emoji, mentions (@Username), retweets, hashtags as (#corona , #كورونا), and punctuation (!@#$%&^ *()_ +<>?:.;-{}c,c) from our SDCT corpus. Secondly, in the normalization phase, we used the Tashaphyne library for the normalization process [21]. This included normalizing letters, such as Alef (أ ، ا ، إ ، آ), Hamza (ئ ، ؤ ، ء), Ya'a (ي ، ى) and Ha'a (ه ، ة), strip repeated letters, elongation (Tatweel), and diacritics (Tashkeel and Harakat). The normalization process is illustrated in Table III.

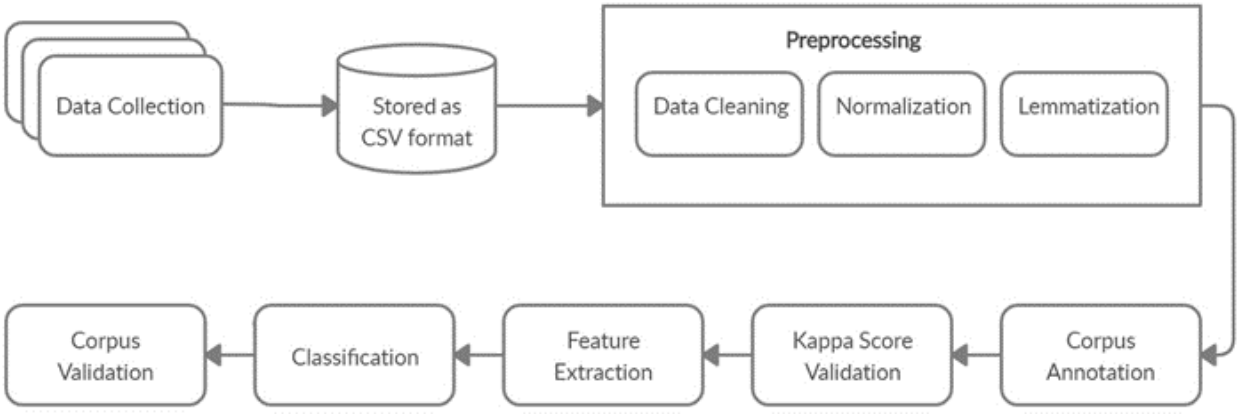


Fig. 1. The Proposed Approach for Building a SDCT Corpus.

TABLE I. This Size of Collected Tweets before and after the Cleaning Process

Prosperities		**Data Collection**		**After Cleaning**	
Dialect	***Collection Process***	***Size***	***Total***	***Size***	***Total***
Hijazi	Time Zone, Keyword	3543	8923	1507	4181
Najdi	Trending Hashtag, Keyword	3450		1341	
Eastern	Keyword	1930		1333	

TABLE II. EXAMPLES OF TWEETS RETRIEVED BY USING THE TIME ZONE, KEYWORDS AND TRENDING HASHTAGS

Dialects	Keyword, Time zone, Hashtag	Example (Arabic)	Example (English Translation)
Hijazi	دحين "Daheen" أهرج "Ahrej" مرة "Marra" كمان "Caman"	دحين كيف مصدقة نفسي اني بقوم الساعة 8 االهرج ماعليك نسمعك اول مرة اشوف الكلمة أصلاً اشتقت للدوام انا كمان	Now, how do I believe myself that I'm going to wake up at 8 Speak, do not worry we hear you Basically, it's the first time I see the word I missed work too
Najdi	عيا" Aiyah" هالحين "Halheen" وش "Wesh" امحق " Emhag"	عطني رابطه تكفا حسابي عيا يفتح يقول حمل الجديد والله كنا عايشين احسن من هالحين والدنيا سهالات وش دا الهجوم لا تدعي على احد كريمة بحق غيري امحق كرم	Give me its link please, my account can't open, it asks to download the new version We were living better than now, and the world was easy What is this attack, do not claim anyone Generous to others, great generosity
Eastern	وايد "waeed" اشفيج"eshfej" شنو "shno" صج "sij"	احسك تحب كرة القدم وايد شفيج معصبه عيوني ترا نمزح شنو اسم اللعبة هذا اللي اشوفه صج ولا جرافيكس	I feel you love football very much Why are you angry we are joking What is the name of the game? This is what I see is true or graphics?

TABLE III. NORMALIZATION PROCESS

Example	Replaced By	Process Name
صدممممتتنننييي	صدمتني	Strip repeated letters
ســــــــلام	سلام	Strip elongation (Tatweel)
ضيّقت	ضيقت	Strip diacritics (Tashkeel and Harakat)
أإآا	ا	Normalizing Alef
ي ى	ي	Normalizing Ya'a
ة ه	ه	Normalizing Ha'a
ئ ؤ ء	ء	Normalizing Hamza

Lastly, we preferred to use Farasa tool, one of the NLP and morphological tools for the lemmatization process. This is a fast and accurate new Arabic segmenter that uses SVM for ranking and is proposed by [22]. Table IV illustrates some examples of lemmatization by using Farasa. Furthermore, Farasa had been used by some previous researches, where it made fewer stemming errors. Some examples of Preprocessing phases are shown in Fig. 2.

TABLE IV. EXAMPLE OF LEMMATIZATION PROCESS USING FARASA

Word	Lemma	English Translation
كتبنا	كتب	Wrote
يشتكون	اشتكى	Complained
قلنا	قال	Said
توقعنا	توقع	Expected
مكتبئة	مكتئب	Depressed

Fig. 2. Examples of Preprocessing Phases.

C. Annotation Methodology

Annotation is the process of adding text for labels that rely on both the classification output and sentiment [15, 23]. In this study, we inducted five annotators who graduated from King Abdulaziz University and are considered to be native speakers of Saudi Arabian. Hence, in the first stage, four of the annotators manually annotated 4,181 tweets corpus for both dialects and sentiment. Firstly, every annotator classified each tweet dialect into one of the following four labels: Hijazi, Najdi, Eastern, and Saudi dialect (SA) for the unknown or the mixed dialect tweets. After that, each annotator labeled the same tweet as being positive, negative, or neutral. This was based on their impression of the tweet and what sentiment they felt it expressed.

The authors provided the annotators with guidelines regarding the annotation process to help them classify the dialects and sentiments clearly and easily. These depended on the conditions for each label. The annotators followed these guidelines when they classified the dialects:

1) The dialects were classified according to the way that the words were written, replacing the writing, and pronouncing the letter in another way. For example:

- In the Hijazi dialect, the letter "zal,ذ " changes to "dal, د" and the letter "The,ث " changes to "T,ت ".
- In the Najdi dialect, the letter "CAF,ك ", changes to "TS,تس ".
- In the Eastern dialect, the letter "CAF, ك", changes to "j, ج".

2) The classification of the dialects was based on the most utilized words in a specific dialect and was associated with a specific dialect. For example:

- In the Hijazi dialect, "Dahin, دحين ", means now.
- In Najdi dialect, "Halhin, هالحين", means now.
- In the Eastern dialect, "Waid, وايد", means much.

3) If the tweet was difficult to classify or contained words that belonged to two or more dialects, it was classified as unknown or mixed dialect tweets.

4) Symbols that should be used when classifying the dialects (Hijazi: hj, Najdi: nj, Eastern: ea, White dialect: sa).

The following guidelines were carried out when categorizing the sentiment:

1) A tweet was positive if the opinion clearly indicated praise, joy, happiness, and any happy emojis.

2) A tweet was negative if the opinion clearly indicated defamation, sadness, anger, disgust, or any sad emojis.

3) A tweet was neutral if it was not positive or negative, such as news, supplications, or general speech.

Thereafter, some examples of the annotation process are shown in Fig. 3. Moreover, we noted the opinions the annotators gave regarding the dialect annotation, as shown in Table V. For the sentiment analysis, there were three different polarities of the entire dataset as shown in Table VI.

Furthermore, to get the final annotation for both the dialects and sentiments we gathered the more frequently used labels from the annotators for each tweet by using the Mode equation. When there was conflict regarding the annotation, which occurred in 27 tweets, we resolved it by taking the opinion of the fifth annotator as the second stage at annotation process. In case the final annotation was labeled as SA, which happened in only one tweet. We decided to eliminate this tweet and the number of SDCT becomes 4180.

After the percentage of sentiment labels for each dialect had been calculated by observing the annotation process on the corpus. We found that the polarity of neutral dominated in the Saudi tweets, followed by negative and, lastly, positive as shown in Fig. 4. We believe the coronavirus situation is the reason why there was an increase in negative polarity, compared to positive, as we observed that a lot of tweets were related to coronavirus during our study of the tweets.

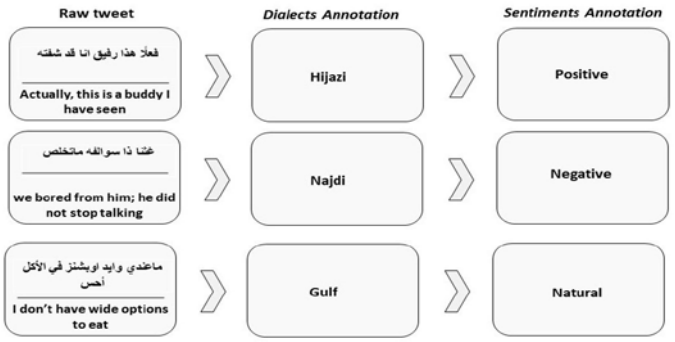


Fig. 3. Examples of Annotations.

TABLE V. ANNOTATION FOR DIALECT LABELS

Label	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Final annotation
Hijazi	1399	1475	1460	1399	1506
Najdi	1369	1353	1351	1341	1341
Eastern	1333	1325	1334	1337	1333
SA	80	28	36	104	1
Total	4181	4181	4181	4181	4181

TABLE VI. ANNOTATION FOR THE SENTIMENT LABELS

Label	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Final annotation
Negative	913	921	890	898	925
Positive	787	794	761	770	771
Neutral	2481	2466	2530	2513	2485
Total	4181	4181	4181	4181	4181

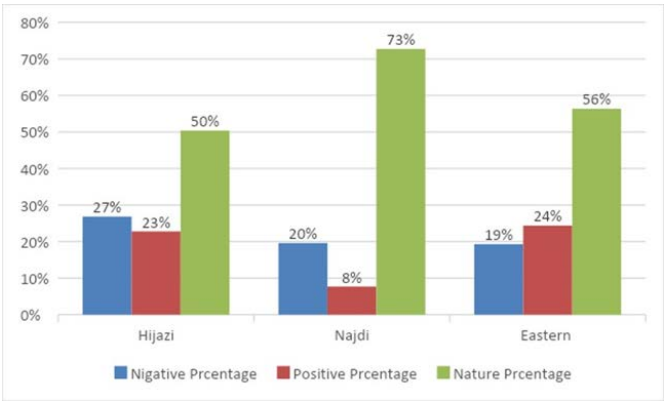


Fig. 4. The Sentiment Labels for each Dialect.

We extracted two hundred words that had been most frequently used in each dialect in the SDTC corpus. This was done graphically as a word cloud of dialect, as shown in Fig. 5, 6 and 7. The font size indicates how frequently the word was used in whole tweets of a specific dialect [24]. Fig. 5 presents the most frequent words in Hijazi Dialect in a word of cloud style. For example, the word "Dahin, دحين "means now and the word "Caman, كمان " means also, they have bigger font in the word cloud that indicates more frequents appearance in the corpus. While, Fig. 6 presents the Najdi Dialect, the word "Ayya, عيا"means refuse, is very common use. From Fig. 7, the Eastern Dialect word cloud shows that the word "Wayed, وايد "means a lot with bigger font to show the frequent used of it.

To ensure the reliability of the results the annotators agreement had to be validated. If the annotators allocated similar labels, we concluded that they all comprehended the annotation instructions in a similar way and would be consistent in their results. The reliability assessments were done to evaluate the level of trust there was in the guidelines and annotation schemes. We used Kappa to calculate the inter-annotator agreements between the four annotators in order to evaluate the quality of the annotations. We calculated the Kappa coefficient for 4,181 tweets that were annotated by the four annotators. The results showed that Kappa was obtained at 0.9382 for the dialect labels, which could be considered almost perfect agreement. The result for the sentiment labels was 0.3199, which indicated fair agreement. Calculation of Kappa coefficient and its interpretation were based on the well-known reliability equations and measurements presented in [25].

D. Validation Test

In this work, two different ways of classifying the dialect were conducted. The aim was to use them to evaluate the accuracy of the SDCT corpus. The first method used all the collected tweets that classified the dialects as Hijazi, Najdi, or

Eastern. The second method established three sub-datasets that had been extracted from the SDCT corpus. Each sub-dataset handled a specific dialect in order to train the model on the characteristics of a particular dialect. Hence, the three sub-datasets were named Hijazi-Dataset, Najdi-Dataset and Eastern-Dataset. The tweets in each sub-dataset were classified as the dialect-name and not-dialect-name. The Hijazi-Dataset contained all the Hijazi tweets in the SDCT corpus that were classified as Hijazi, and Not-Hijazi tweets that had an equal balance of Najdi and Eastern tweets. There was a similar approach with the Najdi-Dataset, which included all the Najdi tweets and an equivalent number of Not-Najdi Tweets. While the Eastern-Dataset was classified as Eastern tweets and Not-Eastern tweets. Table VII illustrates the number of tweets in each dataset.

Fig. 5. Word Cloud of Hijazi Dialect.

Fig. 6. Word Cloud of Najdi Dialect.

Fig. 7. Word Cloud of Eastern Dialect.

TABLE VII. DATASETS OF THE CLASSIFICATION TEST

Dataset	Hijazi	Najdi	Eastern	Not-Dialect	Total
SDCT	1506	1341	1333	-	4180
Hijazi-Dataset	1506	700	700	Not-Hijazi = 1400	2906
Najdi-Dataset	600	1341	600	Not-Najdi = 1200	2541
Eastern-Dataset	650	650	1333	Not-Eastern = 1300	2633

To extract the features, we combined the multi-configurations of the weighting schemes (TF-IDF) and extraction schemes (n-gram), where we tested unigrams, bigrams, trigrams with TF-IDF to present the features vectors. By using these features we trained the most popular classification models, which were named as follows: Linear Support Vector Machines (LSVM), Radial Basis Function Support Vector Machines (RBF SVM), k-Nearest-Neighbors (K-NN), Naive Bayes (NB), Logistic Regression (LR), Gradient Boosting (GB), Random Forest (RF), AdaBoost, Decision Tree (DT), Bernoulli Naive Bayes (BNB), and Stochastic Gradient Classifier (SGC).

This work employed the k-fold cross-validation technique for training classifier models on each of the four datasets. Cross-validation was used to avoid cases of over-fitting and under-fitting and get a better prediction accuracy. K-folds split the dataset randomly to *k* of splitting. This was done by performing loops through the entire data to apply the classification model and getting an accuracy average. Hence, we proposed using two k-folds, 5-folds and 10-folds, that were best practices. Therefore, we performed eight experiments, including two experiments performed on each dataset. Using the first method of dialect classification, we implemented two main experiments (5-folds and 10-folds) on the SDCT dataset. Table VIII shows the experiments' accuracy. In general, we noticed that the accuracy of 10-folds was better than 5-folds. The highest recorded accuracy of 5-folds in the BNB model was 57.36%. This was also the closest result with DT, GB with bigram and trigram, and SGC with unigram. The worst result was shown to be the k-NN models. For 10-folds, the best result was 69.73% of the SGC model with bigram.

TABLE VIII. THE RESULTS OF THE CLASSIFICATION MODELS IN THE SDCT DATASET

	5-Folds			10-Folds		
Model Name	***unigram***	***bigram***	***trigram***	***unigram***	***bigram***	***trigram***
LSVM	55.36	55.4	55.4	68.82	68.87	68.87
RBF SVM	53.75	53.73	53.73	67.53	67.48	67.48
k-NN	41.5	41.57	41.57	50.69	50.69	50.69
NB	49.33	49.33	49.33	62.77	62.77	62.77
LR	55.52	55.52	55.52	69.04	69.06	69.06
GB	56.94	57.05	57.05	69.35	69.3	69.3
RF	58.39	56.7	56.94	69.28	70.26	70.09
AdaBoost	55.02	53.85	53.92	65.23	65.4	65.4
DT	57.13	57.25	57.15	69.13	69.56	69.56
BNB	57.36	57.36	57.36	68.63	68.68	68.68
SGC	57.03	56.39	56.94	69.57	69.73	69.33
Average	54.30	54.01	54.08	66.37	66.52	66.47

The remaining experiments were performed on the three datasets that had been constructed for the second method of classification. Table IX shows the effect of n-gram with 11 different models on the Hijazi-Dataset with the application of cross-validation. As we can see, AdaBoost had the highest accuracy in both 5-folds and 10-folds. Regarding the Najdi-Dataset, AdaBoost with n-gram = 2,3 achieved the best accuracy of 61.77%. This was done by applying 5-folds. The DT achieved a good result with 77.91% of bigram and a 10-folds configuration, as shown in Table X. Table XI shows the accuracy of the Eastern -Dataset, which had an excellent accuracy of 90% in the SGC model with 5-folds. Regarding 10-folds, the SGC and LSVM resulted in a perfect accuracy of 91%.

TABLE IX. THE RESULTS OF THE CLASSIFICATION MODELS IN THE HIJAZI DATASET

	5-Folds			**10-Folds**		
Model Name	***unigram***	***bigram***	***trigram***	***unigram***	***bigram***	***trigram***
LSVM	71.11	71 .18	71.18	80.21	80.14	80.14
RBF SVM	68.74	68.6	68.6	78.53	78.22	78.22
k-NN	55.03	54.75	54.75	65.4	65.4	65.4
NB	68.6	68.53	68.53	78.02	78.19	78.19
LR	70.32	70.25	70.25	80.39	80.25	80.25
GB	69.57	70.39	70.39	81.73	82.12	82.12
RF	69.91	70.29	68.29	82.48	82.03	83.44
AdaBoost	72.46	72.73	72.73	84.72	84.61	84.61
DT	67.12	67.36	67.67	82.41	82.72	82.52
BNB	66.36	65.98	65.98	83.2	83.07	83.07
SGC	72.04	72.42	72.15	81.35	81.14	81.28
Average	68.29	68.40	68.22	79.85	79.80	79.92

TABLE X. THE RESULTS OF THE CLASSIFICATION MODELS IN THE NAJDI DATASET

	5-Folds			**10-Folds**		
Model Name	***unigram***	***bigram***	***trigram***	***unigram***	***bigram***	***trigram***
LSVM	53.62	53.46	53.46	69.76	69.88	69.88
RBF SVM	53.27	53.19	53.19	72.28	72.17	72.17
k-NN	56.93	56.54	56.54	64.09	63.78	63.78
NB	57.4	57.17	57.17	68.62	68.58	68.58
LR	55.47	55.43	55.43	71.38	71.42	71.42
GB	56.61	57.4	57.4	68.54	68.5	68.5
RF	59.37	60.24	59.33	74.13	74.49	75.51
AdaBoost	61.22	61.77	61.77	74.29	74.53	74.53
DT	60.87	60.39	60.79	77.28	77.91	77.81
BNB	59.84	59.45	59.45	72.01	72.24	72.24
SGC	56.85	57.01	57.52	73.7	74.02	74.41
Average	57.40	57.45	57.45	71.46	71.59	71.71

TABLE XI. THE RESULTS OF THE CLASSIFICATION MODELS IN THE EASTERN DATASET

	5-Folds			**10-Folds**		
Model Name	***unigram***	***bigram***	***trigram***	***unigram***	***bigram***	***trigram***
LSVM	88.48	88.52	88.52	91.36	91.36	91.36
RBF SVM	78.06	78.02	78.02	87.84	87.88	87.88
k-NN	60.09	60.16	60.16	61.67	61.74	61.74
NB	63.05	63.28	63.28	69.45	69.71	69.71
LR	74.96	74.81	74.81	86.66	86.7	86.7
GB	89.58	89.58	89.58	90.64	90.64	90.64
RF	89.24	89.5	89.51	89.81	89.92	89.96
AdaBoost	88.03	87.38	87.38	87.72	81.09	87.69
DT	89.96	90.19	90.36	89.47	89.28	89.39
BNB	73.48	74.61	74.61	81.17	81.09	81.09
SGC	90.34	90.68	90.53	91.44	91.48	91.29
Average	80.47	80.61	80.60	84.29	83.71	84.31

In summary, as we mentioned previously, the 10-folds provided greater accuracy than 5-folds across all eight of the experiments. Furthermore, we observed that, on average, the results of unigram, bigram and trigram were close to each other, particularly in short texts such as tweets. The Eastern-Dataset achieved the best result in this paper compared to the other three datasets with an excellent accuracy of 91.48%.

IV. RESEARCH CHALLENGES

Due to the complex nature of the Arabic language, more investigation is needed, especially in the text mining tools that support the Arabic language. We have encountered a number of obstacles and challenges that need to be taken into account in future works. Some of the obstacles that we faced through the different phases of this study were as follows:

- Data Collected: collecting the tweets that were associated with a particular Saudi dialect was not an easy phase, as most Saudis use the general dialect. In addition, the terms used had significant similarities. Furthermore, the Tweepy library has a limitation when tweets older than one week are retrieved. Hence, a massive effort is required to find the most used unique terms in each region. In addition, a lot of the tweets are advertisements and so it takes a long time in the first cleaning phase to filter them manually.
- Data Preprocessing: There is a limited number of libraries that specialize in Arabic text normalization. This area needs to be highlighted and developed in future works. We believe an excellent specialized library could be an important contribution to facilitating the preprocessing phase.
- Lemmatization: Arabic lemmatization libraries and tools need to be improved as there are insufficient libraries that handle the words in Arabic dialects. As mentioned previously in section III, some lemma roots are completely different, and this impacts their overall

meaning. In addition, some libraries don't provide a Python version, which is the most used language in Machine Learning ML.

- Annotation: The similarities between the different dialects means the annotators found it difficult to label some tweets as being in a specific dialect. During the sentiment analysis, some of the tweets were difficult to annotate as being either positive or negative when there was some ambiguity.

V. Conclusion

The objective of this paper was to enrich Arabic, particularly the language used in Saudi Arabia, by constructing a Saudi corpus based on dialects, and make it available for further research in Arabic studies such as NLP applications. This paper presented the methodology used to collect and build a corpus of 4180 multi-dialectal Saudi tweets (SDCT). The corpus was collected by using different keywords, hashtags, and time zones. It was manually annotated into Saudi dialects as Hijazi, Najdi, and Eastern and sentiment as positive, negative, and natural by five native speakers using specific explained guidelines. Cohen's Kappa Coefficient was used to calculate the reliability of the annotations. Eight baseline experiments were performed by using different classifier models with various features and configuration vectors. Four datasets of the corpus were established to fulfill the evaluation of the SDCT corpus that employed a cross-validation mechanism. Further work will be carried out to expand the corpus using other sources for Saudi dialects as well as improve the accuracy of the experiment including various text-features and other factors.

References

[1] Alshutayri and E. Atwell, Exploring Twitter as a source of an Arabic dialect corpus. International Journal of Computational Linguistics (IJCL), 2017. 8(2): p. 37-44.

[2] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, Arabic natural language processing: An overview. Journal of King Saud University-Computer and Information Sciences, 2019.

[3] N. Al-Twairesh, R. Al-Matham, N. Madi, N. Almugren, A.-H. Al-Aljmi, S. Alshalan, R. Alshalan, N. Alrumayyan, S. Al-Manea, and S. Bawazeer, Suar: Towards building a corpus for the Saudi dialect. Procedia computer science, 2018. 142: p. 72-82.

[4] R. Baly, A. Khaddaj, H. Hajj, W. El-Hajj, and K. B. Shaban, Arsentd-lev: A multi-topic corpus for target-based sentiment analysis in arabic levantine tweets. arXiv preprint arXiv:1906.01830, 2019.

[5] S. Harrat, K. Meftouh, and K. Smaïli. Creating parallel Arabic dialect corpus: pitfalls to avoid. 2017.

[6] D. Alahmadi, A. Babour, K. Saeedi, and A. Visvizi, Ensuring Inclusion and Diversity in Research and Research Output: A Case for a Language-Sensitive NLP Crowdsourcing Platform. Applied Sciences, 2020. 10(18): p. 6216.

[7] A. B. Boot, E. T. K. Sang, K. Dijkstra, and R. A. Zwaan, How character limit affects language usage in tweets. Palgrave Communications, 2019. 5(1): p. 1-13.

[8] M. Alruily, Issues of dialectal saudi twitter corpus. Int. Arab J. Inf. Technol., 2020. 17(3): p. 367-374.

[9] N. Al-Twairesh, H. Al-Khalifa, and A. Al-Salman. Subjectivity and sentiment analysis of Arabic: trends and challenges. in 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA). 2014. IEEE.

[10] R. M. Duwairi, N. A. Ahmed, and S. Y. Al-Rifai, Detecting sentiment embedded in Arabic social media–a lexicon-based approach. Journal of Intelligent & Fuzzy Systems, 2015. 29(1): p. 107-117.

[11] R. M. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat. Sentiment analysis in arabic tweets. in 2014 5th International Conference on Information and Communication Systems (ICICS). 2014. IEEE.

[12] E. Refaee and V. Rieser. An arabic twitter corpus for subjectivity and sentiment analysis. in LREC. 2014.

[13] M. Nabil, M. Aly, and A. Atiya. Astd: Arabic sentiment tweets dataset. in Proceedings of the 2015 conference on empirical methods in natural language processing. 2015.

[14] A. Assiri, A. Emam, and H. Al-Dossari, Saudi twitter corpus for sentiment analysis. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 2016. 10(2): p. 272-275.

[15] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. Procedia Computer Science, 2017. 117: p. 63-72.

[16] A. Alshutayri and E. Atwell. Creating an Arabic dialect text corpus by exploring Twitter, Facebook, and online newspapers. in OSACT 3 Proceedings. 2018. LREC.

[17] H. Mubarak and K. Darwish. Using Twitter to collect a multi-dialectal corpus of Arabic. in Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP). 2014.

[18] M. Altamimi and W. J. Teahan, Arabic Dialect Identification of Twitter Text Using PPM Compression. International Journal of Computational Linguistics (IJCL), 2019. 10(4): p. 47-59.

[19] M. Maghfour and A. Elouardighi. Standard and dialectal Arabic text classification for sentiment analysis. in International Conference on Model and Data Engineering. 2018. Springer.

[20] S. Mdhaffar, F. Bougares, Y. Esteve, and L. Hadrich-Belguith. Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. 2017.

[21] Tashaphyne Arabic Light Stemmer. Available online: https://pypi.org/project/ Tashaphyne/. Accessed: Oct. 5, 2020.

[22] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak. Farasa: A fast and furious segmenter for arabic. in Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations. 2016.

[23] M. Altamimi, O. Alruwaili, and W. J. Teahan. BTAC: A Twitter Corpus for Arabic Dialect Identification. in of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-corpora 2018). 2018.

[24] A. L. Uitdenbogerd, World cloud: A prototype data choralification of text documents. Journal of New Music Research, 2019. 48(3): p. 253-263.

[25] J. L. Fleiss and J. Cohen, The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and psychological measurement, 1973. 33(3): p. 613-619.

Using Interdependencies for the Prioritization and Reprioritization of Requirements in Incremental Development

Aryaf Al-Adwan[1]
Department of Computer and Networks Engineering
Faculty of Engineering Technology
Al-Balqa Applied University
Amman, Jordan

An'aam Aladwan[2]
Department of Management Information Systems
Al-Ahliyya Amman University
Amman, Jordan

***Abstract*—There is a growing trend to develop and deliver the software in an incremental manner; to achieve greater consistency in the developed software and better customer satisfaction during the requirement engineering process. Some of the developed increments in the incremental model will be delivered to consumers and run in their environments, so a set of these requirements are evaluated, introduced, and delivered as the first increment. Other requirements are delivered as the next step and so on for the next increment. The priority of requirements plays an important role in each increment, but it is precluded by the interdependences between the requirements and resources constraints. Therefore, this paper introduces a model for requirements prioritization and a reprioritization based on these important factors. The first one is the requirement interdependencies which are described as a hybrid approach of tractability list and directed acyclic graph, and the second factor is the constraints of the requirements resources that are used based on the queuing theory for requirements reprioritization. In order to achieve this, two algorithms namely; Priority Dependency Graph (PDG) and Resources Constraints Reprioritization (RCR), were proposed with a linear time complexity and implemented via a case study.**

Keywords*—Requirement engineering; incremental model; requirement prioritization; requirement interdependencies; dependency graph; queuing theory

I. Introduction

In the incremental software model, small releases of the software are implemented in a series mode instead of providing the whole system after a long period of development. This model efficiently impacts the prioritization of requirements in such a way that the most relevant requirements can be introduced in the system's first releases. On the other hand, later phases are left with less important requirements. When requirements are elicited, a large number of them are often created, which is very difficult to implement them at the same time [1]. This is due to the market impact, the user persisting to have the software finished, and the limitations on cost and staff. Therefore, the requirements need to be prioritized in such a way that the earliest product releases meet the most critical ones, particularly when an incremental model is used, where the product is designed, implemented, and tested incrementally until the product is completed.

The requirements affect each other and are related to each other during software development in a way that prevents treating them separately. This is referred to as the interdependence between requirements. Consequently, requirement interdependency concerns about the relationships between requirements which during software development will influence decisions and activities. This play an important role in the prioritization of requirements, especially with incremental development which requires a careful selection of requirements that meet the growth of the various increments. Choosing one requirement may therefore activate the selection of several other requirements that rely on it.

On the other hand, the literature that discusses the interdependence requirements, limited work has been carried out. Interdependence of requirements is a special form of traceability of requirements that defines the relationships between different requirements. The traceability list, which is a table of relationships describing the dependencies between requirements [1], is one of the techniques for representing requirement interdependencies.

Carlshamre [2] used the directed graph (digraph) to represent the interdependencies between requirements, as well as classifying the interdependencies into five relationships; and, REQUIRES, TEMPORAL, CVALUE, ICOST, and OR, which visualized later by the directed graph. The concern in this approach wasn't to visualize types of interdependencies rather than representing them as dependency graph and apply an algorithm to prioritize the requirements. Another way to represent the interdependencies is to use ontology-based representation and a formal graphical representation to visualize the requirements interdependencies in a proper way [3].

It is possible to prioritize requirements, taking into account several different aspects, such as importance, cost, penalty, time, risk and dependencies [4]. The literature is full of several prioritization strategies for requirements. These include the process of analytical hierarchy (AHP) which is the most common priority-based technique that is designed to permit decision-makers to set priorities and decide the correct decision. Initially, AHP specifies the parameters and substitutes for each requirement and uses them to construct a

hierarchy to activate pair-wise comparisons; then the users can determine their favorites for each pair of attributes by assigning a decision scale. However, this technique requires a quadratic time to prioritize the requirements and suffers from scalability issues particularly when the number of requirements increases [5]. The creation of binary search tree, in which each requirement is shown in a node. The tree needs to be prioritized; the low priority requirements are set on the left side of the tree and high priority needs are positioned on the right. Although this method is fast but it the comparison of BST is typical, only showing which requirement is more desirable [6]. Another technique is the bubble sort in which the principle is similar to AHP, where they both make use of the comparison operation pair-wise and they require a time complexity of n^2. The distinction between them is that it is only possible for the decision-maker to consider which requirement is more significant between the compared requirements in bubble sort. [7]. Cumulative voting or the 100-dollar test is a straight-forward process that gives the system's stakeholders 100 units to be divided between requirements. The higher unit requirement has a higher priority and the lower unit requirement has a lower priority. The stakeholder controls distribution process of these units based on the priority of the requirements. However, if there are quite several requirements, this approach has a downside, because this method will not work well and will calculate the prioritization in wrong way. Also, it can be difficult to be aware the quantity of units that must be allocated and those that must be left [8-9]. Spanning trees technique is similar to AHP, where they both make use of the comparison operation pair-wise, but uses the minimum spanning technique. This can be done by the use of spanning tree architecture in order to eliminate the redundant comparisons, consequently, reducing the total number of comparison. On the other hand, it is not efficient when the number of requirements is large [10]. Numerical assignment (grouping) which provides a scale to all requirements based on separating them into different groups. Each requirement will then be assigned to a 5-point scale to assess its significance, however this technique provides low rate of reliability as well as fault tolerance [11-12]. Wieger Method determines the priority of the requirement by dividing the value of the requirement by the amount of costs and the technological risks associated with its implementation, and by assessing its customer significance, by applying 1-9 scale, as well as its implications, if this requirement were not enforced. It has drawback in which the stakeholders can easily influence it to achieve their objective goals [13]. MoSCoW technique is focused on cooperation between analysts and stakeholders to group the requirements into four categories. The efficiency here is good, but human attempts are required with disagreements between analysts and the views of stakeholders, so this approach would therefore be rated as low scalability and other hybrid techniques [14]. Most of the algorithms mentioned previously require quadratic complexity, so for a large number of requirements, the efficacy of the method becomes poor. Several papers were proposed in the literature in order to compare these method [15-17, 22, 23]. The purpose of this paper is not to compare the different approaches, but to suggest a new algorithm for the prioritization of requirements. However, the proposed approach for prioritizing requirements in this paper differs from the previous methods in achieving linear time complexity as well as the ability to reprioritize the requirement based on the resources availability.

The need for reprioritization has emerged from the fact that despite the effort expended in order to prioritize the requirements, this would be influenced by the constraints of precedent and resources constraints [18]. Therefore, this paper aims to introduce a hybrid approach of Traceability list and Directed Acyclic Graph to represent the requirements interdependencies for the prioritization process. As well as introducing a new algorithm for reprioritizing the requirements based on the queuing theory.

As discussed earlier, there are many types of requirement interdependencies mentioned in the literature. So it is worth to mention that this paper, proposed the prioritization and reprioritization requirement algorithms irrespective of the types of interdependencies and the methods used to identify them between the requirements, which are beyond the scope of this paper. Instead of focusing on the types of interdependencies and how they are described in any software project, this paper focuses on prioritizing the requirements based on the proposed algorithms using the proposed dependency parameters.

The paper is structured as follows: section 1 is the introduction and related work, section 2 gives a description of requirements interdependencies, section 3 presents the proposed approach in requirement prioritization, section 4 illustrates the approach as a case study and section 5 is the paper conclusion.

II. Requirements Interdependencies

The requirements influence each other and are linked to each other in a way that prevents handling them separately. This can be referred to as the dependencies between requirements. Basically, these requirements can also affect the decisions and activities during the development of the software. Requirements can, for example, affect each other through implementation constraints, the cost of implementing other requirements, or the customer satisfaction [19]. This means that in order to make accurate decisions during the development process, it is important to study the interdependencies. Simply stated, requirement interdependencies mean that a dependent relationship exists between the requirements. For instance, it is safer to start developing Ri before Rj if the Rj requirement requires Ri to work.

III. Methodology

A model for the prioritization and reprioritization of requirements based on a hybrid approach of representations of requirements is introduced. Fig. 1. demonstrates this model, which consists of two phases: the phase of prioritization and the phase of reprioritization. The first step can be achieved by prioritizing the requirements using dependency graph, while the next step is used to reprioritize the requirements using the queuing theory. The requirements are presented as a traceability list and then as a dependency graph, as shown in Fig. 1. The dependency graph is subsequently regarded as an input to the Priority Dependency Graph (PDG) algorithm in

order to create a priority list of requirements. This list is then processed on the basis of the resources constraints in the system by the Resources Constraints Reprioritization (RCR) algorithm.

A. Dependency Constraints

Any software can be defined as a set of R requirements where R={R1, R2, ... ,Rn}. When an incremental model is followed, then the first increment is analyzed, implemented and delivered as a set of these requirements. Other requirements are delivered as the next step and so on for the next level. In each increment the priority of requirements is playing an important role, but it is often precluded by requirements interdependencies. The approach to prioritizing the requirements in this paper is therefore focused on the representation of dependencies between requirements using the dependency graph. First of all, a simple description of the dependency factors used is discussed below:

1) Dependency scope: The Dependency Scope determines the scope of the requirements according to their dependencies, two types are available:

A Requirement R2 is an External Dependent on requirement R1 if and only if:

a) Execution of R1 precedes execution of R2.

b) Execution of R1 implies execution of R2 in the future increment.

A Requirement R2 is an Internal Dependent on requirement R1 if and only if:

a) Execution of R1 precedes execution of R2.

b) Execution of R1 implies execution of R2 in the same increment.

2) Dependency volume: The dependency volume determines the number of requirements that are internal dependent on the current requirement.

3) Dependency intensity: The dependency intensity determines the degree of dependency for each requirement, two types are available:

a) "Loose Dependencies is defined as: it would be ok to continue task without awareness of dependencies but would be better with awareness" [20].

b) "Tight Dependencies is defined as: the successor task has to wait until all its precursor tasks finish, the failure of the precursor will block the successor" [20].

The first parameter indicates the two main types of interdependencies in our method. External dependency determines whether the requirement in the future increment is dependent on another requirement, and internal dependency determines whether the requirement in the same increment is dependent on another requirement, as shown in Fig. 2.

The precedence constraint defined in the first parameter illustrates the relationship between the requirements in terms of precedence where in any software; the requirements must be implemented before other requirements. Therefore, for all iterations of increments there must be an order in which the requirements are executed. Loose dependencies can be used when there is no strict use of dependency between requirements, such as some requirements in mobile applications that doesn't require awareness of the context of the mobile user, but if provided the software will behave in an efficient manner. On the other hand, tight dependency is for those requirements that must be executed before other requirements as is the case in most software applications.

As depicted in Fig. 2. , the requirements are represented as a dependency graph where a directed acyclic graph (DAG) is used to define the R requirements as vertices V and the precedence constraints as edges E. In order to calculate the priority for each requirement in each increment based on the dependency types mentioned earlier, a topological sorting [21] with slight modifications is then performed.

An example of a dependency graph is represented in Fig. 3. R1 and R9 do not have dependencies in this DAG, while vertices R2 to R10 are dependent on other vertices; R4, for instance, depends on R1. Note that R4 has a volume of dependency greater than R3 that influences its prioritization process.

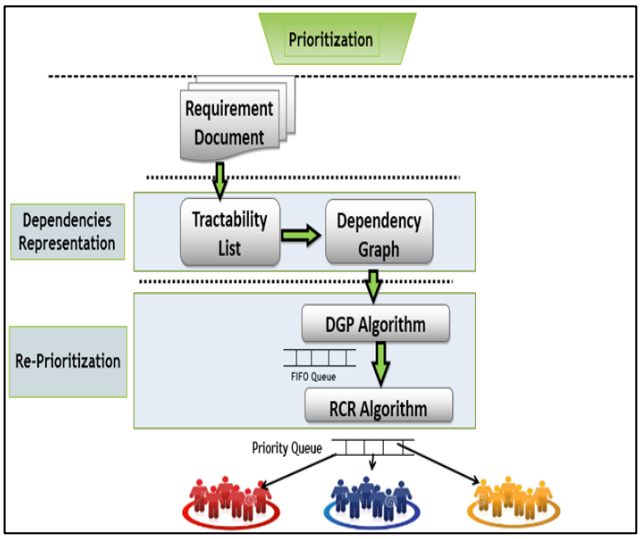


Fig. 1. The proposed Model.

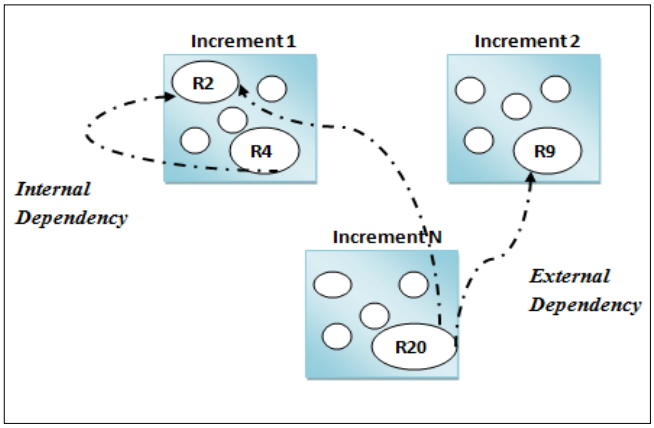


Fig. 2. External and Internal Dependencies.

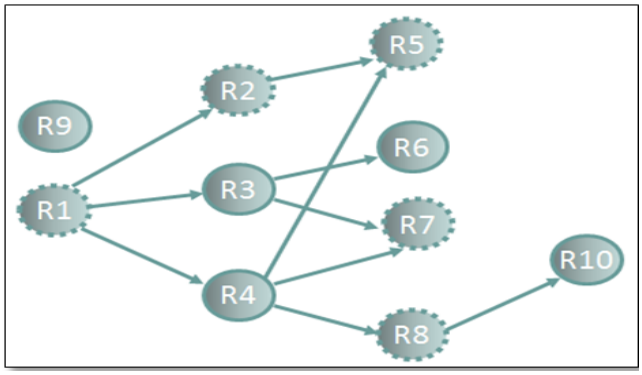


Fig. 3. Dependency Graph for Requirements.

Algorithm: Priority Dependency Graph (PDG) algorithm

Input: Digraph G = (V,E) ,Set of Requirements R as V, Set of Precedence Constraints as E, In_Degree array In_D, Out_Degree array Out_D, Dependency Scope array DS, Dependency Intensity array DI, Queue Q.

Output: list of Requirements R each associated with its priority

1: In_D ← { }
2: Store each vertex's InDegree in In_D array
3: Initialize Q with all in-degree zero vertices
4: While Q is not empty do
5 If DS for vertex v =1
6: Dequeue and output a vertex v
7: Set higher priority for v
8: Reduce In-Degree of all vertices adjacent to v by 1
9: Enqueue v which the In-Degree for it became zero
10: else
11: Dequeue and output a vertex v
12: Set lower priority for v
13: Reduce In-Degree of all vertices adjacent to v by 1
14: Enqueue v which the In-Degree for it became zero
15: repeat
16: end

Four arrays are used to calculate the priority in the proposed algorithm:

- InDegree array that contains the number of InDegree edges for each vertex.
- OutDegree array that contains the number of OutDegree edges for each vertex and represent the dependency volume.
- Dependency Scope array that determine the dependency scope for each vertex in the graph whether it is external or internal based on equation 1.
- Dependency Intensity that determine the intensity of the dependency for each vertex in the graph whether it is tight or loose based on equation 2.

The algorithm starts by initializing queue with all vertices that has zero InDegree, then while this queue has vertices in it.

$$\text{Dependencyscope} = \begin{cases} 1\ Internal\ Dependency \\ 0\ External\ Dependency \end{cases} \quad (1)$$

$$\text{DependencyIntensity} = \begin{cases} 1\ Tight\ Dependency \\ 0\ Loose\ Dependency \end{cases} \quad (2)$$

Higher priority will be granted to the vertices with internal dependency than those with external one; if two requirements are equal in the scope of dependency, then the decision can be made based on the volume of dependency. Such that, the requirements with tight dependency intensity will have higher priority for internal dependency than the requirements with loose dependency intensity; the algorithm will calculate the priority for those vertices until it is empty.

B. Complexity Analysis

Consider the complexity analysis of the proposed algorithm. Hence, a queue is used to store the vertices of zero InDegree. So, each time a node's InDegree is modified, we check if the value of it is 0 we add it to the queue, this will take O (|V|). Now to find a node of zero InDegree it takes O (1). The Dequeue operation will take O (|V|) while reducing the InDegree of all adjacent vertices to a vertex will take O (|E|). Therefore, the algorithm can be implemented to run in O (|V| + |E|) which is a linear running time. Note that most of the prioritization methods in the literatures have a quadratic complexity where our algorithm requires a linear one.

C. Resources Constraints Reprioritization Algorithm

As illustrated earlier, requirement prioritization is precluded by the available resources that are needed to develop the requirement or task, therefore the concept of the queuing theory is used to reprioritize and schedule the requirements to the available teams in the system. The method here is based on the outcome from the previous stage where the requirements are prioritized and added along with their priorities to a list or queue. The service facility may have of one or more teams. So, a requirement at the head of the queue can go to any team that is free. If there is more than one team, then a concurrent development will take place.

Each requirement in the queue must have the following characteristics:

1) Arrival Time λ to the queue for each requirement.
2) Waiting Time wt for each requirement, which indicates the waiting time for each requirement.
3) Status S: either FREEZE or INPROCESS. This parameter used to freeze the requirement and their dependent requirements.
4) Old Priority P_{old}, this parameter is used to indicate the old priority for the requirenment.
5) New Priority P_{new} , this parameter is used to indicate the new priority for the requirenment.
6) Available Resources Flag (ARF), this parameter will be used to indicate whether the resources are available to perform the requirement.

The RCR Algorithm involves the following steps:

1) Select the requirement with the minimum λ.
2) Check ARF whether it is set or not.
3) If ARF is equal to zero, then change the state S of this requirement to FREEZE and increment wt by 1.

4) If ARF is equal to 1 then change the state to IN_PROCESS and change Pold to Pnew and schedule it to the available team.

5) Each time check the ARF for the freeze requirement before reprioritize the new requirement since it has higher priority than it due to the dependency factor.

Fig. 4. demonstrates the prioritizer and scheduler that carries out the previous steps for reprioritizing the requirements based on the resources available and schedules them to the available team.

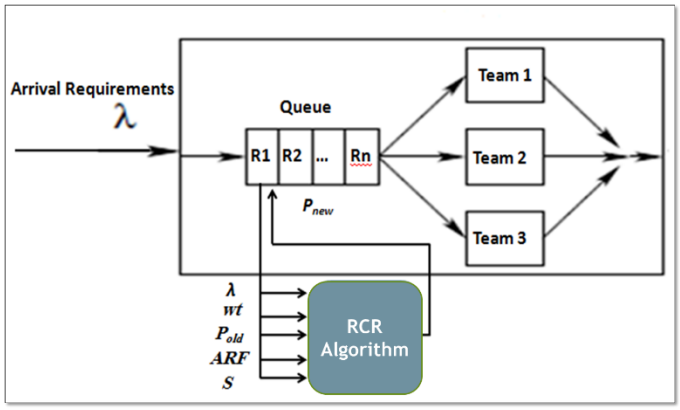


Fig. 4. RCP Algorithm.

IV. Case Study

To demonstrate our approach in a practical way, a sample software project for a library management system is considered, with three increments and twenty requirements, R = {R1,…, R20}. The first increment contains 9 requirements with the following dependencies:

InternalDependency = (R1), (R2, R1), (R3, R1), (R4, R1)

,(R5,R2,R3),(R6,R3),(R7,R3,R4),(R8,R5,R6),(R9)},the tuple (R3, R1) means that R1 depends on R2.

The second increment contains six requirements with the following dependencies:

InternalDependency={(R10),(R11,R10),(R12,R10),(R13,R11, R12),(R14,R11), (R15, R12)}.

The third increment contains five requirements with the following dependencies:

InternalDependency={(R16),(R17,R16),(R18),(R19,R17),(R2 0,R17)}.

Table I represents the traceability list for the twenty requirements. This list is converted into a directed acyclic graph which represents the requirements and interdependencies between them; Table II describes the number of InDegree and OutDegree edges for each vertex in the graph, Table III is the findings of the proposed algorithm which represents all the requirements and its associated priorities, notice that the requirement R1 has the highest priority with 2.0. While the requirement R19 has the lowest priority with 0.1, the scale is based on the total number of requirements in the software. Table IV showed the reprioritization process based on the resources constraints. Owing to the unavailability of the tools needed to develop these requirements, those requirements and their dependent requirements are frozen.

TABLE I. Traceability List

Increments	Requirement	Depends-On
Increment1 (9 Req)	R1	-
	R2	R1
	R3	R1
	R4	R1
	R5	R2,R3
	R6	R3
	R7	R3,R4
	R8	R5,R6
	R9	-
Increment2 (6 Req)	R10	-
	R11	R10
	R12	R10
	R13	R11,R12
	R14	R11
	R15	R12
Increment3 (5 Req)	R16	-
	R17	R16
	R18	-
	R19	R17
	R20	R17

TABLE II. InDegree and OutDegree for Requirements Independency Graph

Increment	Requirement	In Degree	Out Degree
Increment1 (9 Req)	R1	0	3
	R2	1	1
	R3	1	3
	R4	1	1
	R5	2	1
	R6	1	1
	R7	1	0
	R8	2	0
	R9	0	0
Increment2 (6 Req)	R10	0	2
	R11	1	2
	R12	1	2
	R13	2	0
	R14	1	0
	R15	1	0
Increment3 (5 Req)	R16	0	1
	R17	1	2
	R18	0	0
	R19	1	0
	R20	1	0

TABLE III. REQUIREMENT PRIORITIZATION

Requirement	Dependency Volume	Dependency Scope	Dependency Intensity	Priority
R1	3	1	1	2.0
R2	1	0	0	1.6
R3	3	1	1	1.8
R4	1	1	1	1.7
R5	1	1	1	1.5
R6	1	1	0	1.4
R7	0	0	1	1.3
R8	0	0	1	1.2
R9	0	0	1	1.9
R10	2	1	1	1.1
R11	2	1	1	1.0
R12	2	1	0	0.9
R13	0	0	1	0.7
R14	0	1	1	0.8
R15	0	0	1	0.6
R16	1	1	1	0.5
R17	2	1	1	0.3
R18	0	1	1	0.4
R19	0	1	0	0.1
R20	0	1	1	0.2

TABLE IV. REQUIREMENT REPRIORITIZATION

Requirement	Arrival Time λ	Waiting Time *wt*	*ARF*	P_{old}	P_{new}
R1	1	0	1	2.0	2.0
R9	2	0	1	1.6	1.9
R3	3	0	1	1.8	1.8
R4	4	3	0	1.7	1.4
R2	5	0	1	1.5	1.7
R5	6	0	1	1.4	1.6
R6	7	1	1	1.3	1.5
R7	8	1	0	1.2	1.2
R8	9	0	1	1.9	1.3
R10	10	0	1	1.1	1.1
R11	11	0	1	1.0	1.0
R12	12	1	0	0.9	0.8
R14	13	0	1	0.7	0.9
R13	14	0	0	0.8	0.7
R15	15	0	0	0.6	0.6
R16	16	0	1	0.5	0.5
R18	17	3	0	0.3	0.1
R17	18	0	1	0.4	0.4
R20	19	0	1	0.1	0.3
R19	20	0	1	0.2	0.2

V. CONCLUSION

In incremental software model small releases of the software are implemented in a sequence fashion instead of delivering the whole system after a long time of development. Therefore, this model can influence the prioritization of requirements in efficient manner so that the most important requirements can be implemented in the first releases of the system. A model were proposed to achieve requirement prioritization and reprioritization based on requirement interdependencies which represented as a hybrid approach of tractability list and directed acyclic graph, and on the resources constraints of the requirements. The proposed algorithms were introduced, analyzed and implemented using a case study. PDG and RCR algorithms require time complexity of O (|V| + |E|) which is a linear running time compared to the quadratic time complexity provided by the available algorithms that handle requirement prioritization. Also, the proposed approach has the ability to reprioritize the requirement based on the resources availability. Future work may add an improvement to the proposed algorithms or may combine them with other priority algorithms, in order to provide a hybrid solution that enhances the overall process.

REFERENCES

[1] K. Pohl, K.: Process-centered Requirements Engineering. Wiley, New York (1996).

[2] P. Carlshamre, K. Sandahl, M. Lindvall, B. Regnell and J.N och Dag, " An industrial survey of requirements interdependencies in software product release planning" . In Proceedings Fifth IEEE International Symposium on Requirements Engineering ,pp. 84-91, 2001.

[3] S. Soomro , A. Hafeez , A. Shaikh , SH. Musavi, "Ontology based requirement interdependency representation and visualization", InInternational Multi Topic Conference, pp. 259-270, Springer, 2014

[4] I. Sommerville, Software Engineering, Ninth edition, Pearson, 2011.

[5] vestola, " A comparison of nine basic techniques for requirements prioritization". Helsinki University of Technology, 2010.

[6] L. Karlsson, H. Martin, and R. Björn. "Evaluating the practical use of different measurement scales in requirements prioritisation." In Proceedings of the 2006 ACM/IEEE international symposium on Empirical software engineering, pp. 326-335, 2006.

[7] M. Pergher and B. Rossi. "Requirements prioritization in software engineering: a systematic mapping study". In Empirical Requirements Engineering (EmpiRE), 2013 IEEE Third International Workshop on , pp. 40-44, 2013.

[8] A. Asghar, A. Tabassum, Sh. Bhatti and A. Shah. "The impact of analytical assessment of requirements prioritization models: an empirical study." International Journal of Advanced Computer Science and Applications (IJACSA) , vol. 2, pp. 303-313, 2017.

[9] V. Ahl, V. "An experimental comparison of five prioritization methods - investigating ease of use, accuracy and scalability". Master's thesis, Blekinge Institute of Technology, Ronneby, Sweden, 2005.

[10] M. Yaseen, A. Mustapha, N. Ibrahim ,"Prioritization of Software Functional Requirements: Spanning Tree based Approach", International Journal of Advanced Computer Science and Applications (IJACSA) vol, 10, pp.489-497, 2019.

[11] JA. Khan, IU. Rehman , YH .Khan, IJ .Khan, S. Rashid, "Comparison of Requirement Prioritization Techniques to Find Best Prioritization Technique". International Journal of Modern Education & Computer Science. Vol. 11, pp. 53-59, 2015.

[12] C. Duan, P. Laurent, J. Cleland-Huang, and C. Kwiatkowski. "Towards automated requirements prioritization and triage" . Requirements Engineering, vol. 2, pp. 73-89, 2009.

[13] F. Moisiadis, "The fundamentals of prioritising requirements", InProceedings of the systems engineering, test and evaluation conference (SETE'2002) ,2002,

[14] S. Hatton, "Early prioritisation of goals. In Advances in conceptual modeling – Foundations and applications", ER 2007 Workshops CMLSA, FP-UML, ONISW, QoIS, RIGiM, SeCoGIS, Auckland, New Zealand, (pp. 235-244), 2007.

[15] M. Khari, N. Kumar, "Comparison of six prioritization techniques for software requirements", Journal of Global Research in Computer Science. vol. 4, pp. 38-43. 2013.

[16] M. Yousuf, MU. Bokhari, M. Zeyauddin," An analysis of software requirements prioritization techniques: A detailed survey", In2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 3966-3970, 2016.

[17] M. Yaseen, N. Ibrahim , A. Mustapha," Requirements Prioritization and using Iteration Model for Successful Implementation of Requirements", International Journal of Advanced Computer Science and Applications (IJACSA) vol, 10, pp.121-127, 2019.

[18] Z. Racheva, and M. Daneva, "Reprioritizing the Requirements in Agile Software Development: Towards a Conceptual Model from Clients' Perspective". In SEKE ,pp. 73-80, 2009.

[19] I. Bassey,"Towards Release Planning Generic Model: Market-driven software development perspective, IJERT, vol. 2 , 2013.

[20] L. Qi, "value based dependency aware inspection and test prioritization", PHD Dissertation , University of Southern California ,2012.

[21] TH. Cormen, C. Leiserson, E. Rivest, R. L., and C. Stein," Introduction to algorithms". MIT press, 2009.

[22] N. Saher, F. Baharom, and R. Romli, "Guideline for the Selection of Requirement Prioritization Techniques in Agile Software Development: An Empirical Research", International Journal of Recent Technology and Engineering (IJRTE), vol, 8, pp.3381-3388, 2020.

[23] N. Borhan, H. Zulzalil, S. Hassan, N. Mohd Ali, "Requirements Prioritization Techniques Focusing on Agile Software Development: A Systematic Literature Review", INTERNATIONAL Journal of Scientific and Technology Research, vol. 8, pp. 2118- 2125, 2019.

A Novel Geometrical Scale and Rotation Independent Feature Extraction Technique for Multi-lingual Character Recognition

Narasimha Reddy Soora[1]
Associate Professor, Computer Science and Engineering, Kakatiya Institute of Technology and Science Warangal, Telangana, India

Ehsan Ur Rahman Mohammed[2]
Student, Computer Science and Engineering, Kakatiya Institute of Technology and Science Warangal, Telangana, India

Sharfuddin Waseem Mohammed[3]
Assistant Professor, Computer Science and Engineering, Kakatiya Institute of Technology and Science Warangal, Telangana, India

***Abstract*—This paper presents a novel geometrical scale and rotation independent feature extraction (FE) technique for multi-lingual character recognition (CR). The performance of any CR techniques mainly depends on the robustness of the proposed FE methods. Currently, there are very few scale and rotation independent FE techniques present in the literature which successfully extract the robust features from characters with noise such as distortion and breaks in the characters. Many FE methods from the literature failed to distinguish the characters which look similar in their appearance. So, in this paper, we have proposed a novel scale and rotation independent geometrical shape FE technique which successfully recognized distorted, broken, and similarly looking characters. Aside from the proposed FE technique, we've used crossing count (CC) features. Finally, we have combined the proposed features with CC features to make as Feature Vector (FV) of the character to be recognized. The proposed CR technique is evaluated using publicly available media-lab license plate (LP), ISI_Bengali, and Chars74K benchmark data sets and achieved encouraging results. To further assess the performance of the proposed FE method, we've used a proprietary data set containing nearly 168000 multi-lingual characters from English, Devanagari, and Marathi scripts and achieved encouraging results. We have observed better classification rates for the proposed FE method using publicly available benchmark data sets as compared to few of the CR FE methods from the literature.**

Keywords*—Feature extraction; character recognition; crossing count features; edit distance; scale and rotation independent feature extraction

I. Introduction

We the human beings have the beautiful ability to recognize the text present in all sorts of forms such as those printed in different font styles, handwritten, sloppy, and inclined, which are camouflaged with the background, possessing variations in illumination and brightness, of varying sizes, occluded ones, from various viewpoints, written upside down, having characters with missing parts, unwary decorations and marks, broken or even misspelled, having artistic and figurative designs, and many more. It comes as no surprise that the creative computer vision (CV) community despite six and more decades of intensive research could not achieve much in making computers represent images and perform well-defined white-box generic and low-level processes (ANNs are algorithms belonging to the category of the black box) on images thereby making computers capable of detecting and recognizing texts robustly in a way much similar to that of the humans along with many other activities performed by human visual cortex such as classification, and alike.

The class of computer algorithms performing text recognition is known as Optical Character Recognition (OCR). The path tread for image formation and analysis started with mechanical means, followed by optical means, and now uses the concepts of digital representation and processing, has led many works to establish their base on those physical phenomena such as inertia and concepts such as moments which is descriptive of the physical body, etc. Working on the similar lines there is literature which takes heed from physical and biological concepts in order to describe the image using its features, different features being responsible to obtain robustness from different kinds of variations and thereby incorporating as many features as possible gives less space to the algorithms to make mistakes and makes the FV generic. Less computation cost and fastness are also key concerns in image processing (IP) as images are an inherently large collection of data, so care must be taken in making FV's less redundant and compact.

OCR is the most important to the real-world document analysis and storage tasks [10], and it has been used even in technologies that are dealing with automation as in the tasks of natural language processing (NLP) such as machine translation, data mining, along with tasks such as indexing, word spotting, LP recognition, signature verification in banks, sorting postcards based on address, reading aid for blind, etc. FE techniques reduce the dimensionality of the image to be recognized and thereby making the recognition process computationally efficient and mathematically feasible. These features are then checked for similarities with an abstract vector representing a character. General FE techniques from CV are all applicable to an OCR and such OCRs are commonly seen in "intelligent" handwriting recognition and indeed most modern OCR software utilizes such FE techniques which are more based on learning rather than hard-coded mathematics. In the deep learning era, were the internal formulation of ANNs has worked out ways to extract and

learn from features much in a biological way. But robustness could be enhanced to ultra-fine intricacies of data for proper categorization and decision making can be learned if traditional CV and deep learning could go hand-in-hand.

As per the paper [16], there are very few robust scale and rotation independent FE techniques present in the literature which can recognize normal and similarly looking characters from various multi-lingual languages and this inspired us in proposing a novel scale and rotation independent geometrical FE technique for multi-lingual CR and it takes inspiration from some more related, empirically well-established scale and rotation invariant FE techniques, which are briefly summarized in the next section. Most of the existing CR algorithms [16] used a combination of FE techniques from the literature which are elaborated in the paper [16].

The remaining sections of this paper are organized as follows. Section II describes related work from the literature; Section III describes the detailed proposed CR technique; Section IV describes experimental results and discussion and Section V is about conclusion.

II. Related Work

Despite many years of advancement in CV and machine learning, CR is challenging till today [13] because of varying complexity of characters in the form of character graphemes of various languages worldwide, presence of broken, distorted, rotated, varying sized and similarly looking characters. We have observed there are very few scale and rotation independent FE techniques for multi-lingual CR. In this section we will discuss a few of the existing FE techniques (mostly scale and rotation independent) for CR from the literature.

K. Sampath et al. in the paper [1] proposed a feature extraction technique for character recognition using combination of existing features such as histogram oriented Gabor features, grid level features (local gradient), and gray level co-occurrence matrix and reported a success rate of 96% using Chars74K data set. The concept of calculating moments has been central to some IP tasks [2] and applications like pathological brain detection problems, etc. Many moments like Hue, Zernike, and pseudo-Zernike moments (orthogonal radial moments) are invariant to rotation and scale with the help of few geometric transforms. Zernike Moments have the least redundancy of information and hence are less susceptible to noise in the image and it also has better numerical stability. Paper [2] used the magnitude of Zernike moments as rotation invariant features for the classification of grayscale face images and binary character images and reported an accuracy of 99.7% using the Roman proprietary data set of 1560 characters.

Authors in the paper [3] uses a traditional approach wherein a covariance matrix is constructed from a set of rotated versions of each character and an Eigenspace is derived from the matrices obtained, a locus is constructed by projecting the respective rotated characters onto their Eigen sub-space a part of the actual Eigenspace obtained from all the categories. Recognizing a character is done by simply projecting it onto the Eigen sub-spaces and measuring the distance between the projected points and the locus present in each Eigen sub-space. The problem which needs to be acknowledged is that some characters form similar types of locus and hence more rigorous testing needs to be done while recognizing such characters, sometimes this can make the computation cost high due to suggestive interpolation that needs to be applied during the formation of the loci. As this method is dependent on training samples, the samples need to be selected carefully and the number of training samples must also be high for more accurate and precise locus. A strange observation was that few symmetric characters were recognized correctly despite the angle of rotation being wrongly interpreted. The authors have tested the proposed method using a proprietary data set containing 2808 characters with different orientations and reported an accuracy of 99.89%.

T. Hayashi et al., in the paper [4] divided the input character into sub-level patterns which can be mapped to elementary components. The set of sub-patterns derived are used for the task of recognition and this division is based on cross points, angle point's and the system is free from scale and rotation variations as the classification is based on the combination of elementary components which these characters are composed of. The authors in the paper [4] reported 100% accuracy on the proprietary Arabic numerals data set of size 150. For texture images, rotation invariant representation is possible by using dominant orientation which is the orientation with highest total energy across different scale considered during image decompositions as in [6] and finally, rotation-invariance is obtained by circularly shifting the elements of FV within the same given set of scales so that the first elements found at each scale shall have maximum correspondence with the dominant orientations. Representation with the highest total energy across the different orientations (dominant scale) results in the scale-invariant one. The feature alignment process to classify texture is based on the assumption that the images should be rotated so that their dominant orientations/scales are the same. It has been proved that the image rotation in the spatial domain is equivalent to a circular shift of feature vector elements. This paper reported an average recognition accuracy of 98.89% using four image data sets from the Brodatz database.

K. U. Rehman et al. in the paper [7] proposed a feature extraction technique for character recognition using existing moment based features such as raw moments, central moments, hu moments, and Zernike moments and reported a success rate of 96.922% using Urdu proprietary data set. L. A. Torres-Méndez et al., in the paper [8] presented a translation, rotation, and scale invariant method for object recognition by extracting topological object characteristics with the help of novel coding of the normalized moment of inertia. They have tested the proposed method with 238 proprietary images and reported 98% accuracy. Paper [9] presents a rotation and scale invariant multi-oriented CR technique wherein a given character has divided multiple circular zones and each zone is divided into three centroids through the combination of the segments obtained from its constituent character into two clusters and one being the global centroid of the segmented

character. The ordering of centroids as per their distances from the global centroid makes the farther one to be included in one set and another one to be included in the next set and this is crucial to the construction of rotation invariant FV. Observations are that the relative positions and structure is unaltered after clustering is performed and reported the highest accuracy of 99.01% and 99.25% using Bangla and Devanagari data sets of 7874 and 7515 characters respectively.

Parul Sahare et al., in the paper [10] proposed a set of FE techniques for CR based on geometrical properties of characters. The first set of proposed features is based on adaptive center distance based on Euclidean distance from the centroid of each non-overlapping block to each foreground pixels. The second set of features is based on fixed center triangular cut based features for each non-overlapping block is computed. The last set of features is based on neighborhood count based features with the help of a window of size 3x3. The generated features are combined to form a feature vector of a character. The authors in the paper [10] reported an average recognition accuracy of 98.56% using the Chars74K data set containing alphabets and numerals. Rina D. Zarro et al., in the paper [11] proposed a hybrid algorithm using Hidden Markov Model and harmony search algorithm for online Kurdish CR. Authors have classified a group of characters into smaller subgroups based on directional features with the help of Markov model. The small number of group of characters is fed to a harmony search algorithm that uses a common movement pattern for recognition. The proposed system was tested using a proprietary data set having 4500 words structured with 21234 Kurdish characters, and reported an accuracy of 93.52%.

R. P. Kaur et al. in the paper [12] proposed a feature extraction technique for Gurumukhi characters recognition using existing techniques such as zoning features, diagonal features, and parabola curve fitting based features and reported an accuracy of 96.19% using 1605 Gurumukhi characters. J. Chaki et al. in the paper [19] proposed a framework to classify fragmented handwritten digits into three classes based on geometrical functions, grading scheme and fuzzy rules. Authors in the paper [13] have proposed robust geometrical FE techniques for the license plate (LP) CR. They have proposed geometrical shape FV generation using horizontal and vertical scan lines and angular width FV generation using horizontal, vertical, right diagonal, and left diagonal scan lines. They combined these two FVs along with the crossing count FV to form the FV of a character. Authors have tested the proposed FE method with the help of publicly available MediaLab benchmark LP database [5] containing 741 images with 6584 characters (English alphabets and numerals) extracted from these 741 images using LP detection method proposed in the paper [14] and reported an accuracy of 98.8% at the character level.

N. R. Soora et al., in the paper [15] proposed two novel FE techniques namely shape geometry encoding of components of characters with the help of perpendicular distances and encoding of triangular areas computed using four scan lines namely horizontal, vertical, right diagonal and left diagonal scan lines. Authors have tested the proposed method with the help of the MediaLab LP benchmark database [5] containing 6584 characters (English alphabets and numerals) and reported an accuracy of 99.03%. The authors have tested the proposed method using proprietary data sets containing nearly 30000 characters (Devanagari, Marathi, and English alphabets) and reported an accuracy of 98.5%. A good list of FE techniques could be found at [16] and some key points from the authors are provided here. The features which do not have discriminating capabilities to classify an input character when considered alone are called non-shape-based features and are used to eliminate false hits or pooled with other features to recognize characters that look similar in shape. Statistical methods are not generic because they generally involve threshold values that have to be set by the programmers and are not learned or adjusted according to variations in data and the tasks which are to be performed on the given image (data). So, generally statistical methods are not used due to their inflexibility and a great deal of trial and error involved to reach sufficient or good performance.

Tian et al., in the paper [17] proposed two FE techniques for multi-lingual scene CR using co-occurrence of HOG. The first FE technique is based on co-occurrence HOG (Co-HOG) in which authors have encoded co-occurrence of oriented pairs of neighboring pixels. The next FE technique is based on the convolution of Co-HOG (ConvCo-HOG) in which authors have extracted Co-HOG features from all possible images. The performance of the method was tested using five charter data sets in which the ISI_Bengali data set is publicly available and they have reported an accuracy of 92.2% on the ISI_Bengali data set. U Pal et al., in the paper [18] proposed a FE technique for multi-oriented and multi-sized CR using a contour distance based approach. In this paper, the authors have extracted the features by finding the distances from the centroid of the character to the contour points of the characters. They rearranged the extracted features in such a way that the FE is size and rotation invariant. Authors have reported an accuracy of 97.8% and 98.1% using proprietary 2900 Bangla characters and proprietary 3100 Devanagari characters respectively. Authors of the paper [19] tested the proposed method using MNIST, Numta, and Devanagari numerical data bases and reported good recognition accuracies.

We have considered 30000 characters extracted from 280 aged multi-lingual Indian documents having English, Marathi, and Devanagari scripts to assess the performance of the proposed algorithm and achieved a success rate of 98.8% which is almost equivalent the success rate of 98.5% reported in [15]. The advantage of the proposed system as compared to [15] is that it is scale and rotation independent. To test the scale and rotation independent factor of the proposed method, we have generated 168000 characters from proprietary test set of 12000 characters that are different in orientation and sizes and achieved nearly 98% accuracy using the manually generated proprietary data set. As it is not justifiable to compare the performances using proprietary databases, we have considered publicly available MediaLab LP, Chars74K, and ISI_Bengali benchmark data sets to compare the performance of the proposed method with methods from literature and explained the same in detail in the experimental results section.

III. Proposed Work

In this section we have explained in detail about the proposed novel geometrical scale and rotation independent (SRI) FE technique for multi-lingual CR. We have extracted crossing count (CC) features along with the proposed SRI FE techniques and we have combined both SRI and CC features to form the FV of the character.

A. Scale and Rotation Independent Feature Extraction Generation

SRI features are generated with the help of sweep lines which will pass through the centroid of the input character. At first, remove all the background pixels in all directions of the input character so that the input character will properly fit into a rectangular box as shown in Fig. 1. Find the centroid of the character and let it be C(Xc, Yc). Next, find the boundary points of the input character by traversing in 8-directions from the centroid C. Let N (≤ 8) be the number of boundary points that we encounter while traversing from centroid C. Find the distance Di (for i = 1 to N) from centroid C to the N boundary points using Equation (1) where Pi(Xi, Yi) be any general point on the boundary of the input character. Find a point Pi(Xi, Yi) which is closest to centroid C as shown in Fig. 2. Find the slope m of the line joining the points C and Pi using Equation (2). Find the angle θ between the line joining CPi and x-axis using the Equation (3).

$$D_i = (X_c - X_i)^2 + (Y_c - Y_i)^2 \quad (1)$$

$$m = \left((Y_c - Y_i) \;/\; (X_c - X_i)\right) \quad (2)$$

$$\theta = (180/\pi) * \tan^{-1}(m) \quad (3)$$

Extend the line CPi from both sides of the centroid in the direction of slope m up to the boundary of input character using the formula's (4) and (5) to find all points (Xnew, Ynew) on the line by varying the distance d value from centroid C (Xc, Yc) as shown in the Fig. 2. (Xnew, Ynew) are points which are the combination of (Xnew1, Ynew1) and (Xnew2, Ynew2). (Xnew1, Ynew1) are the set of points of the line CPi which reside on one side of the centroid C and (Xnew2, Ynew2) are the set of points of the line CPi which reside on the second side of the centroid C.

$$X_{new} = X_c \pm \left(d \;/\; \sqrt{(1 + m^2)}\right) \quad (4)$$

$$Y_{new} = Y_c \pm \left((m * d) \;/\; \sqrt{(1 + m^2)}\right) \quad (5)$$

Collect all distinct points $P_{new}(X_{new}, Y_{new})$ generated using the Equations (4) and (5) which lie on the line CP_i with slope m as shown in the Fig. 3. At this stage, we have all the points lying on the line CP_i with slope m. Now, find all the starting and ending cut points (from centroid C) from the list of the points computed, which are foreground of the input character that CP_i intersects with each connected component. Let the points of intersection or the cut points of the line CP_i with foreground boundary points of the input character be $P_k(X_k, Y_k)$ as shown in Fig. 4. Find the distances from centroid C to all P_k's using the Equation (1). Preserve all the computed distances in a separate list called DISTANCES. Now change the angle θ that sweep line intersects with x-axis by 2 degrees and find the slope of the new sweep line using the Equation (3) with new θ computed. Generate all the points of the new sweep line passing through centroid C with slope m using Equations (4) and (5) as explained previously and find all the cut points of that the new sweep line that intersects with the foreground boundary points of the input character. Compute the distance from centroid C to the new cut-points of the sweep line with each of the connected components at the boundary points of the input character. Again, preserve the computed distances in the DISTANCES list. Repeat the steps of finding the distances from the centroid to the cut points for each new sweep line formed by varying θ as shown in Fig. 4.

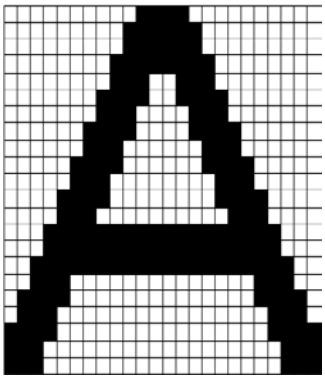


Fig. 1. Original Image Fitted in a Rectangular Box by Removing Background Pixels from All Directions.

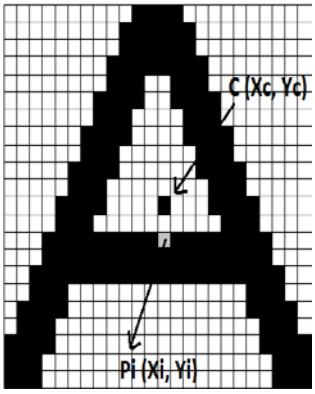


Fig. 2. Example Figure Showing Centroid C (X_c, Y_c) and Nearest Point P_i (X_i, Y_i) on the Boundary.

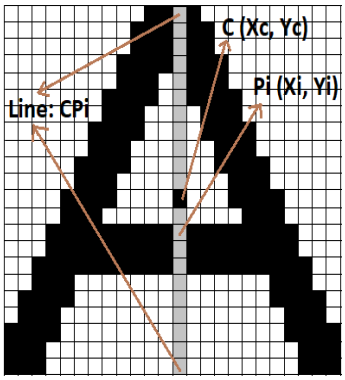


Fig. 3. Example Figure Showing Generation of the Sweep Line CPi.

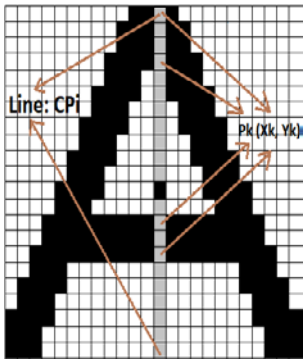


Fig. 4. Example Figure Showing Generation of Foreground Boundary Points P_k (X_k, Y_k) of Input Character Intersection with Sweep Line CP_i.

This process has to be repeated for 90 times (as we are increasing the θ value each time by 2 degrees) so that one end of the initial sweep lines reaches to its opposite position. At this stage, we have all the distances (preserved in Distance list) from centroid C to the cut points of all sweep lines with all the probable connected components of the input character at the boundary points. Normalize the distances present in the DISTANCE list by finding the maximum distance of all the distances in the DISTANCE list and divide all the distances with maximum distance using the Equation (6). The resultant DISTANCE list will contain the values in the range from 0 to 1. Encode DISTANCE list values ranging from 0 to 1 using the Equation (7) which results in SRI features of the input character.

The described FE method extracts the shape of any input character in the form of shape symbols built using Equation (7) with the help of normalized distances from centroid to the boundary points of the input character where each sweep line intersected. The proposed FE method does not use any existing boundary extraction algorithms. To find the next boundary point, we have not traversed through the next available boundary point. The proposed SRI FE method extracts the complete set or subset of the boundary points with the help of sweep lines which retains the shape of the input character. This capability of the proposed FE technique extracts the robust features from the input character even in the presence of the noise and is independent of size, rotation, distortion, and breaks of the input character. It has the ability to distinguish similarly looking characters as mentioned in [16]. The detailed description of the algorithm is shown in Table I.

$$DISTANCE = DISTANCE \;/\; \max(DISTANCE) \tag{6}$$

$$SRIF = \begin{cases} A, & Di \geq 0.0 \;\; and \;\; Di < 0.1 \\ B, & Di \geq 0.1 \;\; and \;\; Di < 0.2 \\ C, & Di \geq 0.2 \;\; and \;\; Di < 0.3 \\ D, & Di \geq 0.3 \;\; and \;\; Di < 0.4 \\ E, & Di \geq 0.4 \;\; and \;\; Di < 0.5 \\ F, & Di \geq 0.5 \;\; and \;\; Di < 0.6 \\ G, & Di \geq 0.6 \;\; and \;\; Di < 0.7 \\ H, & Di \geq 0.7 \;\; and \;\; Di < 0.8 \\ I, & Di \geq 0.8 \;\; and \;\; Di < 0.9 \\ J & Di \geq 0.9 \;\; and \;\; Di \leq 1.0 \end{cases} \tag{7}$$

TABLE I. ALGORITHM TO GENERATE SRI FEATURES FROM AN INPUT CHARACTER

1. Remove background pixels from all directions of the input character and the resultant input character looks as shown in the Fig. 1.
2. Find the centroid C (X_c, Y_c) of the input character.
3. Move from centroid C in 8-directions to find boundary pixels of the input character and let N (≤ 8) be the number of boundary pixels.
4. Compute the distance from C to the N boundary points of the input character using Equation (1).
5. Find a boundary point Pi (Xi, Yi) which is closest to the centroid C and let it be P_i (X_i, Y_i) as shown in Fig. 2.
6. Find the slope m of the sweep line joining C and P_i using Equation (2).
7. Find the angle θ between the sweep line CP_i and x-axis using Equation (3).
8. Find all the distinct points $P_{new1}(X_{new1}, Y_{new1})$, $P_{new2}(X_{new2}, Y_{new2})$ lying on both side of the centroid C of the sweep line CPi as shown in Fig. 3 till the borders of the input image using Equations (4) and (5) by changing the d value. The value of d indicating how far we are moving from centroid C of the input character. P_{new1} are the set of points lying on one side of the centroid C of the sweep line CP_i and P_{new2} are the set of points lying on second side of the centroid of the sweep line CP_i. The set of points P_{new1} and P_{new2} are useful in generating crossing count features.
9. Find the cut points or intersection points that the P_{new1} and P_{new2} making with each connected components of the input character at the boundary points as shown in Fig. 4. Let the intersection or cut points be $P_k(X_k, Y_k)$.
10. Find the distances from centroid C to all P_k's and store them in a separate list called DISTANCES.
11. At this stage we have to find new sweep line. To find new sweep line, increase the value of θ by 2 degrees. Find the new slope m of new sweep line using updated θ with the help of the Equation (3).
12. Repeat steps from 8 to 12 for 90 iterations (which results into moving one side of a sweep line to its opposite side) to compute the distances from centroid C of the input character to all the cut points generated by all the sweep lines.
13. At this stage, we have the shape of the input character in the form of distances from centroid C to the boundary pixels of the input character.
14. Find maximum distance from the set of distances present in DISTANCE list.
15. Divide all the distances present in DISTANCE list with maximum distance. After this step, DISTANCE list contains the values in the range from 0 to 1.
16. Encode the normalized distances present in DISTANCE list using Equation (7).
17. Now, the DISTANCE list contains shape of the input character in the form of shape symbols present in the Equation (7). This DISTANCE list is SRI features.

Let Nnew1i, Nnew2i be the number of connected components for each of the set of points Pnew1i, Pnew2i respectively for sweep line i. The total number of SRI features generated is given by the Equation (8).

$$No.of\ SRI\ features = \sum_{i=1}^{90} \left(2*(N_{new1i} + N_{new2i})\right) \tag{8}$$

B. Crossing Count Features Generation

CC features are generated during the SRI features generation. In Table I step 8, we have set of pixels Pnew1 and Pnew2 on either side of the centroid C of any sweep line CPi.

Using Pnew1 and Pnew2 information, find the count of continuous foreground pixels which indicates the number of connected components on both sides of the centroid C. If there are no foreground pixels, take count as '0'. Store the connected components computed from Pnew1 and Pnew2 into a new list called CCFE. Repeat the process of finding the connected components for each sweep line. Let CCnew1i, CCnew2i be the CC features generated using Pnew1, Pnew2 set of pixels respectively which lie on either side of the centroid C of the sweep line CPi. CC features are generated using the Equation (9) shown below. The steps described above generate CCF of size 180.

$$CCF = \{CC_{new1i}, CC_{new2i}\}_i = 1 \text{ to } 90 \quad (9)$$

Combine the proposed SRI features and CC features to form SRIF vectors. SRIF = {SRI, CCF}. The total size of SRIF can be found by summing up the values from Equations (8) and (9). The size of SRIF differs from one character to other character depending on the complexity of the input character.

IV. Experimental Results and Discussion

We have implemented the SRI FE technique discussed in the previous section using MATLAB 18a on the Intel Core i5 processor machine with 8 GB RAM. We have used the Edit Distance metric for classification because the FV we have generated is the collection of shape symbols from Geometrical SRI features and Statistical CC features which are generated using sweep lines. At first, we have a proprietary data set containing 30000 characters extracted from 280 aged printed multi-lingual Indian documents having Marathi, Devanagari, and English scripts [15]. Out of 30000 characters we have selected 18000 characters for training and 12000 characters for testing. We have stored all 18000 FV's generated using the proposed FE method in a flat file. For each test character, we have generated the proposed geometrical SRI FV and compared the test FV with all the FVs stored in the file. The character which generated minimum edit distance is considered as the classified input character.

With 12000 proprietary data set test characters, we have achieved 98.8% success rate which is almost equivalent to the performance of the method from [15]. To further assess the performance of the proposed system, we have used MediaLab LP benchmark data set having 6584 characters containing English alphabets and numerals and achieved 99.2% success rate which outperforms the performance of the methods from [13], [15], [18] and using ISI_Bengali character data set containing 19530 Bengali characters, achieved 97.72% success rate which outperformed the method proposed in [17].

Table II shows the performance comparison of the proposed method with a few of the methods from the literature that used similar publicly available benchmark data sets. As per the MediaLab LP benchmark data set, we have achieved 99.20% accuracy which outperformance the methods [13] (accuracy: 98.8%), [15] (accuracy: 99.03%), and [18] (accuracy: 95.4%). The disadvantage of the methods from [13], [15] is that they are not scale and rotation independent. Apart from the MediaLab LP benchmark data set and ISI_Bengali data set, we have assessed the performance of the proposed method using the Chars74K benchmark data set containing Kannada and English alphabets and achieved 98.64% which is almost equivalent to 98.56% reported by the method from [10] and better than 96% reported by [1]. The disadvantage of the methods [10] and [1] is that they are not scale and rotation independent.

TABLE II. Performance Comparison of the Proposed Method with Few of the Methods from the Literature that used Publicly Available Benchmark Data Sets

Reference	Name of the data set	Number of characters of data set	Recognition accuracy (%)
[13]	MediaLab LP benchmark data set	6584 characters	98.80
[15]	MediaLab LP benchmark data set	6584 characters	99.03
[18]	MediaLab LP benchmark data set	6584 characters	95.40
Proposed method	MediaLab LP benchmark data set	6584 characters	99.20
[13]	Chars74K benchmark data set	74107 characters	97.24
[15]	Chars74K benchmark data set	74107 characters	97.89
[10]	Chars74K benchmark data set	74107 characters	98.56
[18]	Chars74K benchmark data set	74107 characters	94.20
[1]	Chars74K benchmark data set	74107 characters	96.00
Proposed method	Chars74K benchmark data set	74107 characters	98.64
[13]	ISI_Bengali benchmark data set	19530 characters	94.45
[15]	ISI_Bengali benchmark data set	19530 characters	95.20
[17]	ISI_Bengali benchmark data set	19530 characters	92.20
[18]	ISI_Bengali benchmark data set	19530 characters	91.14
Proposed method	ISI_Bengali benchmark data set	19530 characters	97.72

Table III shows the performance comparison of the method [18] from the literature with the proposed method using publicly available benchmark data sets. We have implemented the method proposed in [18] to test its performance with a few of the benchmark data sets [7]. Even though we have implemented the method proposed in [18], the implementation may not meet the optimizations as per the expectations of the original author [16]. From Tables II and III, it is very clear that the proposed method outperformed many methods from the literature which used publicly available benchmark data sets and proprietary data sets.

Fig. 5. Recognition Results of Few of the Characters from Various Languages by the Proposed Method.

Fig. 5 shows a few of the example CR results by the proposed method. To test the scale and rotation invariance of the proposed method, we have generated manually the characters from MediaLab LP benchmark data set and proprietary data sets using software to rotate the characters with 45°, 90°, 135°, and 180° and resized the set generated with two different sizes. As the FE method in [18] is invariant to scale and rotation, we have compared the performance of the method from [18] with the proposed method using the manually generated characters. We have generated 168000 characters from 12000 proprietary characters containing Marathi, Devanagari, and English scripts. Total we have 180000 characters as part of the proprietary data set. Similarly, we have generated a total of 92176 characters from 6584 characters of MediaLab benchmark LP data set. So, we have a total of 98760 characters as part of the MediaLab benchmark data set.

Table III shows the performance comparison of the proposed method with the method from [18] using newly generated data sets. It is very clear from Table III, that the scale and rotation invariance of the proposed method outperformed the method from [18]. The disadvantage of the method from [18] is that, it is sensitive to the breaks present in the characters. The disadvantage of both the methods (proposed and [18]) is that, if the input character distorted in shape completely, both methods fail to classify the input character. Examples of such characters '5' and 'N' are shown in Fig. 5. Such distorted characters are recognized properly by the methods proposed in [13] and [15]. The advantage of the proposed method is that, it is invariant to scale and rotation. Another advantage of the proposed method is that it was able to classify the similarly looking characters such as {"0", "o", "D", "O"}, {"Z", "z", "2"} and {"8", "B"} from English alphabets and numerals and other scripts as well. Table IV shows the performance comparison of the proposed method with few of the methods from the literature. Table IV gives complete details of the feature extraction methods from the literature such as the number of characters used in the data sets, the language of the data sets, proposed features, whether the data set is benchmark data set or proprietary data set, whether the feature extraction is scale and rotation invariant or not, and recognition rate. Authors of the papers [4], [3] reported 100%, 99.89% accuracies respectively but the authors have not used publicly available benchmark data sets for evaluation purpose. Table IV clearly shows that the proposed method outperformed the methods from literature which used publicly available benchmark data sets for performance evaluation.

TABLE III. PERFORMANCE COMPARISON OF THE PROPOSED METHOD WITH THE METHOD FROM [18] USING MANUALLY GENERATED CHARACTERS FROM VARIOUS DATA SETS

Reference	Name of the data set	Number of characters generated manually from actual data set	Recognition accuracy (%)
[18]	proprietary data set	180000 characters	93.20
Proposed method	proprietary data set	180000 characters	98.00
[18]	MediaLab LP benchmark data set	98760 characters	92.55
Proposed method	MediaLab LP benchmark data set	98760 characters	98.25

TABLE IV. PERFORMANCE COMPARISON OF THE PROPOSED METHOD WITH FEW OF THE METHODS FROM LITERATURE

Reference	Number of characters/images/videos	Language & Hand written / Printed characters	Proposed features	Benchmark data set / proprietary	Independency of FE (Scale, Rotation)	Recognition Rate (%)
[4]	150 images	Arabic Numeral & Printed characters	Shaped based local features (Elementary Components of the input character)	Proprietary	Scale and Rotation Invariant	100
[9]	7874 characters	Bangla & Printed characters	Shape based local features (circular zones)	Proprietary	Scale and Rotation Invariant	99.01
[9]	7515 characters	Devanagari & Printed characters	Shape based local features (circular zones)	Proprietary	Scale and Rotation Invariant	99.25
[6]	Non-character (Texture) Data base	NA	-	Benchmark data set (Non characters)	Scale and Rotation Invariant	98.89
[2]	1560 characters	Roman & Printed characters	Shape based (Moment Based Features)	Proprietary data set	Rotation Invariant	99.70
[8]	238 images	English Alphabets and Numerals & Printed characters	Shape based (Moment Based Features)	Proprietary data set	Rotation and Scale Invariant	98.00
[13]	6584 characters	English Alphabets and Numerals & LP characters (Printed)	Shape based (Geometrical shape encoding and Angular Width features)	MediaLab LP benchmark data set	None	98.80
[15]	6584 characters	English Alphabets and Numerals & LP characters (Printed)	Shape based (Geometrical shape encoding of perpendicular distances and triangular area calculations)	MediaLab LP benchmark data set	None	99.03
[15]	30000 characters	Devanagari, Marathi and English Alphabets & Printed characters	Shape based (Geometrical shape encoding of perpendicular distances and triangular area calculations)	Proprietary data set	None	98.50
[3]	2808 characters	English Alphabets & Printed characters	Shape based (Eigen space)	Proprietary data set	Rotation Invariant	99.89
[11]	21234 characters	Kurdish Alphabets & Hand written	Shape based (Movement patterns of characters)	Proprietary data set	None	93.52
[10]	74107 characters	Kannada and English Alphabets & Printed and Hand written characters	Shape based (Adaptive center distance based features, Fixed center Triangular Cut-based Features, Neighborhood Counts-based Features)	Chars74K benchmark data set	None	98.56
[17]	19530 characters	Bengali Alphabets & Printed characters	Shape based (Various types of HOG features)	ISI_Bengali benchmark data set	None	92.20
[18]	2900 characters	Bangla Alphabets & Printed characters	Shape based (Contour distance based features which are invariant to scale and rotation but sensitive to distortion and breaks)	Proprietary data set	Scale and Rotation Invariant	97.80
[18]	3100 characters	Devanagari Alphabets & Printed characters	Shape based (Contour distance based features which are invariant to scale and rotation but sensitive to distortion and breaks)	Proprietary data set	Scale and Rotation Invariant	98.10
[18]	6584 characters	English Alphabets and Numerals & LP characters (Printed)	Shape based (Contour distance based features which are invariant to scale and rotation but sensitive to distortion and breaks)	MediaLab LP benchmark data set	Scale and Rotation Invariant	95.40
[1]	74107 characters	Kannada and English Alphabets & Printed and Hand written characters	histogram oriented Gabor features, grid level features (local gradient), and gray level co-occurrence matrix	Chars74K benchmark data set	None	96.00
[7]	37440 characters	Urdu ligatures	raw moments, central moments, hu moments, and Zernike moments	Proprietary data set	Scale and Rotation invariant	96.92
[12]	1605 characters	Gurumukhi characters	Zoning features, diagonal features, and parabola curve fitting features	Proprietary data set	None	96.19

Proposed method	30000 characters	Devanagari, Marathi, and English & Printed characters	Shape based (Geometrical distance based features which are invariant to scale, rotation, distortion, and breaks)	Proprietary data set	Scale and Rotation Invariant	98.80
Proposed method	6584 characters	English Alphabets and Numerals & LP characters (Printed)	Shape based (Geometrical distance based features which are invariant to scale, rotation, distortion, and breaks)	MediaLab LP benchmark data set	Scale and Rotation Invariant	99.20
Proposed method	19530 characters	Bengali characters & Printed characters	Shape based (Geometrical distance based features which are invariant to scale, rotation, distortion, and breaks)	ISI_Bengali benchmark data set	Scale and Rotation Invariant	97.72
Proposed method	74107 characters	Kannada and English Alphabets & Printed and Hand written characters	Shape based (Geometrical distance based features which are invariant to scale, rotation, distortion, and breaks)	Chars74K benchmark data set	Scale and Rotation Invariant	98.64

V. Conclusion

In this paper, we have proposed a novel geometrical scale and rotation independent FE technique for multi-lingual CR with the help of various sweep lines. Along with the proposed method we have generated CC features with the help of sweep lines. We have combined both the FE techniques to form as FV of the character to be recognized. The proposed FE technique recognized the multi-lingual characters with noise such as distortion and breaks as shown in Fig. 5. The proposed FE technique has the ability to recognize characters accurately which are similar in shape from various languages. It is evidence from the results of the Tables II, III, and IV that the proposed FE technique outperformed many methods from literature using proprietary data set and various publicly available benchmark data sets. The FE technique proposed in this paper works on any kind of script. The limitation of the proposed method is that, it failed to extract the SRI features properly if the image is too small to distinguish with naked eye.

We have observed that there is still lot of scope in proposing novel robust scale and rotation independent FE techniques for Omni-font character recognition [16] which can recognized similarly looking characters from various multi-lingual languages and can be trained using various types of neural networks which combines the traditional way of extracting the features and new way of training and testing the proposed methods.

Acknowledgment

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] A. K. Sampath, and N. Gothami, “Decision tree and deep learning based probabilistic model for character recognition,” J. of Cent. South Univ., Vol. 24 , No. 12, pp. 2862−2876, Dec. 2017.

[2] C. Singh, E. Walia, and N. Mittal, “Rotation invariant complex Zernike moments features and their applications to human face and character recognition,” IET Comput. Vis., Vol. 5, No. 5, pp. 255-265, Oct. 2011.

[3] H. Hase, T. Shinokawa, M. Yoneda, and C. Y. Suen, “Recognition of rotated characters by Eigen-space,” Proc. of 7th Int. Conf. on Doc. Anal. and Rec., Edinburgh, UK, pp. 731-735, 2003.

[4] H. Takayuki, and T. Noboru, “A Consideration on rotation invariant character recognition,” 2006 World Automation Congress (WAC), Budapest, Hungary, pp. 1-6, July 2006.

[5] I. E. Anagnostopoulos, “MediaLab license plate database,” Multimedia Technology Laboratory, National Technical University of Athens, http: //www.medialab.ntua.gr/ research/ LPRdatabase.htm.

[6] J. A. Montoya-Zegarra, J. P. Papa, N. J. Leite, R. D. S. Torres, and A. X. Falc ˜ao, “Learning how to extract rotation-invariant and scale-invariant features from texture images,” EURASIP J. on Adv. Signal Process, Vol. 2008, no. 1, pp. 1-15, Jan. 2008.

[7] K. U. Rehman, and Y. D. Khan, “A scale and rotation invariant Urdu nastalique ligature recognition using cascade forward backpropagation Neural Network,” IEEE Access, No. 7, pp. 120648-120669, Aug. 2019.

[8] L. A. Torres-Méndez, J. C. Ruiz-Suárez, Luis E. Sucar, and G. Gómez, “Translation, rotation, and scale-invariant object recognition,” IEEE Trans. on Syst., Man, and Cybern.—Part C: Appl. and Reviews, Vol. 30, No. 1, pp. 125-130, Feb. 2000.

[9] N. Tripathy, T. Chakraborti, M. Nasipuri, and U. Pal, “A scale and rotation invariant scheme for multi-oriented character recognition,” Proc. Of 23rd Int. Conf. on Pat. Rec. (ICPR), Cancún Center, Cancún, México, pp. 4041-4046, Dec. 2016.

[10] P. Sahare, and S. B. Dhok, “Robust character segmentation and recognition schemes for multilingual Indian document images,” IETE Tech. Review, Vol. 36, No. 2, pp. 209-222, Apr. 2018.

[11] R. D. Zarro, and M. A. Anwer, “Recognition based on online Kurdish character recognition using hidden Markov model and harmony search,” Eng. Sci. and Technol. an Int. J., Vol. 20, No. 2, pp. 783-794, Apr. 2017.

[12] R. P. Kaur, M. Kumar, and M. K. Jindal, “Recognition of newspaper printed in Gurumukhi script,” J. Cent. South Univ., Vol. 26, No. 9, pp. 2495-2503, Oct. 2019.

[13] S. Narasimha Reddy, and P. S. Deshpande, “Robust feature extraction technique for license plate characters recognition,” IETE J. of Res., Vol. 61, No. 1, pp. 72-79, Dec. 2014.

[14] S. Narasimha Reddy, and P. S. Deshpande, “Color, scale, and rotation independent multiple license plates detection in videos and still images,” Math. Prob. In Eng., pp. 1-14, June 2016.

[15] S. Narasimha Reddy, and P. S. Deshpande, “Novel geometrical shape feature extraction techniques for multilingual character recognition,” IETE Tech. Review, Vol. 34, No. 6, pp. 612-621, Oct. 2016.

[16] S. Narasimha Reddy, and P. S. Deshpande, “Review of feature extraction techniques for character recognition,” IETE J. of Res., Vol. 64, No. 2, pp. 280-295, Jul. 2017.

[17] S. Tian, U. Bhattacharya, S. Lu, B. Su, Q. Wang, X. Wei, Y. Lu, and C. L. Tan, “Multilingual scene character recognition with co-occurrence of histogram of oriented gradients,” Patt. Recogn., Vol. 51, No. 3, pp. 125–34, June 2015.

[18] U. Pal, and N. Tripathy, “A contour distance-based approach for multi-oriented and multi-sized character recognition,” Sadhana, Vol. 34, No. 5, pp. 755–765, Oct. 2009.

[19] J. Chakey, and N. Day, “Fragmented handwritten digit recognition using grading scheme and fuzzy rules,” Sadhana, Vol. 45, No. 1, pp. 1-23, Aug. 2020.

Harmonic Mean based Classification of Images using Weighted Nearest Neighbor for Tagging

Mrs. Anupama D. Dondekar[1], Mr. Balwant A. Sonkamble[2]
Computer Engineering Department, Pune Institute of Computer Technology
Pune, India

***Abstract*—On image sharing websites, the images are associated with the tags. These tags play a very important role in an image retrieval system. So, it is necessary to recommend accurate tags for the images. Also, it is very important to design and develop an effective classifier that classifies images into various sematic categories which is the necessary step towards tag recommendation for the images. The performance of existing tag recommendation based on k nearest neighbor methods can be affected due to the number of k neighbors, distance measures, majority voting irrespective of the class and outlier present in the k-neighbors. To increase the accuracy of the classification and to overcome the issues in existing k nearest neighbor methods, the Harmonic Mean based Weighted Nearest Neighbor (HM-WNN) classifier is proposed for the classification of images. Given an input image, the HM-WNN determines k nearest neighbors from each category for color and texture features separately over the entire training set. The weights are assigned to the closest neighbor from each category so that reliable neighbors contribute more to the accuracy of classification. Finally, the categorical harmonic means of k nearest neighbors are determined and classify an input image into the category with a minimum mean. The experimentation is done on a self-generated dataset. The result shows that the HM-WNN gives 88.01% accuracy in comparison with existing k-nearest neighbor methods.**

Keywords*—Image classification; k-nearest neighbor; weighted nearest neighbor; harmonic mean vector; color and texture features

I. Introduction

Classification is a supervised method that categorizes the unknown data into a specific class or group. In classification, the labels or classes are known in advance. Image classification is used for indexing, categorization and annotation of the images. To classify images into various categories, the features of the images are extracted using different feature extraction techniques and stored in the database with the label. For training the classifier, the images stored in the database are divided into two sets-training and testing images. The classifier learns from the labeled training images stored in a database and predicts the class label of an unlabeled test image.

The k nearest neighbor (kNN) is a memory-based and non -parametric classification algorithm where the algorithm memorizes the training samples to predict the class label of an unknown sample. It is the most widely used classifier in machine learning because it is a very simple, effective and powerful recognition algorithm [1]. In a traditional k nearest neighbor classifier the similarity between unknown sample and training samples are calculated. After similarity calculation, it selects k training samples that are closest to an unknown sample. The class label of an unknown sample is predicted by identifying the dominant category from k training samples. For kNN, the value of the k needs to be specified in advance and it denotes the number of samples that are nearest to an unknown sample. If the k value is too small it is susceptible to overfitting and would result in misclassification.

The steps of the kNN classifier algorithm are as follows:

1) Determine the value of k (no of neighbors).
2) Extract features of a test sample and calculate the distance between the feature vector of the training samples and test sample.
3) Sort the samples as per the distance in ascending order and select the first top k samples.
4) Get the majority voting from top k samples.
5) Predict the class with the majority voting.

However, the traditional k nearest neighbor classifier has the following limitations [2]:

1) It uses only a training set for the classifier model generation. It needs retraining if there is a change in training data.
2) It does not consider the situation in which the distribution of samples may not be equal. All training samples have equal weights.
3) Need to calculate the similarity between all training samples. It takes more time if the training set consists of a large number of samples.
4) The performance of the classifier depends on the number of neighbors (k). The small value of k affects classification results due to noisy and inaccurate samples. The large value of k degrades the performance of classifier due to outliers which comes from incorrect category in the k nearest neighbors.
5) It suffers from outliers if the size of the training sample size is small.

To determine the correct value of k and to reduce the sensitivity of the k value, the authors have proposed adaptive methods to determine the neighborhood for different input samples in [2, 3, 4]. In most of the kNN algorithms, all samples have equal weight but each neighbor may contribute differently in the classification. In [5, 6, 7] the method was

proposed where each nearest sample was assigned varying weights. In [8] the modified kNN method was proposed by assigning weight to the neighbor based on the local value of k. The weighted kNN was used in [9] where inverse Euclidean distance assigned as the weight to each neighbor. But, no weighting scheme perform consistently well under some condition.

To handle the outlier problem, the LMKNN classifier was proposed in [10]. The LMKNN classifier determines the neighbor of an input sample based on the difference between the mean values of the nearest training samples and an input sample. It has good classification accuracy but it has some limitations: the same value of k is used for different classes and all neighbors have uniform weights. To improve the accuracy of the LMKNN method, many local mean-based methods for classification have been proposed. In [11] the PNN method was proposed for classification based on the similarity between the weighted distance of an input sample and pseudo training samples. The LMPNN classifier was proposed in [12] where the weights are assigned to each local mean vector and distance was calculated between mean vector and pseudo neighbors of every class. The MLM-KHNN classification method was proposed in [13] which has used harmonic mean distance instead of Euclidean distance. The LMRKNN method was proposed in [14] where the input sample is represented as a combination of the mean vector of each category and weights are assigned to each local mean vector. In [15] GMDKNN method was proposed which predict the class label based on nested generalized mean distance.

To improve the accuracy of kNN based classification, the harmonic mean based weighted nearest neighbor classifier is proposed which is insensitive to the value of k. Instead of assigning weights to the local mean vector, the weights are assigned to the k nearest neighbor of each class for each feature instead of combined features. The reason behind using separate classifier for each feature is that it may contribute differently to determine number of neighbors depends on input sample which affects the classification accuracy. Also, the harmonic mean is used as it gives more importance to the neighbors closer to the given input image as compared to arithmetic mean.

The work done carried out in the paper can be summarized as follows:

1) In HM-WNN classifier, the separate kNN classifiers are trained for color and texture features and harmonic mean is determined for each category for every feature to reduce the impact of the choice of value k.

2) The HM-WNN classifier assigns different weights to each neighbor of each class instead of uniform weights to reduce the classification error rate.

The structure of the paper is organized as follows: a review of the existing algorithms for classification of the images is described in section 2, section 3 describes feature extraction technique of the images, HM-WNN proposed classifier is explained in section 4, section 5 and section 6 describes performance metric used to check classifier performance and dataset used for experimentation respectively. The classification results are discussed in section 7. Finally, the conclusion and future work is described in section 8.

II. Related Work

The method was proposed for image annotation refinement using a two-pass kNN classifier and group sparse reconstruction algorithm in [16]. The method for annotation of images using visual attention mechanism and SVM particle swarm optimization was presented in [17]. The method to suggest tags for an image based on visual features and tag correlations using a neighbor voting scheme was proposed in [18]. The kNN algorithm is used to find visual neighbors of a given image. The approach for retagging the social images with diverse semantics was presented in [19]. The relevance of a tag to an image was determined using three approaches. The kNN classifier was used to map the tag from textual space to low-level feature space by representing the tag with a set of images containing the tag. The method was proposed for annotation of images using multiple SVM classifiers which combine different visual features and improve the image annotation performance in [20]. The histogram intersection distance was used as the kernel function to reduce the impact of intra-class variations and interclass similarities. A personalized image tag recommendation approach by using users' history was presented in [21]. The approach recommends tags to the users by counting the occurrence of each tag from the visually similar images where the similarities between the images are determined using kNN classifier. In [22] the method was proposed which suggests the tags for labeled as well as unlabeled images. The method identifies similar images based on visual features solely using kNN classifier. A system that classifies images as correct or incorrect with respect to a given tag from a database of annotated images was defined in [23, 24]. The system trained SVM classifier per tag to enable fast classification. In [25] `SheepDog' method was proposed which adds photos into the proper group and also recommends suitable tags for the photos of the Flickr dataset. To predict correct tags of a photo the concepts were detected using SVM classifier. The method for annotation of images using random forest is presented in [26]. The methods find semantic neighbors of a query image from each leaf node and annotate the images using tags of semantic neighbors.

Most of the existing methods for image tag recommendation used a single classifier for the recommendation of tags. A little work is done on the selection of suitable classifier for the purpose of tag recommendation of the images. Also, the tag recommendation methods based on k nearest neighbor has used uniform weight and the same number of neighbors. In the paper, a new classifier is proposed and compared with the existing algorithms as a necessary step in an image tag recommendation.

III. Image Feature Extraction

Color and texture are the most widely used features in retrieval of the images. Color moments are used in color feature representation. The advantage of the color moment method is that it works better when the some region of the two images are similar because the low order moments of the

corresponding region of two images will be different and the similarity score will be small. The color feature extraction is done as follows: i) convert color space of an image into L*a*b* color space ii) divide an image into 2 by 2 blocks. Also, the central image is obtained of the same size as 2 by 2 blocks. ii) The first, second and third moment of an each block are taken as feature vector. For texture feature extraction, the image is decomposed into sub-bands using wavelet packet transform up to level 3 and calculates the energy and its variation of each sub-band at the last level as a feature vector. Both features are combined by assigning weight to them [27]. The color and texture feature values are not in the same scale. Both features are normalized using Min-Max normalization to scale the values in range 0 and 1.

IV. HM-WNN Classification Method

In this section HM-WNN classification is described. The main objective of the proposed algorithm is to improve classification accuracy and to assign different weights to each neighbors of different category to reduce the sensitivity to outlier.

A. Classification

Let $CF = (x_i, C_i)_1^{NF}$ and $TF = (y_i, C_i)_1^{NF}$ represents color and texture features vector of *NF* training images with N number of classes. Let C_1, C_2,…,C_N represent the category. Let $CF_j = (x_i^j, C_i^j)_1^{NF_j}$ and $TF_j = (y_i^j, C_i^j)_1^{NF_j}$ denotes color and texture feature vector of j^{th} category with NF_j training images. In HM-WMM, given an input image first m nearest neighbors are determined from CF_j and TF_j for each category instead of finding m nearest neighbor from entire NF number of training images. The weights are assigned to each neighbor and the harmonic mean is calculated for each category for each feature and combined. Finally, the category j is assigned to an input image with minimum mean.

The steps for prediction of the class of an input image I are as follows:

1) Determine the nearest/closest neighbors for each class C_j from CF_j and arrange m closest neighbors in increasing order according to the distance measure.

2) Determine n nearest/closest neighbors for each class C_j from TF_j and arrange n closest neighbors in increasing order according to the distance measure.

3) Allocate weight to each neighbor of m and n as follows:

$$w_i^j = \frac{1}{i}\ i = 1, \dots\dots, z\ \&\ z \in (m, n) \tag{1}$$

Where w_l^j represents weight assigned to i^{th} neighbor of j^{th} class. It gives more weight to the samples which are nearby and less weight to the samples which are farther away.

4) Compute the harmonic mean of m and n closest neighbor of color and texture for each class C_i.

5) Combine the mean of each class determined for color and texture features to form final mean.

6) Assign the class label C to I with minimum mean distance belongs to among all classes.

B. HM-WNN Algorithm

As explained in section A, the proposed HM-WNN method is brief in Algorithm 1.

Algorithm 1: Harmonic Mean based weighted Nearest Neighbor classifier

Input:

CI: Input image color feature vector
TI: Input image texture feature vector
$TF_j = (y_i^j, C_i^j)_1^{NF_j}$: Texture feature vector of training images of the j^{th} class
$CF_j = (x_i^j, C_i^j)_1^{NF_j}$: Color feature vector of training images of the j^{th} class
$M = C_1, C_2, \dots\dots C_M$: the number of class labels

Result: Prediction of class C_j of an input image I

Step 1: Calculate the distance between CI and CF_j. Find m number of nearest neighbor for each class C_j according to the ascending order of distances to CI. Denote it by CD_j.

for i= 1 to NF_j

$$dist(CI, x_i^j) = \sqrt{(CI_1 - x_{i1}^j)^2 + \dots + (CI_f - x_{if}^j)^2} \tag{2}$$

end for

where $x_i^j \epsilon\ CF_j$

Step 2: Calculate the distance between TI and TF_j. Find n number of nearest neighbor for each class C_j according to the ascending order of distances to TI. Denote it by TD_j.

for i= 1 to NFj

$$dist(TI, y_i^j) = \sqrt{(TI_1 - y_{i1}^j)^2 + \dots + (TI_f - y_{if}^j)^2} \tag{3}$$

end for

Where $y_i^j \epsilon\ TF_j$

Step 3: Assign weight to the i^{th} nearest neighbor of j^{th} class for CD_j and TD_j using Eq. 1 as follows:

$$Cdist(CI, x_i^j) = dist(CI, x_i^j) * w_i^j \tag{4}$$

$$Tdist(TI, y_i^j) = dist(TI, y_i^j) * w_i^j \tag{5}$$

Step 4: Compute the harmonic mean of first m and n nearest neighbor of CD_j and TD_j for the j^{th} class

$$CD\_Mean_j = \frac{m}{\sum_1^n \frac{1}{Cdist(CI, x_i^j)}} \tag{6}$$

Where CD_Mean_j represents color feature mean of class C_j

$$TD\_Mean_j = \frac{n}{\sum_1^n \frac{1}{Tdist(TI, y_i^j)}} \tag{7}$$

Where TD_Mean_j represents texture feature mean of class C_j

Step 5: Calculate the final mean vector as follows:

$$F\_Mean_j = CD\_Mean_j * TD\_Mean_j \tag{8}$$

Step 6: Assign I a class label C which has smallest mean calculated for each class as follows:

$$C = \min_{c_j}(F\_Mean_j, I) \tag{9}$$

V. Performance Matrices

The performance of the classifier is evaluated using confusion matrix on a test data. The confusion matrix analyses the test data to determine how classifier identifies test samples of different categories. It consists of count of predicted values and actual values. By using these count of values the precision, recall, F1-score and accuracy are determined to estimate the performance score of the classifier.

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

$$F1 - Score = \frac{2*precision*recall}{precision+recall} \quad (12)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (13)$$

TP = Test Sample is positive and predicted positive.

TN = Test Sample is negative and predicted negative.

FP = Test Sample is negative and predicted positive.

FN = Test Sample is positive and predicted negative.

VI. Dataset

For experimentation the images are downloaded from Flickr image sharing website using public API. The images are belongs to six different categories actor, clover, fish, autumn, butterfly and aeroplane. Each category consists of 300 images. The 70% of the images are used for training and 30% are used for testing purpose.

VII. Experimental Results

Tables I and II shows the precision, recall, F1-score and accuracy of different classification methods: SVM, traditional kNN, LMKNN, PNN, LMPNN and proposed method HM-WNN.

To implement SVM, m*(m-1)/2 binary classifiers are composed where m is the no of classes/categories and m=6 and one vs. all method is used. In LMKNN, k neighbors are determined for each category and calculated the mean vector of each class [6]. The class label of an input sample is predicted having minimum distance between mean vectors of each among all classes. In PNN, weights are assigned to each neighbor and assign the class of neighbor nearest to an input sample [7]. In LMPNN method, first local mean vector is determined for each category and weights are assigned to each local mean vector [8].

In F-WNN algorithm the features (color & texture) are combined. The nearest neighbors are determined for the combined features and harmonic mean vector of the combined features are calculated and finally predicted the class of an input image I. The HM-WNN shows the performance of proposed method where the nearest neighbors for color feature and texture feature vector are determined separately. We check the value of k between 1 and 50 to identify the correct value of k in order to determine no of neighbors. The value of k at which the highest accuracy is obtained on dataset is selected within the range.

TABLE I. Precision and Recall of different Classification Methods

Method	Precision (%)	Recall (%)
SVM	86.16	86.16
kNN	84.33	84.50
LMKNN	86.16	86.50
PNN	82.50	82.33
LMPNN	84.50	83.50
F-WNN	87.16	87.16
HM-WNN	87.73	88.66

TABLE II. F1-Score and Accuracy of different Classification Methods

Method	F1-Score (%)	Accuracy (%)
SVM	86	85.92
kNN	84	83.87
LMKNN	85.66	85.83
PNN	81.83	82.13
LMPNN	82.66	82.78
F-WNN	86.50	86.50
HM-WNN	88.00	88.01

From Tables I and II, it is observed that the proposed classifier based on harmonic mean vector determined from nearest neighbors of each features separately gives good results for classification of tagged images as compared to the SVM, traditional kNN, LMKNN, PNN, LMPNN classification algorithm. Also, classification error rate of the HM-WNN method is less than other classification algorithms as shown in Fig. 1.

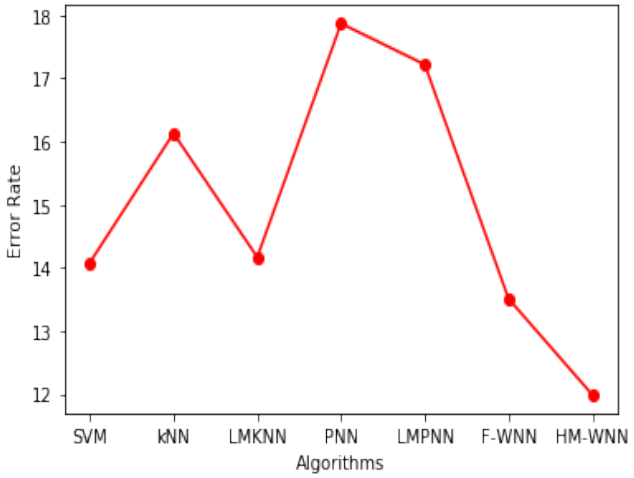


Fig. 1. Error Rate of different Classification Methods.

VIII. Conclusion

In the paper, HM-WNN classifier is proposed for classification of the images. The main objective of the HM-WNN is to improve classification accuracy by creating a separate classifier for each feature and combining the

harmonic mean obtained by classifier designed for each feature. The proposed method also used the information of neighbors per category over the entire training samples. The experimental result shows that the HM-WNN provides good performance as compared to the existing kNN methods. In future we need to develop a method to assign tags to the new images using tags assigned to the visually similar images.

REFERENCES

[1] Cover, Thomas, and Peter Hart. "Nearest neighbor pattern classification." IEEE transactions on information theory 13, no. 1 (1967): 21-27.

[2] Zhang, Shichao, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Ruili Wang. "Efficient knn classification with different numbers of nearest neighbors." IEEE transactions on neural networks and learning systems 29, no. 5 (2017): 1774-1785.

[3] Zhang, Shichao, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Debo Cheng. "Learning k for knn classification." ACM Transactions on Intelligent Systems and Technology (TIST) 8, no. 3 (2017): 1-19.

[4] García-Pedrajas, Nicolás, Juan A. Romero del Castillo, and Gonzalo Cerruela-García. "A proposal for local k values for k -nearest neighbor rule." IEEE transactions on neural networks and learning systems 28, no. 2 (2015): 470-475.

[5] Li, Ping, Jianping Gou, and Hebiao Yang. "The distance-weighted k-nearest centroid neighbor classification." J. Intell. Inf. Hiding Multimedia Sig. Process 8, no. 3 (2017): 611-622.

[6] Gou, Jianping, Lan Du, Yuhong Zhang, and Taisong Xiong. "A new distance-weighted k-nearest neighbor classifier." J. Inf. Comput. Sci 9, no. 6 (2012): 1429-1436.

[7] Dudani, Sahibsingh A. "The distance-weighted k-nearest-neighbor rule." IEEE Transactions on Systems, Man, and Cybernetics 4 (1976): 325-327.

[8] Liu, Shenglan, Ping Zhu, and Sujuan Qin. "An improved weighted KNN algorithm for imbalanced data classification." In 2018 IEEE 4th International Conference on Computer and Communications (ICCC), pp. 1814-1819. IEEE, 2018.

[9] Fan, Guo-Feng, Yan-Hui Guo, Jia-Mei Zheng, and Wei-Chiang Hong. "Application of the weighted k-nearest neighbor algorithm for short-term load forecasting." Energies 12, no. 5 (2019): 916.

[10] Mitani, Yoshihiro, and Yoshihiko Hamamoto. "A local mean-based nonparametric classifier." Pattern Recognition Letters 27, no. 10 (2006): 1151-1159.

[11] Zeng, Yong, Yupu Yang, and Liang Zhao. "Pseudo nearest neighbor rule for pattern classification." Expert Systems with Applications 36, no. 2 (2009): 3587-3595.

[12] Gou, Jianping, Yongzhao Zhan, Yunbo Rao, Xiangjun Shen, Xiaoming Wang, and Wu He. "Improved pseudo nearest neighbor classification." Knowledge-Based Systems 70 (2014): 361-375.

[13] Pan, Zhibin, Yidi Wang, and Weiping Ku. "A new k-harmonic nearest neighbor classifier based on the multi-local means." Expert Systems with Applications 67 (2017): 115-125.

[14] Gou, Jianping, Wenmo Qiu, Zhang Yi, Yong Xu, Qirong Mao, and Yongzhao Zhan. "A local mean representation-based K-nearest neighbor classifier." ACM Transactions on Intelligent Systems and Technology (TIST) 10, no. 3 (2019): 1-25.

[15] Gou, Jianping, Hongxing Ma, Weihua Ou, Shaoning Zeng, Yunbo Rao, and Hebiao Yang. "A generalized mean distance-based k-nearest neighbor classifier." Expert Systems with Applications 115 (2019): 356-372.

[16] Ji, Qian, Liyan Zhang, Xiangbo Shu, and Jinhui Tang. "Image annotation refinement via 2P-KNN based group sparse reconstruction." Multimedia Tools and Applications 78, no. 10 (2019): 13213-13225.

[17] Hao, Zhangang, Hongwei Ge, and Long Wang. "Visual attention mechanism and support vector machine based automatic image annotation." PloS one 13, no. 11 (2018): e0206971.

[18] Cui, Chaoran, Jialie Shen, Jun Ma, and Tao Lian. "Social tag relevance learning via ranking-oriented neighbor voting." Multimedia Tools and Applications 76, no. 6 (2017): 8831-8857.

[19] Qian, Xueming, Xian-Sheng Hua, Yuan Yan Tang, and Tao Mei. "Social image tagging with diverse semantics." IEEE transactions on cybernetics 44, no. 12 (2014): 2493-2508.

[20] Wu, Wei, Jianyun Nie, and Guanglai Gao. "An improved SVM-based multiple features fusion method for image annotation." Journal of Information & Computational Science 11, no. 14 (2014): 4987-4997.

[21] Qian, Xueming, Xiaoxiao Liu, Chao Zheng, Youtian Du, and Xingsong Hou. "Tagging photos using users' vocabularies." Neurocomputing 111 (2013): 144-153.

[22] Li, Xirong, Cees GM Snoek, and Marcel Worring. "Learning social tag relevance by neighbor voting." IEEE Transactions on Multimedia 11, no. 7 (2009): 1310-1322.

[23] Li, Xirong, and Cees GM Snoek. "Classifying tag relevance with relevant positive and negative examples." In Proceedings of the 21st ACM international conference on Multimedia, pp. 485-488. 2013.

[24] Shen, Yi, and Jianping Fan. "Leveraging loosely-tagged images and inter-object correlations for tag recommendation." In Proceedings of the 18th ACM international conference on Multimedia, pp. 5-14. 2010.

[25] Chen, Hong-Ming, Ming-Hsiu Chang, Ping-Chieh Chang, Ming-Chun Tien, Winston H. Hsu, and Ja-Ling Wu. "Sheepdog: group and tag recommendation for flickr photos by automatic search-based learning." In Proceedings of the 16th ACM international conference on Multimedia, pp. 737-740. 2008.

[26] Fu, Hao, Qian Zhang, and Guoping Qiu. "Random forest for image annotation." In European Conference on Computer Vision, pp. 86-99. Springer, Berlin, Heidelberg, 2012.

[27] Anupama D. Dondekar and Balwant A. Sonkamble, "Analysis of Flickr Images using Feature Extraction Techniques," 4th IEEE International Conference on Computer and Communication Systems (ICCCS 2019), Singapore, pp. 278-282.

A Design Study to Improve User Experience of a Procedure Booking Software in Healthcare

Hanaa Abdulkareem Alzahrani[1], Reem Abdulaziz Alnanih[2]
Computer Science Department, Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—In the era of technology-driven healthcare delivery and the proliferation of e-health systems, procedure booking software is becoming common. Procedure booking software (PBS) affects healthcare delivery by improving health care efficiency and outcomes, while cutting costs. Therefore, poor software design for PBS, especially if it is designed for important and critical appointments such as cardiac catheterization operations, creates stress for physicians and may result in their rejection of this technology. Moreover, if the system design forces them to spend more time documenting health information, physicians would then tend to prefer face-to-face interaction with patients. Software with poor usability increases the workload of physicians thus reducing system efficiency. So designing a useful and effective web user interface for such software is an essential requirement for health websites. The aim of this paper is to design and develop a PBS as a case study using the health systems design (HSD) tool. HSD is a validated design tool for creating PBS based on physician behavior and persona. The applicability of a PBS design is explored by physicians evaluated. The PBS design was evaluated in terms of objective and subjective characteristics and user experience attributes. Test participants were divided into two groups: specialists and fellows. The results show that there was no significant difference between participants in either group. All were able to complete the tasks successfully with a minimum amount of time, clicks, and errors indicating that the effectiveness, efficiency and cognitive load were similar for all participants. User satisfaction yielded a score of 86 on the System Usability Scale (SUS), putting it in the A Grade. Also, user experience attributes demonstrated that participants were satisfied using the proposed design system.**

Keywords*—*Procedure booking software; health systems design tool; cardiac catheterization; user experience; usability evaluation; system usability scale

I. Introduction

In the emerging era of technologically advanced health care and the proliferation of e-health systems, effective procedure booking software (PBS) is imperative [1]. PBS impacts health-care delivery by improving health-care efficiency, reducing costs and improving health findings [2]. Designing a useful and effective web user interface (UI) for such software is an essential requirement for health websites. Even a minor change in a UI could lead to usability problems. For example, changing the background or foreground color of buttons could create difficulties for physicians. One of the reasons physicians are spending more time documenting patient health information is poor UI design for PBS which could lead them to reject the technology and prefer face-to-face interactions with patients [3].

Using a system for the first time, especially an outdated system, can be challenging. Unfortunately, it can be very difficult to replace or upgrade specific-purpose systems that require retraining physicians and integrating it with other services. Therefore, often new physicians asked to learn outdated systems that aligned with UI difficult to use.

The UI should be designed based on proven design principles, such as consistency between pages, ease of navigation, and also be eye-catching and user-friendly. All these characteristics are included in the usability definition. Usability refers to the degree to which a specific user often uses a product to attain specific aims with effectiveness, efficiency and satisfaction in a specific use context [4].

In the pediatric cardiology department at King Abdulaziz University Hospital (KAUH), the booking for cardiac catheterization is paper-based and can only be viewed and modified by visiting the clinic. There is no communication system between the cardiologist, the calendar and the hospital information data. As a result, cancellation or adjustment is likely to cause confusion in the cardiac catheterization team and even loss of bed appointments. Taking healthcare user experience (UX) into account in the creation of software systems can provide substantial benefits for the creation of successful healthcare systems, and also improve user overall well-being. However, designers must be familiar with medical expressions and understand human cognition [5].

The importance of this research is to propose a solution to the problem of health systems that are designed without taking into account the users requirements, preferences and goals in terms of speed and ease of use. The emergence of a UX that focuses on user requirements as the basis for designing UI, is the ideal solution for this problem.

The purpose of this study is to design an adaptable UI for PBS based on healthcare UX using the proposed health system design (HSD) tool. HSD is a tool-based website, whose design is based on physician persona and behavior. This tool maps healthcare personae from real life to the design space for healthcare UI systems, especially for PBS. The proposed PBS UI is evaluated and validated as proof of concept for the healthcare domain based on physician persona and behavior. This will assist the software designer to optimize UX and create persuasive software by understanding the type of behavior that needs to be considered when designing a UI [6].

The research question address in this study is, how can special software be designed for physicians? And how can its

design be validated to measure physician work quality who lack a fundamental interest in technology?

The objectives of this research are as follow:

1) To create a PBS using the HSD tool (Sections III and IV).

2) To validate the PBS design and measure the quality of physician work as follows:

a) Measure objective attributes in terms of effectiveness, efficiency, and cognitive load.

b) Measure subjective characteristics in terms of the System Usability Scale (SUS).

c) Measure UX attributes in terms of completing the task - the task's success and errors per task.

The rest of the paper is organized as follows: Section II presents the literature review. Section III describes the HSD tool. In Section IV details the PBS method and materials process. In section V the experimental design and evaluation are discussed and the results are illustrated in section VI. Finally, the conclusion and future of the work are presented in section VII.

II. Literature Review

In this section, the related literature is explored from two main streams, booking systems and usability evaluation.

A. Booking Systems

Web-based booking systems are important for reducing errors that can arise using paper-based systems. Researchers Zhao et al. searched PubMed seeking to identify advantages and barriers to implementation of web-based medical scheduling [7]. A total of 36 articles discussing 21 web-based appointment systems were selected. The results of this review suggests there are advantages to a variety of patient outcomes from web-based scheduling such as reducing no-show rates, waiting time, and staff work, while improving satisfaction. Otherwise, as barriers, patient reluctance to adopt web-based appointment scheduling is mainly affected by past experience with computer technology and the Internet. Primary and specialty clinics use appointment scheduling systems to manage access to service providers, and hospitals can also use an appointment system to plan elective surgery.

Gupta & Denton described the most common healthcare delivery systems paying particular attention to the many factors that make appointment arrangements more challenging [8]. Factors including variability of arrival and service times, patient and provider preferences, available information technology, and the scheduling staff's level of experience [8]. In addition, the key bottleneck was determined to lie in the application of Industrial Engineering and Operations Research (IE/OR) technology. They provided the latest technology roadmap in the design of appointment management systems and identified future opportunities for novel applications of IE/OR model [8].

In Taiwan, Yang et al. revealed that more than half of hospitals have public online booking systems. However, they note that most systems perform only the registration function and rarely seize the opportunity to collect other information, such as related medical history or reasons for consultation [9]. They indicated further efforts should be made to strengthen the functions of these online booking systems in order to improve the efficiency of consultation. Some methods for information extraction and retrieval of unstructured medical records are needed to improve the efficiency of the appointment process.

Gamma et al. regard design patterns as a general solution to the recurring problems in software development, and therefore provides basic support to deal with these problems [10]. Appling the patterns of human-computer interaction provides basic design support and enhances development of the UI. The UI development model remains an important key theme. However, further research is needed to optimize its application.

Sinnig et al. applied the pattern concept to the analysis design of UI for online booking applications [11]. The application was tested with a usability evaluation to discover any other usability issues, thus optimizing its use. Their paper constitutes a step forward to integrating UI design patterns into the software development process.

According to Murray et al., the measurement of the performance of health systems relates goal attainment to the resources available [12]. They show that variety in performance is a function of how a health system organizes four key functions: stewardship, financing, service supply, and resource generation [12]. By investigating these four functions and how they are combined, they show it is possible to observe major policy challenges and understand the proximate determinants of health system performance.

The difference between the proposed approach and today's prevailing systems is: the proposed approach is based on user personas that plays an important role in the design orientation. Also, the proposed approach includes user's behavior. By considering persona and behavior that helps to understand the mental language of the target users and makes the whole experience in a systematic way quite humane.

B. Usability Evaluation

In 1998, the International Standards Organization released an original document containing requirements that describe the extent to which specific types of users can use a system to achieve specific goals of effectiveness, efficiency and satisfaction, as shown below [13]:

- Effectiveness: The completeness and accuracy in achieving a specific goal.
- Efficiency: Resources are used to improve the accuracy and completeness of users in achieving their goals.
- Satisfaction: Users don't feel any discomfort and have a positive attitude when using this product.

DeLone and Ephraim [14] described success terms of an information system (IS) defined as "a purposeful entity composed of interdependent computer-based technology and human components unified by design to accomplish one or more objectives." Interface satisfaction is defined as a dimension that captures the user's general satisfaction with the interface in terms of presentation, format, usability, and efficiency [14]. Decision support satisfaction is the level of

satisfaction a user has with the help provided by the IS tool in planning or controlling a business process. Operational efficiency focuses on improvement in internal customer performance and is estimated with respect to flexibility, productivity, consistency, and process duration. Quality of work-life satisfaction is the level of user satisfaction as a result of the impact of the IS on their emotions, physical requirements, personal goals, and psychological states [14]. Finally, task support satisfaction is "the level of user satisfaction from the help provided by the IS tool toward the goal of achieving job and task requirements.

J. Rinder in her dissertation examined literature investigating website usability testing [15]. She reviewed 31 peer-reviewed articles, conference proceedings, and books identifying 10 usability testing categories, across a variety of testing methods, she noted that the most frequently listed categories addressed (a) navigation, (b) search features, and (c) content.

Usability evaluation has become a very important issue for websites since usability strongly affects users. Shasha & Weideman undertook a usability study of Cape Town hotel reservation systems [16]. Employing usability testing as a research instrument to evaluate the system, they found that more than 52% of participants indicated key factors were content simplicity and an easy-to-understand booking system. 18% felt the websites were confusing, while 12% found the booking process frustrating. Their results provide a clear understanding how website usability affects user satisfaction.

A number of international standards on usability are available, but rarely used for useful usability evaluation. Hussain et al. used standard ISO 9241-11 to evaluate a web-based health awareness portal within the smartphone mobile context [17]. Their results uncovered some usability issues as well as confirming that the web-based awareness portal is relatively usable on smartphone devices within components defined in the models.

Gustafsson discussed how to design a UI for booking sunbeds to be both efficient and effective, while meeting the needs of first-time users [18]. Two personas were created to facilitate design of initial paper prototypes using user estimates. Then they iterated the design of the paper prototype and created a high-fidelity prototype using Adobe XD, which was evaluated using a task-based usability test. The prototype turned out to be very effective with additional qualitative data from participants helping to create an interesting experience and a system easy to learn and use.

Bangor et al. evaluated the system usability scale (SUS) from information collected on various systems during the various stages of the development life cycle over a 10-year period [19]. The SUS reflects a strong demand in the usability community for tools that can swiftly and efficiently collect users' subjective evaluations of usability. The information gathered in their study demonstrates that SUS can meet this need. Their analysis of SUS scores showed that for usability experts SUS is an exceptionally powerful and multifaceted tool.

Sauro conducted a five-year three-part study analyzing the experience of 4,000 users on more than 100 websites, thereby generating an eight-item website standardized UX percentile questionnaire featuring four factors: usability, trust, appearance and loyalty [20]. He concludes that questionnaires create reliable scores in benchmark websites, while standardized scores help designers to understand how the website scores compare with other scores in the database.

Feedback from developers, managers, and users is needed to optimize a system. In addition to basic software qualities, usability and UX are important attributes to improve. Usability is well known and can be tested, for example, by usability testing or expert reviews. On the other hand, UX describes the overall impact of the system on the end user before, during and after use.

Rauschenberger et al. introduced a tool that can easily evaluate the UX [21]. The tool is existing in multiple languages. They showed how to use the tool for continuous UX evaluation. Their work involved a validation study analyzing the Spanish version of the UX questionnaire, which they show can quickly evaluate the UX of interactive systems. Their tool measures usability aspects (such as efficiency, perspicuity, and dependability) and UX aspects. As the User Experience Questionnaire (UEQ) has a semantically different form, it is important that participants can use their natural language to rate products.

According to Kaur et al., the most important aspect of regulating the value of a website is usability [22]. Website designers need to understand the usability level of their website. Measurement techniques can be used to improve the usability level of a website. To determine usability level, they used two automated tools: a site analyzer which calculates multiple parameters and the Qualidator tool to check usability, accessibility, search engine optimization and technical quality. They evaluated educational universities in Punjab and provided rankings based on some evaluation criteria.

Based on the literature review the authors conclude the following:

1) There are many advantages and challenges to using PBS [7- 12].

2) There are different dimensions through which the usability of websites can be evaluated. The SUS is the most reliable option, even with a small number of participants, because of its reliability and low cost [15-22].

III. THE HSD TOOL

The HSD tool is a dynamic website based on physicians' persona. This tool provides various objects to designers of the health website. Each user has a dashboard that contains a set of different templates, or one can start a project from scratch. The tool provides the technologies and instruments used in the project in workspace panels such as page layout, adding an object (components, layouts, media, typography, buttons, and forms) and adding new pages. The user also has various options such as: change the setting, preview site and save code. The tool also contains code editors for maximum control. The HSD website implementation is based on the bootstrap

framework. The programming languages used in the HSD website are HTML, CSS, SCSS, JavaScript, jQuery.

The benefits of this design tool include making it possible to create an early health site UI prototype which addresses user requirements and avoids many of the usual difficulties. Based on this tool, the authors designed a UI for a PBS prototype and evaluated its usability creating a system suited to the personae of healthcare providers that is effective, efficient, and user-friendly while also reducing stress levels and improving time management.

IV. Procedure Booking Software (PBS) Materials and Methods

In this section, the procedure for designing the PBS is described as follows:

- Define the design requirements: follow the manual PBS used by the pediatric cardiologist at (KAUH), Jeddah, Saudi Arabia. Define persona that characterize physician behavior and the scenarios used in real life. [5].
- Design and develop an HSD tool, and validate this tool by measuring usability and UX.
- Based on the validation, the PBS was designed and evaluated (see section V).

After designing the HSD tool and prior to using it for designing the PBS UI empirical research was conducted to gather information from six physicians, three specialists and three fellows, in the pediatric cardiology department at (KAUH). Specialists use the manual PBS to book patients and work under the fellows, the experts in the field, who perform the surgeries. A questionnaire was designed to collect important information from two areas: general information and specific questions related to the participant's experience using technology devices. For example: Do they think that using electronic healthcare booking is easier than paper records? Do they prefer to manage the booking procedure electronically? Do they feel comfortable using an electronic healthcare booking system? Do they require knowledge about how to use the electronic healthcare booking system? This step helped to reveal any assumptions about user knowledge and experience and created a starting point for decomposing any stereotypes. All the response results were collected and expressed as a percentage as follows:

- 71.4% of the responses were from the age group 30 to 39 years.
- 86% of the responses have a medium level of experience in using technology devices.
- All the responses their specialty was pediatric cardiology, 50% specialist and 50% fellow.
- 100% of the responses prefer to manage the booking procedure electronically.
- 57% of the responses used a paper-based booking procedure method and 43% a computer-based method.
- 86% of the responses believe using electronic healthcare booking is easier than paper records.
- 86% of the responses have the required knowledge to use the electronic healthcare booking system.
- 86% of the responses feel comfortable using the electronic healthcare booking system.

Based on the above results, the researchers decided to implement the PBS based HSD tool designed to suit the physicians' personalities and meet their requirements.

Establishing a PBS means designing a system that mainly deals with the process of booking operations. Operation room reservations are linked to the system so a physician can make reservations from his private clinic through the system. The method revolves around the physician's ability to enter the electronic system from any computer, mobile phone or tablet to enter the patient's data. Then the physician chooses the day of the operation, expected duration and type, either as a one-day surgery or hypnotic session. After this, the physician schedules the process electronically and approves it. The system is characterized by several features as follows:

- Friendly: It easily and accurately deals with the information.
- Arrangement: If the operation is canceled by the physician, the canceled day will be automatically carried over to the next available operation day.
- Notification: The patient is notified of the new appointment by letter.
- Reminder: Reminders are sent to the physician and the patient about the date of the operation, with the attendance confirmed by the patient, or the reservation is canceled, making it available to another person.
- Availability: The physician is able to enter the system and see all the private information, the date of the surgery, and all the observations of the operation.
- In addition, there is a special schedule for anesthesiologists, their names, hours of operation, and numbers so they are notified by email.

A. Tools and Technologies

Libraries and programming languages used for the PBS website are the same as for the HSD tool. The framework consists of HTML, CSS, SCSS, JavaScript, jQuery, and the bootstrap framework [23].

B. Architecture and Implementation

Fig. 1 illustrates the architecture applied to PBS website.

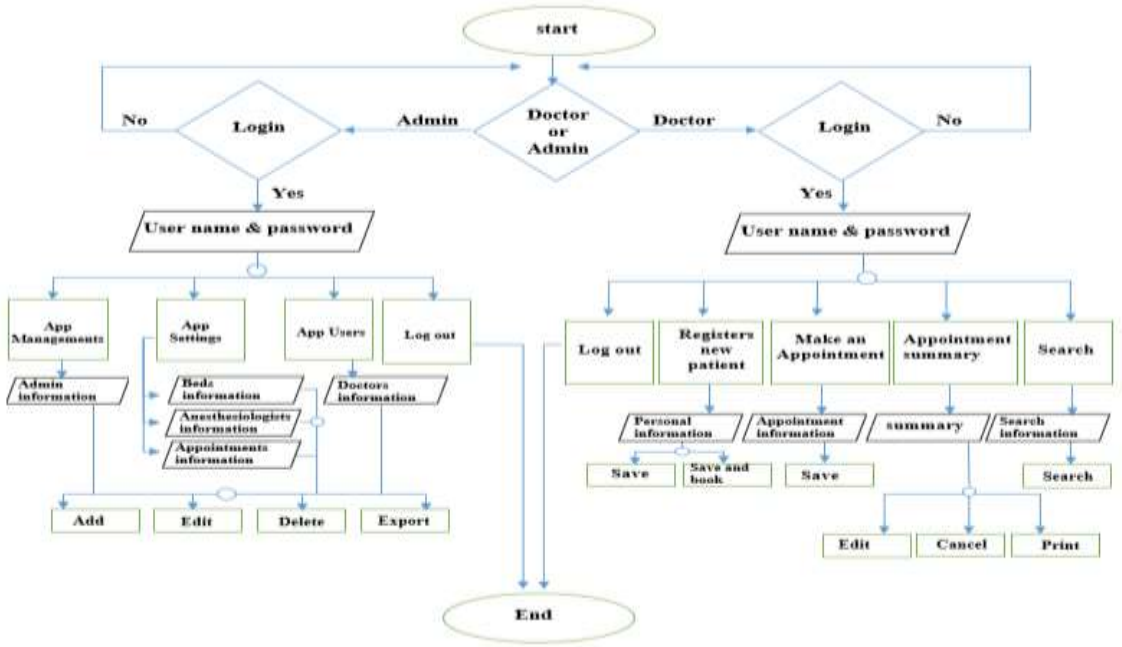


Fig. 1. Architecture of the Procedure Booking Software (PBS) Website.

C. Prototype

The majority of studies involving prototypes discuss the fidelity aspect of prototypes [24]. The concept of fidelity denotes the similarity between a prototype and the final product. Based on the degree of refinement, prototypes are categorized as low-fidelity or high-fidelity (hi-fi) [25]. The final hi-fi prototype was designed based on the outcomes from the low-fidelity early version, and the personae. The set of features provided by PBS allows users to register a new patient, book an appointment, display a summary of the appointment, and search for patient information throughout a website [26]. Two sections are harmonious all around the UI: menu section on the left bar, and the action section (dashboard) on the right bar. The site is divided into two basic pages. The user page presents materials for booking procedures. Here the user can register new patients and book appointments. The admin page is used to modify and control input and relevant processes and communicate and exchange with the user page (Fig. 2). During the design process, the website went through various revisions considering different parameters before reaching the final version.

Fig. 2. The Admin main Page.

V. Experimental Design and Evaluation for the PBS

The experimental evaluation presented in this section involves the usability testing of PBS. Before conducting the testing, the following key issues require consideration for the information-gathering to be fruitful [27]:

A. Defining the Goal

The goal of this study is to design a PBS based on the HSD tool to improve effectiveness, efficiency, and physician satisfaction —as usability definition in ISO 9241-11 [28].

B. Pilot Test

A pilot study is a preliminary run of the principle study. The point is to ensure that the proposed strategy is practical before setting out on the real examination. The equipment and instructions that should be used were prepared.

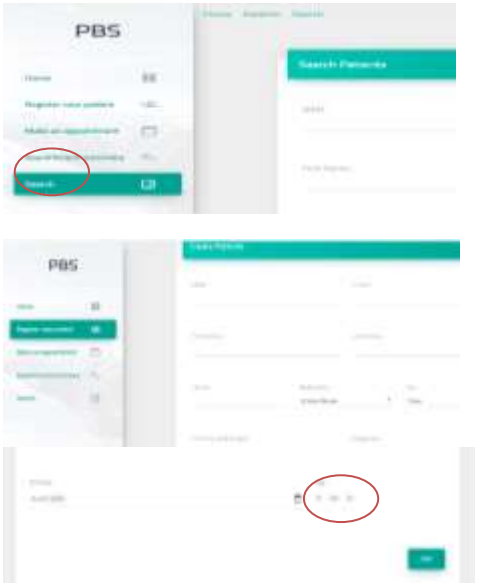
The proposed PBS was tested first with one expert designer and one expert physician. Both were asked to complete the list of tasks in Table II to ensure the design is clear and works well. The researchers received feedback before conducting the test. Several adjustments were made before conducting the pilot test and after, with the clarification of the actions. Table I contains sample of Pilot test 1 with several changes: (added the necessary items, changed some terms name, added searching). Pilot test 2 with several changes: (add content to the home page, change the format of the entered lists, resize all buttons). Pilot test 3 with several changes: (changed some terms, make the date a full (day, month and year), change some terms, add an advanced search).

C. Defining the Participants

Participant samples were comprised of six physicians selected at random for the study belonging to the department of paediatric cardiology at KAUH. Three participants were

fellows (Novice) and three specialists (Expert) in using the existing booking method. All the participants were asked to perform the same list of tasks individually, and then answer the post-test questionnaire. The attributes collected: the time spent to complete each task, the number of clicks, and the number of errors for each task. All the attributes were tracked using video recordings. After completing the tasks, the participants answered SUS questionnaire [29], and were requested to offer inputs related to their experience on a short note.

TABLE I. THE PILOT TEST

Before Pilot test	After Pilot test	Action
		Added the necessary item (Search), (MRN - First Name - Last Name) and added some necessary terms (Primary Cardiologist – Diagnosis (Consultant performing the procedure -Procedure - Admission Site) Made the date a full day, month and year instead of just a year Changed some terms such as book an appointment instead of making an appointment, anaesthesiologist to Anaesthesia
		Change the page shape to one like iPad where all icons and pages appear on the same screen Added content to the Home page Resized all buttons to a smaller size Changed the format of the entered lists and put them in a smaller size
		Added MRN search and display patient information. Add an advanced search

To determine the effectiveness of the proposed tool, the two groups were defined as follows:

- Specialists group (Expert): The least experienced in the field but the most familiar with electronic systems.
- Fellow group (Novice): The most experienced in the field but the least familiar with electronic systems.

The test was conducted in the summer 2020. Due to the outbreak of COVID 19, the evaluation was conducted in online environment. The researchers contacted the participants first by sending the consent form to obtain their approval and explain the reason for conducting the study. All the participants received the same set of instructions. For example: 1) Participants were asked to open the screen recorder to record each session before starting to perform the test. 2) Participants were asked to fill out the SUS. Then, asked to perform a list of tasks (Table II) on the PBS link on the web-based system and download the recording screen to record their performance, then return it to the researchers by email.

The participants were asked to perform the tasks without explanation of the system. All the participants received four tasks to complete reflecting different functionality as follows (Table II):

TABLE II. FOUR TASKS TO EXAMINING THE PBS SYSTEM

	Task	***Description***
1	Register a new patient.	Input MRN (any 7 number), E-Mail (any email), First Name (any name), Family name (any name), Tel No (any mobile number), Nationality (Saudi Arabia), Sex (male or female), Primary Cardiologist (Khadijah Maghrabi), Diagnosis (Double outlet right ventricle), Birthday (any date) and click SAVE.
2	Book an appointment	Choose MRN that you registered then input Bed (1 or 2), Anesthesia (YES | NO), Consultant performing the procedure (Jameel Al-Ata), Procedure (Closure of PDA), Admission Site (Pediatric ward), Date (any date), Time (any time) and click SAVE.
3	Display a summary of the appointment.	Display all appointments.
4	Search about the patient	Search about the patient that you enter by his MRN or by his name.

D. Triangulation Role

The triangulation of data is drawn from a different sampling technique (observation recorded video and questionnaire) that was used to collect and interpret the data. The authors chose video recording because video has the advantage of capturing both visual and audio data. Also, results from a video session can be taken away and analyzed. Participants had to use their own desktop computers supported with Google Chrome to test the PBS.

VI. RESULTS

This section first provides the results of the usability study in terms of objective characteristics. Then, the subjective characteristics in terms of SUS results and finally, the UX attribute measures.

A. The Usability Study Result

The usability testing was conducted to measure the effectiveness, efficiency, and cognitive load.

To make the test completely scientific, the authors set two hypotheses and performed a T-test to determine whether there was a significant difference between the two groups [30].

- Effectiveness:

The effectiveness was measured using the total min correct clicks divided by the sum of correct clicks and incorrect clicks per task.

$$\textbf{Effectiveness} = \frac{\text{Min correct clicks}}{\text{correct clicks} + \text{incorrect clicks}}$$

The assumption was relied on the data to support or reject the hypotheses in effectiveness characteristics based on the following hypotheses:

- Hypothesis 01: There is no significant difference between the effectiveness of the specialists group and fellow group.
- Hypothesis 1: There is a significant difference between the effectiveness of the specialists group and the fellow group.

Based on the average result for all tasks, the P-value was (0.16) greater than the value of alpha (0.05), which means that the hypothesis 01 is not refuted. Fig. 3 shows that the fellow group is better at completing all tasks than the specialist group, except task4.

- Efficiency:

The efficiency was measured using the effectiveness divided by the total time spent per task.

$$\textbf{Efficiency} = \frac{\text{Effectiveness}}{\text{Time}}$$

The assumption was relied on the data to support or reject the hypotheses in efficiency characteristics measured based on the following hypotheses:

- Hypothesis 02: “There is no significant difference between the efficiency of the specialists group and fellow group.”
- Hypothesis 2: There is a significant difference between the efficiency of the specialists group and the fellow group.

Based on the average result for all tasks, the P-value (0.10) is greater than the value of alpha (0.05). The result conclude that there is no significant difference in efficiency values between the specialists group and the fellow group. Fig. 4 shows that the fellow group had higher efficiency compared to the specialist group in completing the tasks. This indicates the fellow group performed better than the specialist group.

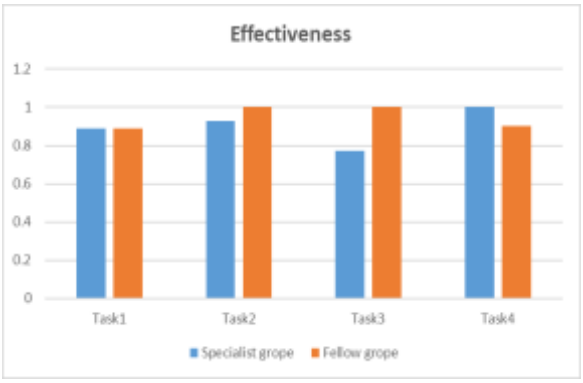


Fig. 3. Calculations for Effectiveness.

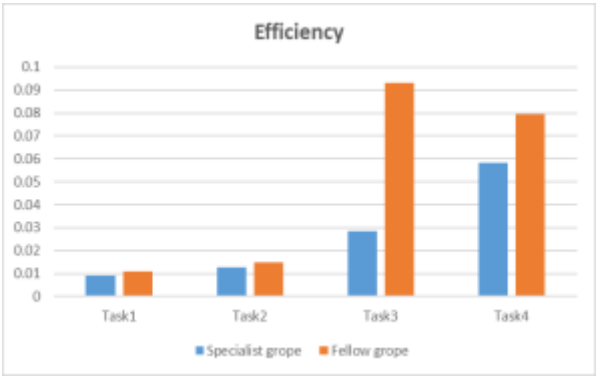


Fig. 4. Calculations for Efficiency.

- Cognitive load:

To accomplish a 'task', a user may have to switch from one screen to another. This will depend on the complexity of tasks, screen size, or the way the designer has filled information on various screens [31]. It was measured by:

$$\textbf{Cognitive load} = \frac{\text{Pages' view}}{\text{correct clicks} + \text{incorrect clicks}}$$

The assumption was relied on the data to support or reject the claims in cognitive load characteristics measured based on the following hypotheses:

- Hypothesis 03: "There is no significant difference between the efficiency of the specialists group and fellow group."
- Hypothesis 3: There is a significant difference between the efficiency of the specialists group and the fellow group.

Based on the average result for all the tasks, the P-value obtained (0.21) is greater than the value of alpha (0.05), which means that cognitive load hypothesis 03 is not rejected.

Fig. 5 shows that the fellow group had equal or higher cognitive load than the specialist group in completing the tasks except for task 4 which means the specialist group was better in cognitive load than the fellow group. This result explains the specialist's familiarity with using the technology since their cognitive load was better than the fellow in most of the tasks except Task 4.

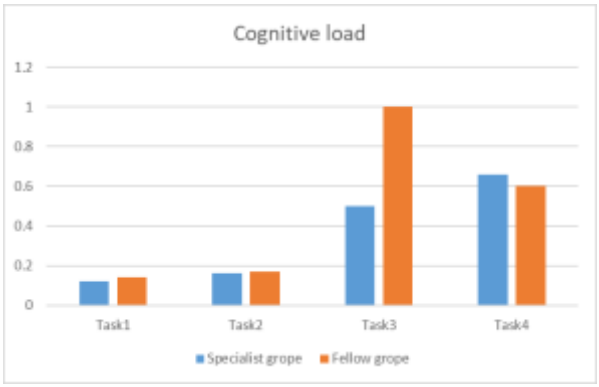


Fig. 5. Calculations for Cognitive Load.

In terms of effectiveness and efficiency, all participants in the two groups were able to successfully complete each task, which means that they understand the task to be performed using the system and have the ability to perform that task.

Overall, there is no significant difference between the two groups for either of the tasks. However, there is a slight difference in the number of clicks and the time spent between the participants, as the fellows were quicker than the specialists in terms of time and required less clicks. See Fig. 6 (The average number of clicks required per task), and Fig. 7 (The average completion time per task).

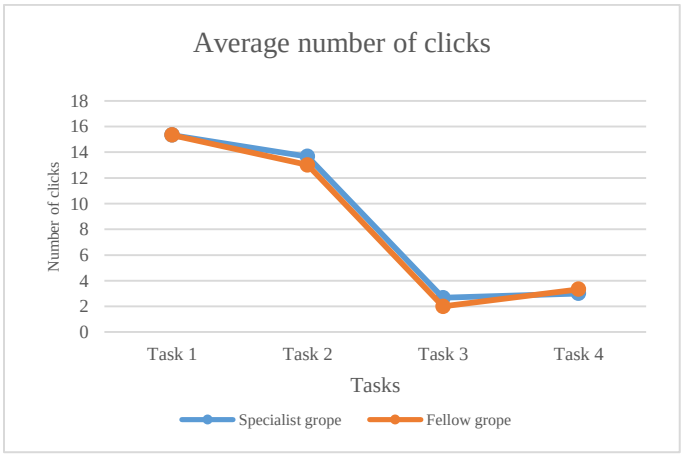


Fig. 6. Total Number of Clicks.

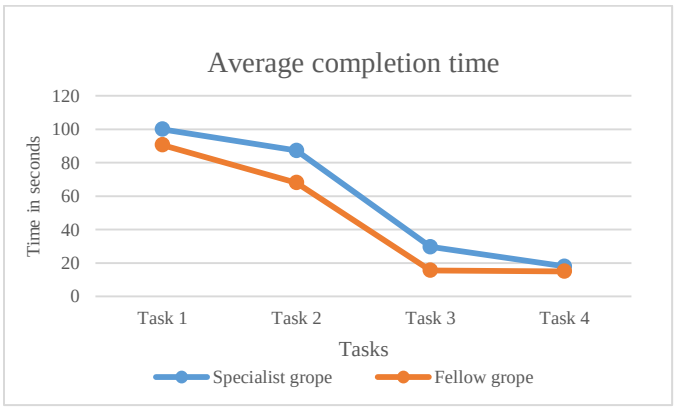


Fig. 7. Total Time to Complete the Tasks.

B. System Usability Scale Result

The questionnaire was followed up with a short Google survey. For all participants in both groups the lowest SUS score landed on 65 while the highest was 100. The average SUS score was 86 setting it in the A Grade.

The SUS questionnaire includes the following criteria:

1) I prefer to use the PBS regularly.
2) I found the PBS unnecessarily sophisticated.
3) I found the PBS was easy to use.
4) I would need the help of a technical person to use the PBS.
5) I found the different functions in the PBS integrated well.
6) I found there was a lot of discrepancy in the PBS.
7) I can claim that most people would learn to use the PBS very quickly.
8) I found the PBS very burdensome to use.
9) I felt very confident using the PBS.
*10)*I need to learn many items before working with the PBS.

Comments mentioned by most of the physicians are as follow:

- *"User-friendly"*
- *"Easy to learn and organized system"*
- *"This system is good for its ability to save booking data from being lost, and it is good for easy search /recall saving a lot of time needed to look for a booked patient. Thank you for developing this system to help us"*
- *"Its efficient, user friendly and saves time"*
- *"So easy and comfortable"*
- *"Maybe we need additional digits like additional notes"*

C. User Experience Attributes Result

The PBS was evaluated by measuring the UX attributes in terms of time -success - errors [32]. The results are as follows:

- Time to complete the tasks:

Fig. 8 shows that all participants in both groups completed all tasks in the best time. The fellow group was faster than the specialists group in all tasks.

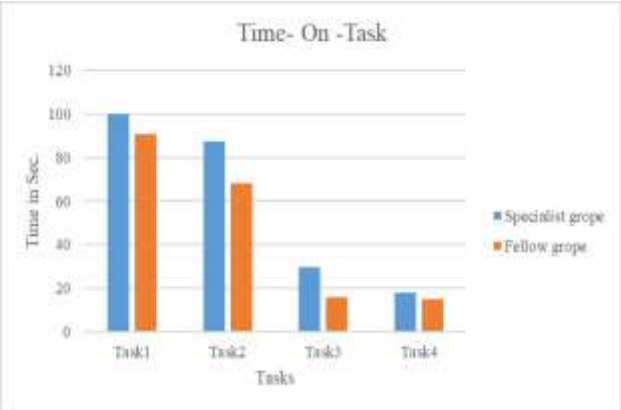


Fig. 8. Total Time to Complete the Tasks.

- Successful tasks percentage:

Fig. 9 shows that all participants in both groups had a success rate of 100%.

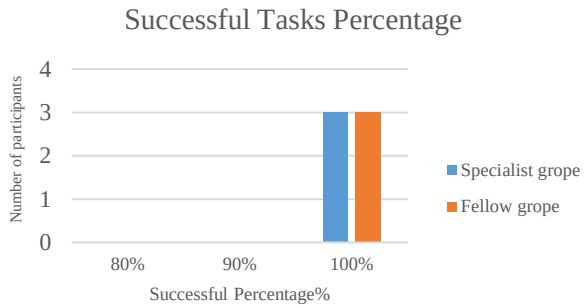


Fig. 9. Total Successful Tasks Percentage.

- Number of errors percentage:

Fig. 10 shows that participants in both groups had less than 10% error in completing all tasks. Just one specialist had errors close 20% in task 1.

From the results on the usability study, SUS and UX attributes, the authors can confirm that the proposed PBS is user friendly, easy to use, easy to learn and satisfy the user with minimum effort and time. This is supported by the observation that most of the results of the fellow (novice user) is better or close to the specialist (expert users).

Fig. 10. Total Number of Errors.

VII. Conclusion and Future Work

This research aimed to design and develop a suitable PBS UI as a case study by using the HSD tool based on physicians' persona to prove the study design concept in the pediatric cardiology department KAUH. It was then measured the objective characteristics (effectiveness, efficiency, and cognitive load) and subjective features (satisfaction) by applying the SUS measurement. The results confirm that: there is no significant difference between specialists and fellow groups, which supports the study claim. Specifically, the fellow group achieved similar results to the specialists' group indicating the PBS was perceived as efficient, effective, and satisfactory when tested on six participants in a task-based usability test that obtained a SUS score of 86.

UX attributes and user feedback confirmed that the test created a good experience that helped them perceive the system as efficient, user-friendly, and time saving.

In future work, the authors are looking to enhance the tool by generating the UI automatically based on the physicians' personality and expanding the persona, for example, adding a persona for nurses. Also, make it adapted for other health systems and testing with a larger sample.

ACKNOWLEDGMENT

The authors gratefully acknowledge the pediatric cardiology staff in the department of Cardiac Catheterization at KAUH.

REFERENCES

[1] Ellingsen G, Obstfelder A. Collective expectations—Individual action implementing electronic booking systems in Norwegian health care. International Journal of Medical Informatics. 2007 Jun 1;76:S104-12.

[2] Zhao P, Yoo I, Lavoie J, Lavoie BJ, Simoes E. Web-based medical appointment systems: a systematic review. Journal of medical Internet research. 2017;19(4):e134. R. Keefer, “Reducing Stress in Healthcare : Evidence from Using an Integration Design Model,” pp. 705–709, 2017.

[3] M. Y. Ivory and M. A. Hearst, “Improving Web Site Design,” no. April, 2002.C.

[4] International Organisation for Standardisation ISO 9241–11:1998 Guidance on usability.

[5] Alzahrani, Hanaa, and Reem Alnanih. "The Effect of User Experience on the Quality of User Interface Design in Healthcare." International Conference on Computing. Springer, Cham, 2019.

[6] B. Fogg, “The Behavior Grid,” Proc. 4th Int. Conf. Persuas. Technol. - Persuas. ’09, p. 1, 2009.

[7] Zhao, Peng, et al. "Web-based medical appointment systems: a systematic review." Journal of medical Internet research 19.4 (2017): e134.

[8] Gupta D, Denton B. Appointment scheduling in health care: Challenges and opportunities. IIE transactions. 2008 Jul 21;40(9):800-19.

[9] Yang PC, Chu FY, Liu HY, Shih MJ, Chen TJ, Chou LF, Hwang SJ. Features of online hospital appointment systems in Taiwan: a nationwide survey. International Journal of Environmental Research and Public Health. 2019 Jan;16(2):171.

[10] E. Gamma and R. Helm and R. Johnson and J. Vlissides, Design Patterns - Elements of Reusable Object-Oriented Software, Addison-Wesley, 2002. 24th edition.

[11] D. Sinnig, A. Gaffar, D. Reichart, P. Forbrig and A. Seffah, Patterns in Model-Based Engineering, Proc. CADUI 2004.

[12] Murray, Christopher JL, and Julio Frenk. "A framework for assessing the performance of health systems." Bulletin of the world Health Organization 78 (2000): 717-731.

[13] Iso IS. 9241-11 (2018) Ergonomics of human-system interaction—part 11: usability: definitions and concepts. International Organization for Standardization. https://www. iso. org/obp/ui/# iso: std: iso. 2018;9241(11).

[14] DeLone, William H., and Ephraim R. McLean. "Information systems success measurement." Foundations and Trends® in Information Systems 2.1 (2016): 1-116.

[15] Rinder, Julie. "The Importance of Website Usability Testing." (2012).

[16] Shasha,ZT and Weideman M. Usability measurement of web-based hotel reservation systems. Conference Proceedings of the 1st International TESA Conference, September 2016, Cape Town, South Africa.

[17] Hussain, Azham, Emmanuel OC Mkpojiogu, and Zakaria Hussain. "Usability evaluation of a web-based health awareness portal on Smartphone devices using ISO 9241-11 model." Jurnal Teknologi 77.4 (2015): 1-5.

[18] Gustafsson, Filip. "An Explorative Design Study of a Booking System: Evaluating the Usability and Experience of a User Interface for Novice Admin Users." (2019).

[19] Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Intl. Journal of Human–Computer Interaction. 2008 Jul 29;24(6):574-94.

[20] Sauro J. SUPR-Q: A comprehensive measure of the quality of the website user experience. Journal of usability studies. 2015 Feb 1;10(2).

[21] Rauschenberger, Maria, et al. "Efficient measurement of the user experience of interactive products. How to use the user experience questionnaire (UEQ). Example: Spanish language version." (2013).

[22] Kaur S, Kaur K, Kaur P. Analysis of website usability evaluation methods. In2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) 2016 Mar 16 (pp. 1043-1046). IEEE.

[23] Spurlock, Jake. Bootstrap: Responsive Web Development. O'Reilly Media, Inc.", 2013.

[24] Nissinen, Tuomas. "User experience prototyping–a literature review." University of Oulu, Oulu (2015).

[25] Coyette, Adrien, Suzanne Kieffer, and Jean Vanderdonckt. "Multi-fidelity prototyping of user interfaces." IFIP Conference on Human-Computer Interaction. Springer, Berlin, Heidelberg, 2007.

[26] Alzahrani, H., 2020. *PBS Procedure Booking Software*. [online] Booking-ps.com. Available at: <http://booking-ps.com/> [Accessed 13 October 2020].

[27] Sharp H and Preece J,, Interaction Design: Beyond Human-Computer Interaction, Wiley, p. 841, 2015.

[28] “ISO - ISO 9241-11:2018 - Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts.” [Online]. Available: https://www.iso.org/standard/63500.html. [Accessed: 11-Apr-2020].

[29] Peres, S. ., Peres, C., Pham, T., & Phillips, R. (2013). Validation of the System Usability Scale (SUS): SUS in the wild. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (pp. 192–196). Santa Monica, CA: HFES.

[30] Talib MA, Alnanih R, Khelifi A. Application of quality in use model to assess the user experience of open source digital forensics tools. International Journal of Electronic Security and Digital Forensics. 2020;12(1):43-76.

[31] Reem Alnanih and Olga Ormandjieva and Radhakrishnan, T. (2013 C). A New Quality-in-Use Model for Mobile User Interfaces. Proceedings of the 23nd International Workshop on Software Measurement, IWSM-MENSURA.

[32] Albert W, Tullis T. Measuring the user experience: collecting, analyzing, and presenting usability metrics. Newnes; 2013 May 23.

Validation Analysis of Scalable Vector Graphics (SVG) File Upload using Magic Number and Document Object Model (DOM)

Fahmi Anwar[1]
Department of Informatics
Universitas Ahmad Dahlan
Yogyakarta, Indonesia

Abdul Fadlil[2]
Department of Electrical Engineering
Universitas Ahmad Dahlan
Yogyakarta, Indonesia

Imam Riadi[3]
Department of Information Systems
Universitas Ahmad Dahlan
Yogyakarta, Indonesia

***Abstract*—The use of technology is increasing rapidly, such as applications or services connected to the Internet. Security is considered necessary because of the growing and increasing use of digital systems. With the number of threats to attacks on digital form or server systems is required to handle the risk of attacks on the server, the file upload feature. The system usually processes the file upload feature on a website or server with server-side (back-end) validation or filtering of digital object file types or a client-side (front-end) web browser in HTML or Javascript. Filtering techniques for Scalable Vector Graphics (SVG) usually files only see the file extension or Multipurpose Internet Mail Extension (MIME) type of an uploaded file. However, this filtering can still manipulate, for example, in ASCII prefix checking, which has two writes, namely "<?xml" and "<svg ". SVG files do not contain metadata such as image encoded in JPEG or PNG files. This problem can overcome by adding filtering techniques to check the validation of a file with validation of eXtensible Markup Language (XML) using magic numbers and the Document Object Model (DOM). This research developed using the waterfall method and black-box security testing refers to a software security testing method in which security controls, defense, and application design are tested. Handling of security validation for uploading SVG files using file extensions and MIME types has a success rate of 75 percent from the eight tested scenarios while handling using file extensions, magic numbers, and Document Object Model (DOM) produces a success rate of 100 percent from 8 test scenarios. Testing uses a black-box so that handling using the file extension, magic number, and Document Object Model (DOM) is better than using only file extensions and mime types.**

***Keywords*—Magic number; Scalable Vector Graphics (SVG); security; upload; validation**

I. INTRODUCTION

The website is an Internet service that can be used by various users in the world, which usually has an upload feature. The file upload feature or file upload is a feature that is generally functionally needed in applications for users [1]. However, without proper filtering, file selection, and validation processes during upload can present a significant security risk for website security [2] with three critical characteristics: integrity, input validation, and correct logic required for security applications DDoS attacks.

Distributed Denial of Service (DDoS) is a network security problem that continues to develop dynamically and increases significantly until now. DDoS is a type of attack performed by draining the network resources by flooding packets with significant intensity until they become overloaded and servers stop functioning. DDoS assault characterization depends on network traffic movement utilizing the Neural Networks and Naïve Bayes Methods. Because of the trials led, it discovered that the aftereffects of exactness in counterfeit Neural Networks were 95.23%, and Naïve Bayes Methods was 99.9%. The trial results show that the Naïve Bayes Methods are superior to Neural Networks. The examination and investigation consequences as proof in the primary cycle [3]. Another research is Artificial Neural Network (ANN) can be used as a viable device for network parcel arrangement with the proper blend of learning, move, concealed layer, and preparing capacities. ANN with two concealed layers gives generally predictable MSE, combination speed, higher right grouping rate at 99.04%, and a Quasi-Newton preparing capacity strategy (Matlab-trainlm) suited for the arrangement task, given the estimation of relapse both in the preparation and approval stage [4]. Another technique in detecting these attacks is monitoring but found several problems [5], including difficulty distinguishing the attack and regular data traffic using Density K-Means Method.

Cyber attacks by sending large data packets that deplete computer network service resources using multiple computers when attacking are called DDoS attacks. Total data packet and essential information in the form of log files sent by the attacker can be observing and captured through the port mirroring of the computer network service. The classification system must distinguish network traffic into two conditions, the first normal condition, and the second attack condition. The Gaussian Naive Bayes classification is a method that can use to process numeric attributes as input and determine two decisions of access that occur on the computer network service [6]: "normal" access or access under "attack" by DDoS as output using Numeric Attribute-based Gaussian Naive Bayes.

Another research about forensic analysis and prevent Cross-Site Scripting (XSS) using the Open Web Application Project (OWASP) Framework covers three essential stages: Attacking stages, Analysis, and Patching. Stages Attacking is doing exercises with Single-Victim-strategy utilizing the

OWASP Xenotix XSS Attack Exploit Framework v6.2 to incorporate assaults Information Gathering, Keylogger, Download spoofer, and Live webcams screen capture to the casualty through the internet browser [7]. Stages Analysis led utilizing Live Forensic by Wireshark, live HTTP Header, and Tcpdump.

Other studies provide reviews of techniques, stages, approaches, and tools to detect web servers is vulnerabilities [8]. Challenges and problems during application security testing prove to provide testers and managers input about application projects [9]. This study also highlighted various authors based on the Open Web Application Security Project (OWASP) Top 10 [10]. As per a report by the Web Application Security Consortium, about 49% of the web applications investigated contain weaknesses of high-hazard levels, and beyond what 13% of the sites can be undermined altogether consequently.

According to the 2018 White Hat report, analysis results using data from more than 20,000 applications indicate a decline in security [11]. More than 75% of malicious attacks are mostly cross-site scripting (XSS) attacks [12]. XSS attacks pose serious threats, especially servers in the financial and economic fields. Such attacks pay considerable attention to developing methods of protection against XSS attacks and proactively detecting vulnerabilities. These aspects determine the emergence of scientific research, especially in studying XSS vulnerabilities in graphic content files.

Several vulnerabilities in the exploitable file upload feature [13] include:

- No validation is performed on the client-side or server-side.
- Client-side validation skippable using developer options in the web browser and not using validation checking file contents.
- No validation is performed to check file size, as validation is based only on content type.
- Attacks can be carried out by manipulating file content types.
- It is allowed to use more than one type of file extension.
- Some conditions may use forbidden file extensions along with file extensions not permitted by the application.

Scalable Vector Graphics (SVG) is a language based on the eXtensible Markup Language (XML) for describing two-dimensional vectors and mixed vector/raster graphics. Stylized SVG content can also be scaled to different display resolutions and can be viewed alone or mixed with HTML content or embedded using XML namespaces in other XML languages. SVG also supports dynamic changes with form scripts. The script can create interactive documents and animation using the declarative animation feature or a script [14].

The Document Object Model (DOM) is a programming Application Programming Interface (API) for accessing and modifying XML documents. The DOM defines the document as a logical structure and the various ways of accessing or manipulating the document. XML is also used to represent different formats of information stored in a heterogeneous system and is mostly interpreted as data rather than documents [15].

Previous research that analyzed the main problems related to web applications and Internet services in several web applications from various organizations, such as banking, health care, financial services, retail, developed a systematic grouping of XSS protection techniques [16] such as rules for protecting website graphic content. Web and prevent XSS vulnerabilities [17] and An Analysis of Vulnerability Web Against Attack Unrestricted Image File Upload [18].

Numerous elements make it trying to make sure about applications that have been mulled over to improve application security. Unreliable applications work because of the weaknesses of multiple components. For example, security testing is done past the point of no return in the SDLC, avoiding security testing in light of the delivery surge, spending limitations. All the more usually, the absence of security mindfulness by designers. The lack of designer consciousness of secure coding norms and the absence of spending plans spent on application security are two of the most alarming issues. This present examination's essential objective is for designers and analyzers to comprehend the fundamental weaknesses of record transfer usefulness, prompting assaults, and their particular alleviations for future secure turns of events. This study conducted a series of tests and performed a graphical content vulnerability analysis against XSS attacks. In contrast to image encoded such as JPEG and PNG, which have metadata and can be processed using image processing [19], it can also be classified based on color and pattern values [20]. This study utilizes different magic numbers and DOM to validate SVG files in the file upload feature in the appropriate XML format scriptwriting structure.

II. Research Methodology

File upload is transferring the files (photos, audio files, etc.) to a server on the website. To upload data to the server, the client first initiates communication with the server by initiating a TCP/IP connection from the client to the server called a handshake. In this communication, the client initiates any communication and not the server. When a connection is established between a client and a server, data transfer can occur between them. It does not require port forwarding to send/receive data to/from the server. Now the client needs a file to upload and a form on a Web page where the file is sent to the server. This allows the user to enter one or more files into the form submission like the code shown in Fig. 1.

After the HTML tag from Fig. 1 sends through the server data, it is often processed to save the file to the webserver disk. The server-side script executing the file is received on the server. The server knows how to handle such requests and store data. It saves files to server disks with multiple names and process data by simply extracting some information from them [13]. This study uses the Waterfall model development method. Waterfall model development methods usually suggest a systematic and sequential approach to software

development that starts with customer specifications and progresses through planning, modeling, construction, and completed software deployment [21].

Fig. 2 is the stage of the Waterfall Model which reflects the main points of development activities such as:

- The communication contains system services, system limitations, and objectives set after consultation with system users, defined in detail, and used as system specifications.
- Planning contains creating a system to identify and explain the system abstraction and its relationship, estimated processing time, and scheduling.
- Modeling contains a system design that will be made in the form of a flowchart.
- The construction contains designing the system into a program or unit program and then unit testing, which involves verification to ensure whether each unit meets system specifications. Each program unit and existing programs are integrated and tested as system integrity to confirm whether the system requirements have been met. After testing, the new system is deployed to users.
- The deployment contains stages of installation of the system and is used in practice.

Black-box security testing refers to a software security testing method in which security controls, defense, and application design are tested from the outside, with little or no prior knowledge of how the application is internals work. Essentially, black-box testing takes a similar approach to a real attacker. Since black-box security testing does not assume or know the target being tested, it is a technology-independent testing method [22]. This makes it ideal for various situations, mainly when testing for vulnerabilities arising from deployment issues and server configuration errors.

```
<!DOCTYPE html>
<html>
<body>

<form action="upload.php" method="post" enctype="multipart/form-data">
 Select image to upload:
 <input type="file" name="fileToUpload" id="fileToUpload">
 <input type="submit" value="Upload Image" name="submit">
</form>

</body>
</html>
```

Fig. 1. Sample of File upload HTML Tags.

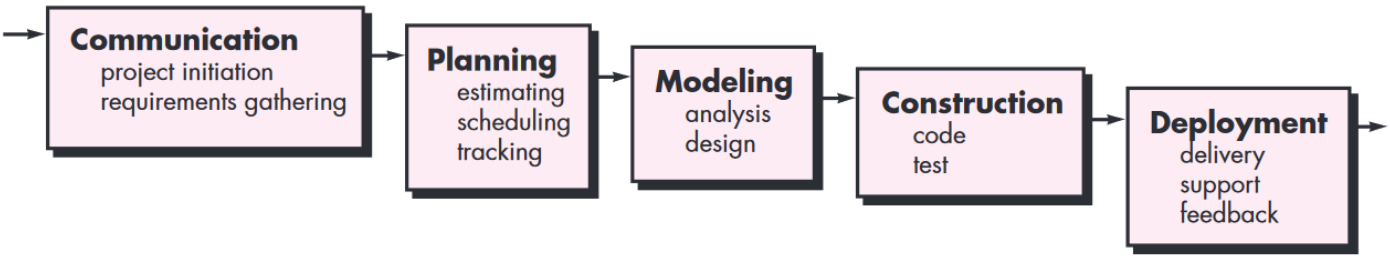


Fig. 2. Waterfall Method.

III. Result and Discussion

This study uses Waterfall as a method of developing the implementation of handling the SVG file upload validation using magic numbers and DOM and black-box security testing refers to a software security testing method with the following stages.

A. Communication

The communication stage contains the preparatory stages for making system services, system boundaries, and implementation objectives in detail, according to the system is specifications. SVG files contain tags in the form of XML with file extensions ".svg" and MIME "image/svg+xml" but in checking the ASCII prefix, there are two standards, namely "<?xml" and "<svg ". Bug #79045 in PHP about Incorrect SVG MIME types detected, PHP when rendered SVG may be without an XML header, will return the mime type “image/svg” instead of “image/svg+xml”. However, IANA doesn't list “image/svg” as an existing MIME type, nor does the browser display it [23].

According to SVG standards, the MIME-type returned must always be “image/svg+xml”, PHP must also behave consistently for a single file type. If this is used to distinguish a properly header-fed SVG from non-strict ones, it should be done as part of a file validity check, not a mime type check. A previously reported bug #76543 is indicated as "Not a bug" because it is considered an upstream error with libmagic. However, libmagic returns an svg that doesn't have an xml header as “text/plain”. This indicates that the error is in the adaptation PHP took to eliminate the error [24]. Expected result for MIME type of SVG is “image/svg+xml” but actual result is “image/svg”. SVG files do not contain metadata such as image encoded in JPEG or PNG files.

B. Planning

The planning stage contains steps that are carried out using a sample made as presented in Fig. 3, which includes the SVG code as in Fig. 4 with the prefix "<svg " while Fig. 5 contains the SVG code with the prefix "<?xml".

Fig. 3 displays the SVG code presented in Fig. 4 and Fig. 5, showing the same visual appearance with different magic number values.

Fig. 3. Sample SVG File.

1) SVG Code with SVG tag first: Fig. 4 contains the SVG code that uses the standard prefix "<?xml" in the form of ASCII hexadecimal value 3C 73 76 67 20, which is the SVG 1.1 standard (Second Edition), which became the W3C recommendation on 16 August 2011 [13]. However, many SVG code writings use the prefix "<svg" in ASCII form, which is hexadecimal 3C 3F 78 6D 6C in Fig. 5.

2) SVG Code with XML tag first: The two standards of the prefix "<?xml" and "<svg " are the standards that are often used and then sampled according to the file extension parameters, magic number, and XML format as in Table I.

```
<svg height="512" viewBox="0 0 128 128"
width="512"
xmlns="http://www.w3.org/2000/svg">
    <g>
        <path d="m7.157 61.039s-8.924-
30.295 31.234-27.477 36.453-15.94 66.258-
11.57c38.141 5.591 6.432 57.391-37.553
63.545-43.805 6.128-56.88-14.101-59.939-
24.498z" fill="#f2e7cb" />
        <path d="m64.006 50.121c-11.747-
6.451 2.471-8.9 2.471-8.9 18.195-2.442
20.9-11.164 32.278-10.175 20.09 1.747 9.23
22.95.881 26.146 0 0-7.734 2.76-19.6-
.637a87.6 87.6 0 0 1 -16.03-6.434z"
fill="#ef3829" />
        <path d="m33.546 43.2s-19.075-.978-
15.81 13.551 25.466 24.3 55.5 17.283-7.336-
11.407-16.483-15.978-4.44-13.196-23.207-
14.856z" fill="#ef3829" />
        <path d="m121.7 38.052c0 17.53-
24.409 43.261-54.6 47.485-43.805 6.128-
56.88-14.1-59.939-24.5a26.44 26.44 0 0 1 -
.844-6.82v19.859a26.436 26.436 0 0 0 .844
6.961c3.059 10.4 16.134 30.626 59.939 24.5
30.14-4.217 54.516-29.867 54.6-47.4v-
20.086z" fill="#ef983b" />
        <g fill="#422002">
```

Fig. 4. Sample SVG Script with SVG Tag First.

```
<?xml version="1.0" standalone="no"?>
<svg height="512" viewBox="0 0 128 128"
width="512"
xmlns="http://www.w3.org/2000/svg">
    <g>
        <path d="m7.157 61.039s-8.924-
30.295 31.234-27.477 36.453-15.94 66.258-
11.57c38.141 5.591 6.432 57.391-37.553
63.545-43.805 6.128-56.88-14.101-59.939-
24.498z" fill="#f2e7cb" />
        <path d="m64.006 50.121c-11.747-
6.451 2.471-8.9 2.471-8.9 18.195-2.442
20.9-11.164 32.278-10.175 20.09 1.747 9.23
22.95.881 26.146 0 0-7.734 2.76-19.6-
.637a87.6 87.6 0 0 1 -16.03-6.434z"
fill="#ef3829" />
        <path d="m33.546 43.2s-19.075-.978-
15.81 13.551 25.466 24.3 55.5 17.283-7.336-
11.407-16.483-15.978-4.44-13.196-23.207-
14.856z" fill="#ef3829" />
        <path d="m121.7 38.052c0 17.53-
24.409 43.261-54.6 47.485-43.805 6.128-
56.88-14.1-59.939-24.5a26.44 26.44 0 0 1 -
.844-6.82v19.859a26.436 26.436 0 0 0 .844
6.961c3.059 10.4 16.134 30.626 59.939 24.5
30.14-4.217 54.516-29.867 54.6-47.4v-
20.086z" fill="#ef983b" />
        <g fill="#422002">
```

Fig. 5. Sample SVG Script with XML Tag First.

TABLE I. SAMPLES OF FILES UPLOAD

No.	File Extension	Magic Number	XML/DOM
1	✕	✕	✕
2	✓	✕	✕
3	✕	✓	✕
4	✕	✕	✓
5	✓	✓	✕
6	✓	✕	✓
7	✕	✓	✓
8	✓	✓	✓

Table I contains eight SVG samples prepared for SVG validation in the system by removing some of the three parameters (validation file extension, the magic number, and XML) from the file, as in Table II.

TABLE II. SVG TYPES FILES

File Extension	ASCII		Magic Number	
	Start of File	End Of File	Start of File	End Of File
svg	<?xml	</svg>	3C 3F 78 6D 6C	3C 2F 73 76 67 3E
svg	<svg	</svg>	3C 73 76 67 20	3C 2F 73 76 67 3E

Table II contains the types of information contained in the SVG file used based on the prefix tag "<?xml" with the magic number value "3C 3F 78 6D 6C" and MIME "text/xml" while "<svg " has a magic number value "3C 73 76 67 20" and MIME is" image/svg". The scenarios in Table III described as follows :

3) PHP Code in TXT file: Fig. 6 illustrates a PHP file renamed file extension from ".php" to ".txt" with file contents like Fig. 7. In this scenario, do not use the file extension, magic number, and XML format.

Fig. 7 contains PHP code by displaying PHP settings or server information by changing the file extension from ".php" to ".txt".

4) PHP Code in SVG file: Fig. 8 illustrates a PHP file renamed file extension from ".php" to ".svg" with file contents like Fig. 9. In this scenario, use the SVG file extension but do not use the magic number at the start of the file ("3C 3F 78 6D 6C" or "3C 73 76 67 20") and the end of the file (3C 2F 73 76 67 3E) and XML format.

Fig. 9 contains PHP code by displaying PHP settings or server information by changing the file extension from ".php" to ".svg".

5) PHP Code in SVG tag TXT file: PHP Code file renamed file extension from ".php" to ".txt" with file contents like Fig. 10. In this scenario, use the magic number at the start of the file ("3C 3F 78 6D 6C" or "3C 73 76 67 20") and the end of the file (3C 2F 73 76 67 3E) but do not use XML format and the SVG file extension.

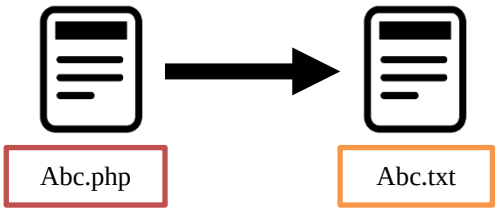


Fig. 6. Illustration of Changing the File Extension from PHP to TXT.

```
<?php
phpinfo();
?>
```

Fig. 7. PHP Code in TXT File.

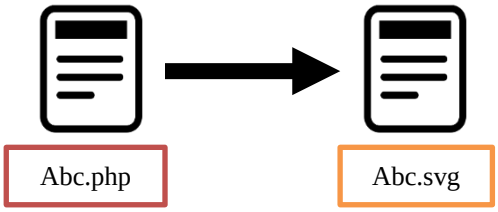


Fig. 8. Illustration of Changing the File Extension from PHP to SVG.

```
<?php
phpinfo();
?>
```

Fig. 9. PHP Code in SVG File.

```
<svg height="512" viewBox="0 0 128 128" width="512"
xmlns="http://www.w3.org/2000/svg">
<?php
phpinfo();
?>
```

Fig. 10. PHP Code in SVG Tag TXT File.

Fig. 10 contains PHP code by displaying PHP settings or server information with SVG tag by changing the file extension from ".php" to ".txt".

6) XML tag in TXT file: SVG tags with XML tag first or magic number value "3C 3F 78 6D 6C" with file contents like Fig. 11. In this scenario, use XML format but do not use the SVG file extension and magic number at the start of the file ("3C 3F 78 6D 6C" or "3C 73 76 67 20") and the end of the file (3C 2F 73 76 67 3E).

7) PHP Code in SVG tag first: Fig. 12 contains PHP code by displaying PHP settings or server information in SVG code. In this scenario, use the file extension and magic number ("3C 3F 78 6D 6C" or "3C 73 76 67 20") but do not use or invalid XML format.

8) XML Code in SVG file: Fig. 13 contains XML tags in SVG file. In this scenario, use the file extension and XML format but do not use the magic number at the start of the file ("3C 3F 78 6D 6C" or "3C 73 76 67 20") and the end of the file (3C 2F 73 76 67 3E).

9) SVG Code in TXT file: Fig. 14 contains SVG tags in the TXT file. In this scenario, do not use the SVG file extension but use the magic number at the start of the file ("3C 3F 78 6D 6C" or "3C 73 76 67 20") and the end of the file (3C 2F 73 76 67 3E) but XML format.

10)SVG Code with SVG/XML first tag in SVG file: Fig. 15 contains SVG tags in SVG file extension. In this scenario, use the file extension and magic number at the start of the file ("3C 3F 78 6D 6C" or "3C 73 76 67 20") and the end of the file (3C 2F 73 76 67 3E) and valid XML format.

```
<?xml version="1.0" standalone="no"?>
<svg height="512" viewBox="0 0 128 128" width="512"
xmlns="http://www.w3.org/2000/svg">
    <g>
        <path d="m7.157 61.039s-8.924-30.295 31.234-
27.477 36.453-15.94 66.258-11.57c38.141 5.591 6.432
57.391-37.553   63.545-43.805   6.128-56.88-14.101-
59.939-24.498z" fill="#f2e7cb" />
.................................................
        </g>
    </g>
</svg>
```

Fig. 11. XML Tag in TXT File.

```
<?xml version="1.0" standalone="no"?>
<svg height="512" viewBox="0 0 128 128" width="512"
xmlns="http://www.w3.org/2000/svg">
    <g>
.....................................................
    </g>
</svg>
<?php
phpinfo();
?>
```

Fig. 12. PHP Code in SVG Tag First.

```
<?xml version="1.0" encoding="UTF-8"?>
<svg height="512" viewBox="0 0 128 128" width="512"
xmlns="http://www.w3.org/2000/svg">
<note>
  <to>Fahmi</to>
  <from>Anwar</from>
  <heading>Reminder</heading>
  <body>Don't forget me this weekend!</body>
</note>
```

Fig. 13. XML Code in SVG File.

```
<svg height="512" viewBox="0 0 128 128" width="512"
xmlns="http://www.w3.org/2000/svg">
    <g>
        <path d="m7.157 61.039s-8.924-30.295 31.234-
27.477 36.453-15.94 66.258-11.57c38.141 5.591 6.432
57.391-37.553   63.545-43.805   6.128-56.88-14.101-
59.939-24.498z" fill="#f2e7cb" />
.................................................
        </g>
    </g>
</svg>
```

Fig. 14. SVG Code in TXT File.

```
<svg height="512" viewBox="0 0 128 128" width="512"
xmlns="http://www.w3.org/2000/svg">
    <g>
        <path d="m7.157 61.039s-8.924-30.295 31.234-
27.477 36.453-15.94 66.258-11.57c38.141 5.591 6.432
57.391-37.553   63.545-43.805   6.128-56.88-14.101-
59.939-24.498z" fill="#f2e7cb" />
.................................................
        </g>
    </g>
</svg>
```

Fig. 15. SVG Code with SVG/XML First Tag in SVG File.

C. Modeling

The modeling stage contains the process carried out by the system in the form of a flowchart, as shown in Fig. 16 for general validation and Fig. 17 for new validation.

1) General Validation SVG File: Fig. 16 is an SVG image validation flowchart utilizing file extensions and MIME type using the function mime_content_type($file) in PHP programming language.

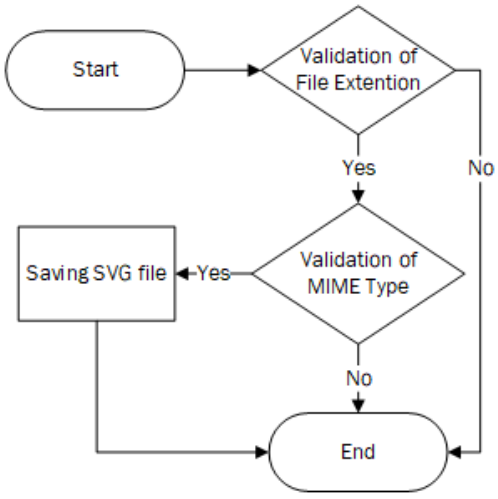


Fig. 16. Flowchart of General Validation SVG File.

2) New Validation SVG File: Fig. 17 is an SVG image validation flowchart utilizing file extensions, magic numbers, and DOM using the PHP programming language.

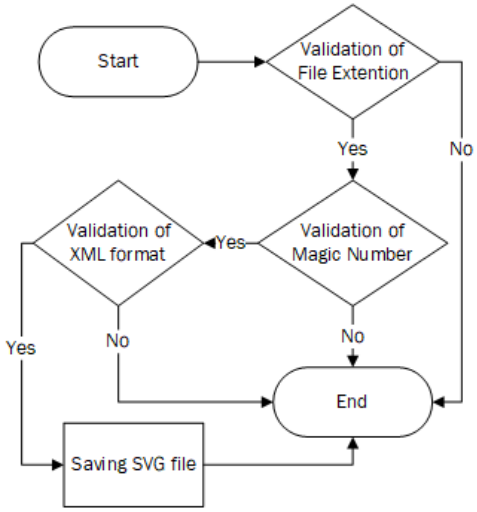


Fig. 17. Flowchart of New Validation SVG File.

D. Construction

The construction stage contains the system's steps in the form of a flowchart converted into Algorithm 1.

1) General Validation SVG File: Algorithm 1 contains the PHP programming algorithm, which functions to filter or validate SVG images by utilizing file extensions and mime using the PHP programming language. The algorithm created is tested using black-box testing with the test scenarios presented in Table III.

Algorithm 1 General Validation SVG File Upload

```
<?php
$filePath = "Abc.svg";
$errors = array();
$fileExtensionWhitelist = array("SVG");
$imageFileExtension = pathinfo($filePath, PATHINFO_EXTENSION);

if (!in_array(strtoupper($imageFileExtension), $fileExtensionWhitelist)){
 array_push($errors,"File Extension is not allowed");
}

if(mime_content_type($filePath)!="image/svg+xml"){
 array_push($errors,"MIME is not allowed");

}

if(empty($errors)){
 echo "SVG is valid";
}else{
 foreach ($errors as $data) {
 echo $data."<br />";
 }
}
?>
```

The algorithm tested using black-box testing by validation types with the results in Table III. All test scenarios have been tested by comparing the expected results with the actual results, then the conclusion of the desired results has been successful or appropriate. The test results in Table III contain one scenario

that matches expectations and 6 scenarios that do not match expectations from 8 scenarios so that the success rate is 75%.

2) New Validation SVG File: Algorithm 2 contains the PHP programming algorithm, which functions to filter or validate SVG images by utilizing file extensions, magic numbers, and DOM using the PHP programming language.

Algorithm 2 New Validation SVG File Upload

```
<?php
$filePath = "Abc.svg";
$errors = array();
$fileExtensionWhitelist = array("SVG");
$imageFileExtension = pathinfo($filePath, PATHINFO_EXTENSION);

if (!in_array(strtoupper($imageFileExtension), $fileExtensionWhitelist)){
 array_push($errors,"File Extension is not allowed");
}

function magicNumberStartOfFile($filename) {
 if(file_exists($filename)){
 $handle = fopen($filename, 'r');
 $bytes = strtoupper(bin2hex(fread($handle, 5)));
 fclose($handle);
 return $bytes;
 }else{
 return false;
 }
}

function magicNumberEndOfFile($filename) {
 if(file_exists($filename)){
 $handle = fopen($filename, 'r');
 fseek($handle, -6, SEEK_END);
 $bytes = strtoupper(bin2hex(fread($handle, 6)));
 fclose($handle);
 return $bytes;
 }else{
 return false;
 }
}

$magicNumberStartOfFileWhitelist = array("3C3F786D6C", "3C73766720");
if (!in_array(magicNumberStartOfFile($filePath),
$magicNumberStartOfFileWhitelist)){
 array_push($errors,"Magic Number (Start of File) is not allowed");
 }

$magicNumberEndOfFileWhitelist = array("3C2F7376673E");
if (!in_array(magicNumberEndOfFile($filePath),
$magicNumberEndOfFileWhitelist)){
 array_push($errors,"Magic Number (End of File) is not allowed");
 }

libxml_use_internal_errors(TRUE);
$dom = new DOMDocument;
$dom->Load($filePath);
if ($dom->validate()) {
array_push($errors,"XML format is not valid");
var_dump(libxml_get_errors());
}

if(empty($errors)){
 echo "SVG is valid";
}else{
 foreach ($errors as $data) {
 echo $data."<br />";
 }
}
?>
```

TABLE III. BLACK-BOX TESTING OF GENERAL VALIDATION

No.	Scenarios of Validation			Expected Result	Actual Result
	File Extension	*Magic Number*	*Document Object Model*		
1	×	×	×	Uploaded failed	[✓] Succeed [ ] Failed
2	✓	×	×	Uploaded failed	[✓] Succeed [ ] Failed
3	×	✓	×	Uploaded failed	[✓] Succeed [ ] Failed
4	×	×	✓	Uploaded failed	[✓] Succeed [ ] Failed
5	✓	✓	×	Uploaded failed	[ ] Succeed [✓] Failed
6	✓	×	✓	Uploaded failed	[ ] Succeed [✓] Failed
7	×	✓	✓	Uploaded failed	[✓] Succeed [ ] Failed
8	✓	✓	✓	Uploaded succeed	[✓] Succeed [ ] Failed

All test scenarios have been tested by comparing the expected results with the actual results, then the conclusion of the desired results has been successful or appropriate. The algorithm tested using a black-box of general validation with the results in Table IV.

The test results in Table IV contain six scenarios that match expectations and all scenarios that match expectations from 8 scenarios so that the success rate is 100%.

TABLE IV. BLACK-BOX TESTING OF NEW VALIDATION

No.	Scenarios of Validation			Expected Result	Actual Result
	File Extension	*Magic Number*	*Document Object Model*		
1	×	×	×	Uploaded failed	[✓] Succeed [ ] Failed
2	✓	×	×	Uploaded failed	[✓] Succeed [ ] Failed
3	×	✓	×	Uploaded failed	[✓] Succeed [ ] Failed
4	×	×	✓	Uploaded failed	[✓] Succeed [ ] Failed
5	✓	✓	×	Uploaded failed	[✓] Succeed [ ] Failed
6	✓	×	✓	Uploaded failed	[✓] Succeed [ ] Failed
7	×	✓	✓	Uploaded failed	[✓] Succeed [ ] Failed
8	✓	✓	✓	Uploaded succeed	[✓] Succeed [ ] Failed

E. Deployment

This research can be applied to the SVG file upload algorithm using magic numbers and DOM after other validations, such as validating file extensions to check SVG validation in the file upload process by performing superior filtering with validation of writing XML structures so that they can filter. SVG text that conforms to a standard XML format is incompatible.

IV. Conclusion

This research produces an application that can provide security in uploading files to web-based applications, especially SVG files. The Waterfall method is used to develop or build software because of the many preparatory stages before the software development stage. Handling of security validation for uploading SVG files using file extensions and MIME types has a success rate of 75% from the eight tested scenarios while handling using file extensions, magic numbers, and Document Object Model (DOM) a success rate of 100% from 8 test scenarios. Testing uses a black-box so that handling using the file extension, magic number, and Document Object Model (DOM) is better than using only file extensions and mime types. Subsequent research work is that the proposed method must be validated by various unique classifications of SVG files or other file formats.

Acknowledgment

This research is supported by Direktorat Riset dan Pengabdian Masyarakat, Direktorat Jenderal Penguatan Riset dan Pengembangan Kementerian Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia. Surat Kontrak Pelaksanaan Penelitian Kementerian Riset dan Teknologi/ Badan Riset dan Inovasi Nasional (KEMENRISTEK/BRIN) Tahun Tunggal Tahun Anggaran 2020 Nomor: PTM-019/SKPP.TT/LPPM UAD/VI/2020.

References

[1] H. Chen, L. J. Zhang, B. Hu, S. Z. Long, and L. H. Luo, “On Developing and Deploying Large-File Upload Services of Personal Cloud Storage,” Proc. - 2015 IEEE Int. Conf. Serv. Comput. SCC 2015, pp. 371–378, 2015, doi: 10.1109/SCC.2015.58.

[2] X. Li and Y. Xue, “A survey on web application security,” Nashville, TN USA, 2011, [Online]. Available: http://isis.vanderbilt.edu/sites/default/files/main_0.pdf.

[3] A. Yudhana, I. Riadi, and F. Ridho, “DDoS classification using neural network and naïve bayes methods for network forensics,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 11, pp. 177–183, 2018, doi: 10.14569/ijacsa.2018.091125.

[4] I. Riadi, A. W. Muhammad, and Sunardi, “Neural network-based ddos detection regarding hidden layer variation,” J. Theor. Appl. Inf. Technol., vol. 95, no. 15, pp. 3684–3691, 2017.

[5] A. Iswardani and I. Riadi, “Denial of service log analysis using density K-means method,” J. Theor. Appl. Inf. Technol., vol. 83, no. 2, pp. 299–302, 2016.

[6] A. Fadlil, I. Riadi, and S. Aji, “DDoS Attacks Classification using Numeric Attribute-based Gaussian Naive Bayes,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 8, pp. 42–50, 2017, doi: 10.14569/ijacsa.2017.080806.

[7] A. Kurniawan, I. Riadi, and A. Luthfi, “Forensic analysis and prevent of cross site scripting in single victim attack using open web application security project (OWASP) framework,” J. Theor. Appl. Inf. Technol., vol. 95, no. 6, pp. 1363–1371, 2017.

[8] S. B. Almi, “Web Server Security and Survey on Web Application Security,” Int. J. Recent Innov. Trends Comput. Commun., vol. 2, no. 1, pp. 114–119, 2014, [Online]. Available: http://ijritcc.org/IJRITCC Vol_2 Issue_1/Web Server Security and Survey on Web Application Security.pdf.

[9] A. Jaiswal, G. Raj, and D. Singh, “Security Testing of Web Applications: Issues and Challenges,” Int. J. Comput. Appl., vol. 88, no. 3, pp. 26–32, 2014, doi: 10.5120/15334-3667.

[10] OWASP, “OWASP Top 10 2017 - The Ten Most Critical Web Application Security Risks Release Candidate 2,” pp. 1–25, 2017, [Online]. Available: https://www.owasp.org/images/b/b0/OWASP_Top_10_2017_RC2_Final.pdf.

[11] WhiteHatSec, “2018 Application Security Statistics Report - The Evolution of the Secure Software Lifecycle,” 2018. https://info.whitehatsec.com/Content-2018StatsReport_LP.html.

[12] V. Subramaniyaswamy, G. Venkata Kalyani, and N. Likhitha, “Securing web applications from malware attacks using hybrid feature extraction,” Int. J. Pure Appl. Math., vol. 119, no. 12, pp. 13367–13385, 2018.

[13] K. Pooj and S. Patil, “Understanding File Upload Security for Web Applications,” Int. J. Eng. Trends Technol., vol. 42, no. 7, pp. 342–347, 2016, doi: 10.14445/22315381/ijett-v42p261.

[14] W3C, “Scalable Vector Graphics (SVG) 1.1 (Second Edition),” W3C, 2011. https://www.w3.org/TR/SVG11/ (accessed Oct. 20, 2020).

[15] S. D. Ankush, “XSS Attack Prevention Using DOM based filtering API XSS Attack Prevention Using DOM based fitering API,” 2014.

[16] M. Johns, B. Engelmann, and J. Posegga, “XSSDS: Server-side detection of Cross-site Scripting attacks,” Proc. - Annu. Comput. Secur. Appl. Conf. ACSAC, pp. 335–344, 2008, doi: 10.1109/ACSAC.2008.36.

[17] D. Zubarev and I. Skarga-Bandurova, “Cross-Site Scripting for Graphic Data: Vulnerabilities and Prevention,” Conf. Proc. 2019 10th Int. Conf. Dependable Syst. Serv. Technol. DESSERT 2019, pp. 154–160, 2019, doi: 10.1109/DESSERT.2019.8770043.

[18] I. Riadi and E. I. Aristianto, “An Analysis of Vulnerability Web Against Attack Unrestricted Image File Upload,” Comput. Eng. Appl. J., vol. 5, no. 1, pp. 19–28, 2016, doi: 10.18495/comengapp.v5i1.161.

[19] I. Riadi, A. Fadlil, and T. Sari, “Image Forensic for detecting Splicing Image with Distance Function,” Int. J. Comput. Appl., vol. 169, no. 5, pp. 6–10, 2017, doi: 10.5120/ijca2017914729.

[20] R. A. Surya, A. Fadlil, and A. Yudhana, “Identification of Pekalongan Batik Images Using Backpropagation Method,” J. Phys. Conf. Ser., vol. 1373, no. 1, 2019, doi: 10.1088/1742-6596/1373/1/012049.

[21] R. S. Pressman and B. R. Maxim, Software Engineering : a practitioner’s approach, 8th ed. New York: McGraw-Hill Education, 2014.

[22] J. Bau, E. Bursztein, D. Gupta, and J. Mitchell, “State of the art: Automated black-box web application vulnerability testing,” Proc. - IEEE Symp. Secur. Priv., pp. 332–345, 2010, doi: 10.1109/SP.2010.27.

[23] IANA, “Media Types,” 2020. https://www.iana.org/assignments/media-types/media-types.xhtml.

[24] Sloth at 0k dot vc, “PHP :: Bug #79045 :: Incorrect svg mimetypes detected,” 2019. https://bugs.php.net/bug.php?id=79045 (accessed Nov. 21, 2020).

A Pilot Study of an Instrument to Assess Undergraduates' Computational thinking Proficiency

Debby Erce Sondakh[1], Kamisah Osman[2], Suhaila Zainudin[3]
Faculty of Education, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia[1, 2]
Faculty of Computer Science, Universitas Klabat, Manado, Indonesia[1]
Faculty of Technology and Information Science, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia[3]

***Abstract*—The potentiality of computational thinking (CT) in problem solving has gained much attention in academic communities. This study aimed at developing and validating an instrument, called Hi-ACT, to assess CT ability of university undergraduates. The Hi-ACT evaluates both technical and soft skills applicable to CT based problem solving. This paper reports a pilot study conducted to test and refine the initial Hi-ACT. Survey method was employed through which questionnaire comprising of 155 items was piloted among 548 university undergraduates. Structural equation modeling with partial least squares was applied to examine the Hi-ACT's reliability and validity. Composite reliability was used to assess internal consistency reliability, while convergent validity was evaluated using based on items' outer loadings and constructs' average variance extracted. As a result, 41 items were excluded, and an instrument to assess CT ability comprising 114 items and ten constructs (abstraction, algorithmic thinking, decomposition, debugging, generalization, evaluation, problem solving, teamwork, communication, and spiritual intelligence) was developed. The reliability and validity of the Hi-ACT in its pilot form have been verified.**

Keywords*—Computational thinking; assessment; skills; attitudes; undergraduates; self-assessment

I. Introduction

The ability to solve a complex problem is demanded, regardless of the field in which we work. Wing [1] introduced computational thinking (CT) as problem solving approach that using the way computer scientist think. Her vision is, the set of CT skills and attitudes will be beneficial for everyone, not only computer science majors. Further studies had reinforced that CT enables one to become a technology builder rather than a mere technology consumer [2], develops logic, creativity, innovative thinking [3], and analytical skills [4], The World Economic Forum [5] found these attributes are increasingly in demand in the digital world workplaces.

Further, recognition of CT as an essential skill for all students is expanding rapidly. Accordingly, initiatives are underway to bring CT into educational institution around the world. Among them, a number of recent studies focus on incorporating CT in classroom/curriculum [6], [7], some on creating artifacts with which to teach CT principles [8], [9], as well as on assessment [10]–[12]. Other studies [13]–[16] highlighted teachers' conception of CT.

In this work, we focus on CT assessment at the undergraduate level. Some studies have initiated CT assessments for undergraduates. An instrument, which tries to test the correlation between CT and critical thinking has been developed [17]. Specifically, this instrument assesses simple algorithms, sorting method, file structure, and digital information storage. The author used multiple choice questions and short answer questions. However, it has not been validated.

In another study [18] a test to identify CT skill of first-year computer science students was developed. It was based on six classes of CT skills and practices defined in 'Computational Thinking Framework,' i.e. models and abstractions, patterns and algorithms, processes and transformation, tools and resources, inference and logic, and evaluations and improvements.

A paper-based test, called 'The Testing Algorithmic and Application Skills' is presented [19]. This test measures algorithmic skills, computer science terminology used, and problem solving abilities. Particularly, it comprises questions related to the students' computer usage habits, self-assessment on their knowledge in informatics, and tasks of traditional programming, numerical system calculation, handling files, word processing, and spreadsheet programming.

The aforementioned CT assessment studies mostly highlighted the skills, and little has been done to include attitudes. Contrariwise, according to Wing [1], CT comprises both skill and attitudes necessary in solving problems. There is thus a need for an instrument that includes items and constructs to measure students' CT competency, in terms of skills and attitudes. Therefore, in the light of Wing's original conception of CT as a set of skills and attitudes to solve problem, this work proposes an instrument, the "Holistic Assessment of Computational Thinking" (Hi-ACT), to test undergraduates' perceptions of their CT competency. We use the term 'holistic' to describe the inclusion of both skills and attitudes in the CT assessment framework. This paper reports a pilot study that was conducted to assess and refine the initial Hi-ACT by examining its reliability and validity. This is an extended version of the work published in [20].

II. Computational Thinking

CT has been noticed as a major research field since the publication of Wing's remarkable article in 2006. However, several researchers noted the long history of CT [21]–[24], as presented in Fig. 1. As early as 1945, George Polya emphasized the application of disciplined manner, decomposition, and generalization (reuse common techniques)

to solve the everyday problem [24]. In 1962, Perlis proposed his vision that programming concept would foster the ability to understand various topic outside computer science and become a vital part of education [21]. As noted by Denning [22], from the field of science, L.K. Wilson introduced the 'computational science,' a computation-based approach to exploit existing knowledge and discover the new one. Thereafter, Seymour Papert, in 1980, found that 'thinking like a computer' was a useful component of thinking skills to teach mathematics to children [23]. Programming symbols and representation were used in solving mathematical problems.

Further, in 2006, Wing introduced CT, a way of thinking to solve problems, design systems, understand human behavior, using computer science based concepts. Several researchers [2], [25]–[27] revisited Wing's definition of CT provide a definite understanding of CT and to perceive its core principles. Denning [26] defined CT as a mental orientation to formulate problems through what so-called 'conversion'. Algorithms are applied to convert some input into an output. Other studies described CT as problem solving process [2], and the essence of is 'thinking like a computer scientist' [27]. Wing refined her early delineation of CT to be "the thought processes involved in formulating problems and their solutions that can be effectively carried out by an information-processing agent; a human or machine, or combinations of humans and machines" [28]. Aho [25] then simplified Wing's refinement by defining CT as the thought processes to formulate solutions to the problems, which represented as computational steps and algorithms. Put simply, the core of CT is to approach a problem using computer scientists' way of thinking.

CT adopts some fundamental concepts of computer science as its skills [1]. There are varying views considering CT skills (Table I). Along with the skills, attitudes are also required in CT-based problem solving [1]. Barr et al. [2] used the term 'dispositions' to describe the values, motivations, feelings, stereotypes and attitudes' appropriate to CT. It, therefore, can be said that, in CT, attitudes is indeed necessary for solving problems using. Nevertheless, as shown in Table II, only a few works of literature that considered attitudes.

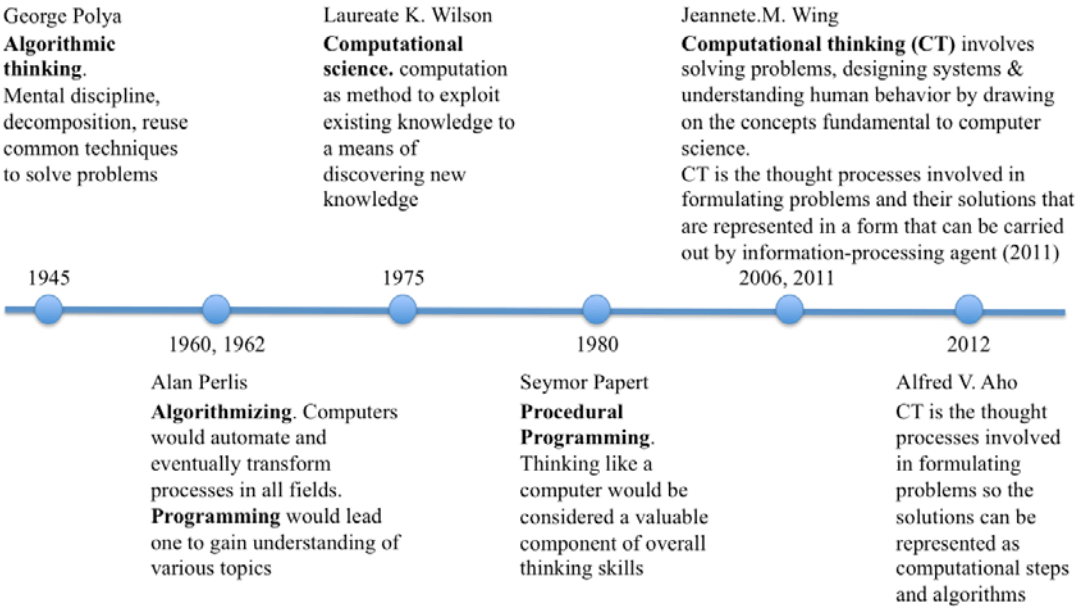


Fig. 1. Computational Thinking Evolution.

TABLE I. CT SKILLS DISCUSSED IN THE LITERATURE

	Abstraction	Algorithmic	Automation	Decomposition	Generalization	Logical reasoning	Evaluation	Debugging
Wing [1], [29]	√	√	√					
Barr et al. [2]	√	√	√	√	√	√		
Selby and Wollard [30]	√	√		√	√		√	
Grover and Pea [27]	√	√		√			√	√
Angeli et al. [31]	√	√		√	√			√
Bocconi et al. [32]	√	√	√	√	√			√
Mueller et al. [33]	√	√		√	√		√	√

TABLE II. CT ATTITUDES (SOFT-SKILLS) DISCUSSED IN THE LITERATURE

	Teamwork	Confidence	Ambiguity handling	Persistence	Coordination/Cooperation	Self-awareness	Creativity	Problem solving	Communication
Hambrusch et al. [34]	√								
Barr et al. [2]	√	√	√	√					
Kazimoglu et al. [35]					√				
Vergara et al. [36]	√					√			
Korkmaz et al. [10]					√		√	√	
Allsop [37]				√	√		√	√	√

III. RELATED WORKS

Korkmaz et al. [10] developed the 'Computational Thinking Scale' (CST), an instrument comprising 29 five-point Likert type items. The CST, which was tested on undergraduate students, assesses algorithmic thinking, critical thinking, creativity, problem solving, and cooperation. In a more recent study, a scale has been developed to assess high school students' computational thinking skill [11]. In the same way as [10], this study also develops the scale based on ISTE (2015)'s definition of computational thinking skill. The scale takes in five skills, i.e. problem solving, algorithmic thinking, critical thinking, cooperative learning, and creative thinking as the initial factors. Subsequent to validity and reliability examinations, the resulting scale consists of five-point Likert scale of 42 items categorized into four factors, i.e. problem solving, cooperative learning and critical thinking, creative thinking, and algorithmic thinking.

The aforementioned instruments evaluate some soft-skills relevant to CT (creativity, problem-solving, and cooperation); however, both fail to consider the abstraction, decomposition, pattern recognition, and generalization. In contrast, abstraction is a basic tool of reasoning in CT [38]. That is, abstraction allows one to make simpler the large and complex problems [39]. In the same way, the ability to recognize patterns and generalizing solutions is an invaluable skill for computer scientists [18]. Decomposition is needed to break-down a problem into smaller, simpler and more manageable sub-problems [39]. That is, these skills are essential in CT.

This work proposes the Hi-ACT, a CT proficiency assessment instrument that takes both skills and attitudes into consideration. The constructs measured are elaborated in detail in the next section.

IV. DEFINING HI-ACT

There is still little unanimity on CT definition, as [18], [40] inferred. Besides, its underlying skills are still being debated and redefined. Notwithstanding, to develop the Hi-ACT, we define CT as the thought process of formulating solutions to a problem that entails some skills and attitudes. The term 'skills' refers to computer science-based concepts used in CT, whereby this work draws on the work of [30], [32] to define CT core skills, including abstraction, algorithmic thinking, decomposition, debugging, evaluation, and generalization. The term 'attitudes' refers to soft-skills. Soft-skills are personal specific skills which include attitudes, character traits, and behaviors [41]. In this work, the attitudes were drawn from the Computer Science Curricula 2013, i.e. problem solving, ambiguity tolerance, teamwork, communication, and personal attributes [42], and the operational definition of CT attitudes previously stated by [2]. This collection of work on attitudes was synthesized into three categories: problem solving, teamwork, and communication.

Furthermore, this work suggests one additional element, i.e. spiritual intelligence, to be included as one of the attitudes (soft-skills) of CT. The justification for such inclusion is: spiritual intelligence comprising a set of abilities that encourage people's ability to solve problems, achieve goals, and enhance decision-making capability [43]. In this way, spiritual intelligence might be beneficial to CT as a way of thinking about solving problems. Moreover, there are some attitudes that demonstrate spiritual intelligence, i.e., self-awareness, creative reasoning, integrity, and asking 'why' questions, which are found helpful when confronting challenging problems, including artificial intelligence problems [44]. Hence, including spiritual intelligence would be beneficial in CT-based problem solving process.

In summary, this work develops an instrument, the Hi-ACT, which evaluates undergraduates' perceptions of their CT competency. Based on the following literature [2], [28], [45]–[48], [29]–[32], [35], [39], [43], [44], the authors define each associated CT skills included in the Hi-ACT.

- Abstraction: the ability to simplify a problem by removing unnecessary details or information, then create a representation of the solution.
- Algorithmic Thinking: the ability to thinking algorithmically in formulating the instructions (procedure) through logical thinking to solve a problem.
- Decomposition: the ability to simplify a problem by dividing it into smaller, simpler, and easier to manage sub-problems.
- Debugging: the ability to identify and remove errors in the designed solutions (the algorithm).

- Evaluation: the ability to assess the solution's correctness, performance, resource usage, and the action of refining to improve the solution's quality.
- Generalization: the ability to identify similar patterns between the problems and generalizing solutions of previous problems to similar ones.
- Problem solving: the characters applicable to problem solving process, including self-confidence, persistence, ambiguity handling, and willingness to solve the problem.
- Teamwork: the ability to work in a team.
- Communication: the ability to exchange information and knowledge, by means of verbal and non-verbal, within the member of teamwork.
- Spiritual intelligence: copes with the ability to use spiritual abilities, including self-awareness, integrity, and creative reasoning, in enhancing an individual's personal characters to facilitate problem-solving process.

V. Methodology

A. Hi-ACT Initial Instrument

The Hi-ACT was firstly designed with 172 7-point Likert scale candidate items. These items address one of the sub-construct presented in Table III. Sub-construct is construct categories among the candidate items that are defined to ensure the items' convergent validity. Further, the first version underwent a content validation process through experts' judgment in a three-round Fuzzy Delphi Study, as reported in [49]. As a result, the initial Hi-ACT comprising 155 items was ready for validity and reliability assessment.

B. Participants

The initial Hi-ACT was administered on a total sample of 713 undergraduate students, from STEM and non-STEM major of specializations. The participants were recruited from different departments (Computer Science, Economics, Social Sciences and Humanities, Design, Linguistic, Natural Sciences, Health, Engineering, Law, Medicine, and Education), from two universities located in two different cities in Indonesia and one university located in Malaysia. After removing the surveys that have not been completely filled in, the final usable sample size is 548. Prior to data collection, universities' approval was obtained. All participants were notified of their voluntary participation, anonymity and confidentiality were assured. The percentage of participants in term of gender was equal, 274 (50%) were male, and 274 (50%) were female. Regarding the major of specialization, from the total sample, 363 (66%) were registered as STEM-based.

C. Data Analysis

This work aimed at refining the initial Hi-ACT by examining its validity and reliability. To do so, the structural equation modeling with partial least squares (PLS-SEM) was chosen. This choice was made for two reasons. First, factor analysis is a common statistical method for conceptualizing the constructs when refining a new instrument [50]. Exploratory factor analysis is specifically intended to refine a set of items in a new instrument. In that regard, as argued by Hair, Hult, Ringer and Sarstedt [51], PLS-SEM is mainly used to develop theories in an exploratory study.

Second, PLS-SEM is suitable for a complex model [52]. Based on the literature analyzed in this work and the result of content validation, Hi-ACT comes up as a multi-dimension construct, i.e. the constructs and sub-constructs described in Table III. Thus, Hi-ACT was modeled as reflective-formative higher-order constructs, as shown in Fig. 2. This model comprises of 29 first-order constructs, i.e. the sub-constructs (AR, AC, ATPr, and so forth) and ten second-order constructs, i.e. the constructs (Abstraction, Algorithmic Thinking, Decomposition, and so forth). Finally, the second-order constructs are formative to the Hi-ACT construct. Each of the 155 items in the initial Hi-ACT was modeled as a reflective indicator of one of the 29 first-order constructs.

TABLE III. CT Skills and Attitudes in Hi-ACT

Constructs	**Sub-construct**
Abstraction	Remove unnecessary detail (AR)
	Choose the right model (AC)
Algorithmic Thinking	Procedural thinking (ATPr)
	Sequence action (ATS)
	Conditional (ATC)
	Repetition (ATR)
	Parallelism (ATPa)
	Logical thinking (ATL)
Decomposition	Divide and conquer (DD)
	Modularizing (DM)
Debugging	Debugging (DE)
Evaluation	Performance evaluation (EP)
	Iterative refinement (EI)
	Optimizing (EO)
Generalization	Pattern recognition (GP)
	Reuse (GU)
	Remix (GM)
Problem solving	Confidence (PSC)
	Persistent (PSP)
	Ambiguity handling (PSA)
	Willingness (PSW)
Teamwork	Cooperation (TCp)
	Coordination (TCd)
	Participation (TP)
	Conflict management (TCM)
Communication	Communication (COM)
Spiritual intelligence	Self-awareness (SIS)
	Integrity (SII)
	Creative reasoning (SIC)

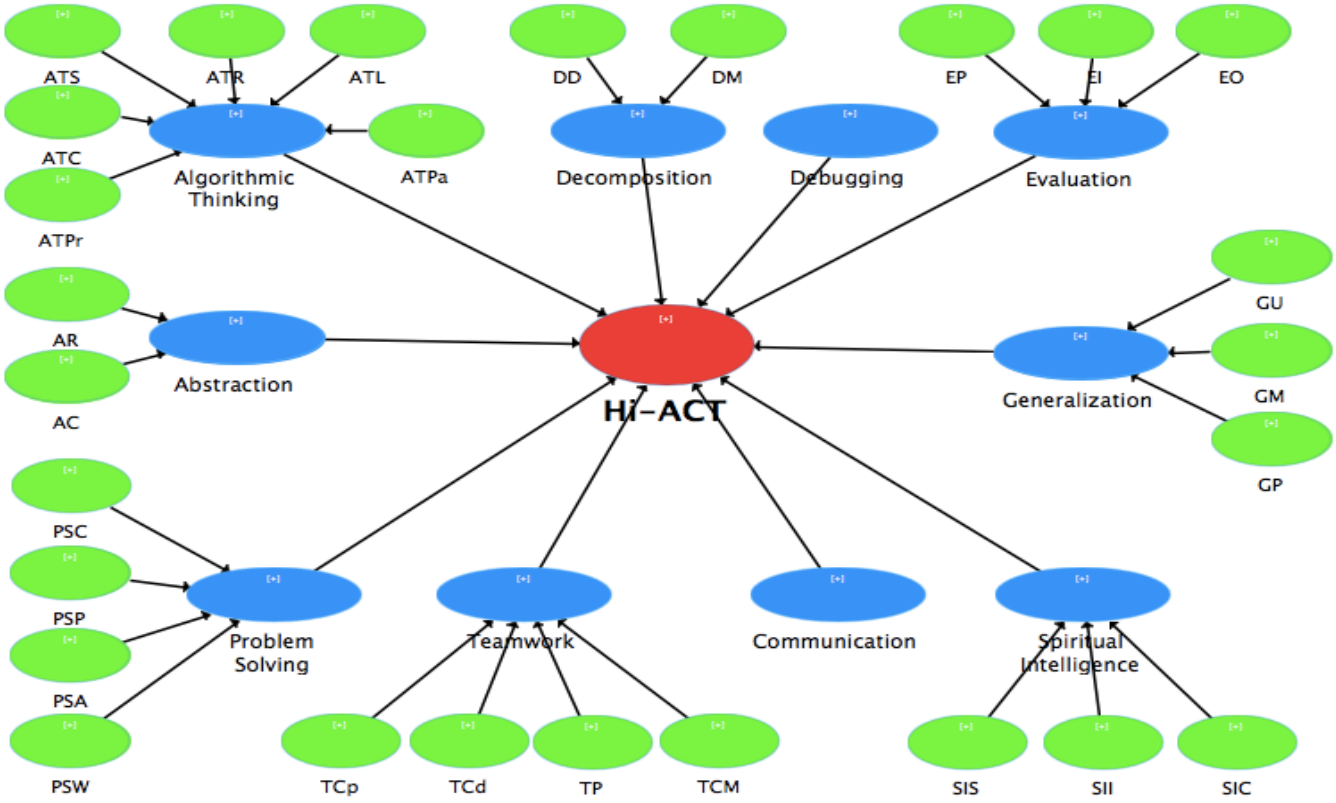


Fig. 2. The Hi-ACT.

In order to evaluate the instrument validity and reliability, the first-order constructs were evaluated. Evaluating reflective first-order constructs involves the examination of reliability and construct validity (convergent validity), which we deemed acceptable to analyze the result of new instrument pilot study. By referring to Hair, Hult, Ringer and Sarstedt [51], the following analyses were conducted:

- Internal consistency reliability as an evaluation of reliability. The internal consistency reliability was assessed by composite reliability (CR). It is desirable to have value within the range of 0.7 to 0.9, and it should not exceed 0.9.
- Convergent validity was assessed based on two criteria, i.e. items outer loadings and constructs' average variance extracted (AVE). Item's outer loading should be ≥ 0.7 and the AVE to be ≥ 0.5. When the AVE does not meet the required threshold, the item with the smallest loadings should be removed.

VI. Result

Table IV presents the internal consistency and convergent validity results. In the first run, the constructs CR value ranged from 0.7 to 0.93 exceed the threshold of 0.7. However, the TCp, SII, and SIC constructs' CR are higher than 0.9. CR value above 0.9 indicates that all the indicators measuring the same phenomenon, which is not a valid measure of the construct, and therefore is not desirable [51]. The second test was convergent validity. The loadings of all items ranged from 0.55 to 0.85, while the AVE ranged from 0.43 to 0.66. The AR, AC, ATS, ATR, DD, COM, SIS, SII, and SIC constructs' AVEs fall short of the threshold value of 0.5, indicating that the conditions of convergent validity were not met. Accordingly, items with low loadings were eliminated.

In summary, the initial instrument was refined by removing 38 items to improve each particular construct's AVE, and three items to shrink the CR value of TCp constructs to 0.9. Hence, the total of items eliminated was 41. This increased the AVEs while keeping the CRs in the threshold (Table V), then subsequently support the internal consistency and convergent validity conditions.

TABLE IV. RELIABILITY AND CONVERGENT VALIDITY RESULTS

Construct	Item Code	Item	Loading	AVE	CR	Reason for Exclusion
AR	AR-1	Reducing complexity by removing unnecessary detail	0.61	0.45	0.83	*
	AR-2	Evaluate what is valuable information and what is not	0.69			
	AR-3	Filtering the information when developing solution	0.69			
	AR-4	Separate the important from the redundant information	0.71			
	AR-5	Add or remove details to clarify a problem	0.62			*
	AR-6	Find appropriate level of detail to define and solve a problem	0.67			*
AC	AC-1	Understand that a model depicts general idea of the problem	0.72	0.48	0.74	
	AC-2	Create a model to solve problem	0.72			
	AC-3	Choosing a way to represent an artifact, allow it to be manipulated in useful ways	0.64			*
ATS	ATS-1	Formulate instructions to achieve a desired effect	0.65	0.46	0.81	*
	ATS-2	Create a set of precise steps to solve a problem	0.72			
	ATS-3	Elaborate particular activity or task as a series of individual steps or instructions	0.69			
	ATS-4	Explicitly wording the steps to solve problem	0.64			*
	ATS-5	Put instructions in the correct sequence	0.67			*
ATPr	ATPr-1	Select and execute appropriate steps to solve problem	0.76	0.54	0.82	
	ATPr-2	Identify the steps required to solve a problem	0.75			
	ATPr-3	Identify the sequence of steps including possible decisions and branching	0.77			
	ATPr-4	Understand normal and exceptional behaviors of a solution	0.65			*
ATC	ATC-1	Make decisions based on certain conditions	0.71	0.54	0.78	
	ATC-2	Think of possibility of different procedures for a problem	0.77			
	ATC-3	Produce many options while thinking of the possible solution regarding a problem	0.73			
ATR	ATR-1	Implement the same design plan for a specified number of times	0.56	0.43	0.75	*
	ATR-3	Repeat design processes to refine solutions until the ideal result is achieved	0.65			*
	ATR-4	Identify all possibilities of procedures that can be executed more than once	0.70			
	ATR-5	Decide, based on certain condition, when to execute a procedure and when to stop	0.70			
ATPa	ATPa-1	Running different sequences of instructions at the same time	0.76	0.64	0.78	
	ATPa-2	Dividing up resources and task in such a way to be processed in parallel	0.85			
ATL	ATL-1	Explain why something happens	0.73	0.54	0.87	
	ATL-2	Infer a conclusion based on existing knowledge	0.76			
	ATL-3	Explain how a conclusion is drawn	0.72			
	ATL-4	Provide the reason for my own thinking	0.72			
	ATL-5	Using existing knowledge to make reliable predictions	0.73			
	ATL-6	Elaborate logical connections between cause and effect	0.73			
DD	DD-1	Breaking apart problem into smaller subproblems to make it easier to solve	0.70	0.48	0.79	
	DD-2	Breaking down a problem into simpler version enables the same problem to be solved in the same way	0.72			
	DD-3	Apply the order of mathematical operations properly	0.62			*
	DD-4	Do classification	0.72			
DM	DM-1	Combine smaller parts to produce something larger	0.78	0.61	0.76	
	DM-2	Develop a solution by assembling together the smaller parts	0.77			
DE	DE-1	Think of anticipation plan for a problem	0.78	0.6	0.75	
	DE-2	Recognize problem when procedures do not correspond to solutions	0.77			
EP	EP-1	Determine whether the procedures in a solution is complete to solve the problem	0.75	0.53	0.85	
	EP-2	Assess whether the solution is suitable for solving the problem	0.73			
	EP-3	Assess whether the solution does the right thing	0.72			
	EP-4	Compare the performance of different procedures that solve the same problem	0.74			

	EP-5	Assess whether the solution is easy for people to use	0.68			*
EI	EI-1	Refine the solution procedure to improve its precision	0.78	0.58	0.8	
	EI-2	Evaluate solution against the success criteria	0.74			
	EI-3	Adjust the design and implementation of a solution when necessary	0.75			
EO	EO-1	Analyze the solution for efficient use of resources	0.79	0.56	0.84	
	EO-2	Develop a solution that can utilize the available resources	0.79			
	EO-3	Adapt the solution that can utilize the available resources	0.76			
	EO-4	After the problem solved, I analyze what went right and what went wrong	0.65			*
GU	GU-1	Applying an existing solution in a given problem to cover more possibilities	0.73	0.5	0.8	
	GU-2	Use sequence of instructions previously employed to solve a new problem	0.74			
	GU-3	Transfer ideas and solutions from one problem to another	0.70			
	GU-4	Building on other people's work	0.66			*
GM	GM-1	Embed other's work into my own work in a meaningful way	0.72	0.54	0.82	
	GM-2	Efficient in researching relevant information	0.75			
	GM-3	Constructively builds on contributions of others & integrates my own works with others'	0.75			
	GM-4	Combine and build on the ideas of others	0.72			
GP	GP-1	Identify patterns, similarities, and connections between prior and current problems	0.76	0.54	0.7	
	GP-2	Solve similar problems with the same set of steps	0.72			
PSC	PSC-1	I am a good problem solver	0.73	0.59	0.89	
	PSC-4	Confident to solve most problems	0.77			
	PSC-5	Given enough time, I believe I can solve most problems that confront me	0.77			
	PSC-7	Can solve new and difficult problems	0.79			
	PSC-8	Have a systematic method for comparing alternatives and making decisions	0.80			
	PSC-9	When I make plans to solve a problem, I am almost certain that I can make them work	0.78			
PSP	PSP-1	Can work a long time on a problem	0.68	0.55	0.88	*
	PSP-2	Keep working on a problem until I get the right answer	0.73			
	PSP-3	Keep trying even when a task becomes difficult	0.76			
	PSP-4	When a solution to a problem was unsuccessful, I will examine why it didn't work	0.75			
	PSP-5	When I'm confronted with a complex problem, I develop a strategy to collect information so that I can define exactly what the problem is	0.76			
	PSP-6	When my first effort to solve a problem fail, I still have certainty about my ability to handle the situation	0.76			
PSA	PSA-1	Anticipate impact of change and direct myself and others in smoothly shifting gears	0.74	0.52	0.89	
	PSA-3	Can guide others to cope with ambiguous situation effectively	0.7			
	PSA-4	Rise to challenge, accepting risk and uncertainty as normal	0.72			
	PSA-5	Remain calm and focus during time of change	0.74			
	PSA-6	Willing and open to change	0.69			*
	PSA-7	Adaptable with the unknown	0.68			*
	PSA-8	Have no problems with demonstrating the solution of a problem in my mind	0.71			
	PSA-10	Apply the solution I plan respectively	0.78			
PSW	PSW-1	Like to try to solve problems	0.84	0.66	0.85	
	PSW-2	It is fun to try to solve problems	0.79			
	PSW-3	Will try to solve almost any problems	0.81			
TCp	TCp-1	Enjoy working together with other	0.7	0.54	0.93	/
	TCp-2	Share the power with others	0.78			
	TCp-3	Understand that there are shared knowledge and skills between team members	0.79			
	TCp-4	Create a cooperative atmosphere among the members when addressing problems	0.79			
	TCp-5	Listen to and consider other's opinions	0.73			
	TCp-6	Willing to ask others for help	0.69			/
	TCp-7	Trust other team members	0.6			/

	TCp-8	Set aside differences when work with others to achieve a common goal	0.71			
	TCp-9	Like to experience cooperative learning together with team members	0.76			
	TCp-10	In cooperative learning, I think I attain more successful results	0.73			
	TCp-11	Solve problems related to group project in cooperative learning	0.76			
	TCp-12	More ideas occur in cooperative learning	0.76			
TCd	TCd-2	Work together harmoniously	0.77	0.59	0.81	
	TCd-3	Communicates actively and constructively	0.78			
	TCd-4	Acknowledge contribution of others	0.75			
TP	TP-1	Stay focus on the task during group work	0.75	0.58	0.87	
	TP-2	Fulfill the role assigned by the group	0.75			
	TP-3	Participate actively and accept a fair share of group work	0.81			
	TP-4	Work skilfully on the task assigned to me and complete them on time	0.72			
	TP-5	Share responsibilities for the team's success or failure	0.77			
TCM	TCM-1	Respond to and manage direct/indirect conflict constructively and effectively	0.71	0.51	0.86	
	TCM-2	Fully accept each other's strengths and weaknesses	0.73			
	TCM-3	Try to achieve harmony by avoiding conflict	0.7			
	TCM-4	Takes criticism in a friendly way	0.73			
	TCM-5	Avoid using put-down or blaming others	0.68			*
	TCM-6	Accept compromise to deal with conflict	0.74			
COM	COM-1	Like to share feelings, ideas, or opinions with others	0.66	0.48	0.88	*
	COM-2	Speak clearly with acceptable vocabulary	0.73			
	COM-3	Use a variety of communication means (written message, e-mail, phone, informal discussion)	0.68			
	COM-4	Limit length of comments so others can talk	0.64			*
	COM-5	Listen to everyone and respect their view	0.68			
	COM-6	Contribute appropriately in healthy debate	0.74			
	COM-7	Ensure consistency between words and tone	0.72			
	COM-8	Ensure consistency between facial expression and body languages	0.68			*
SIS	SIS-1	Aware of one's abilities and weaknesses	0.63	0.46	0.89	*
	SIS-2	Live with self-respect	0.66			*
	SIS-3	Satisfied with who I am	0.55			*
	SIS-4	Do any work with self-confidence	0.66			*
	SIS-5	Can decide my own goal	0.71			
	SIS-6	Consistently look for and try to discover my blind spots	0.68			*
	SIS-7	In negotiating, I try to see things from other's perspective even when I disagree	0.68			
	SIS-8	During an activity or conversation, I monitor and notice my thoughts and emotions	0.73			*
	SIS-9	My actions are aligned with my true essential nature	0.74			
	SIS-10	Aware of one's inner truth, what one know inside to be true	0.73			
SII	SII-1	Proud of one's country culture	0.58	0.49	0.93	*
	SII-2	Believe that character is one's real strength	0.68			*
	SII-3	Be aware of my own values and beliefs	0.74			
	SII-4	Keep the promises given to others	0.73			
	SII-5	My actions are aligned with my values	0.76			
	SII-6	I accept myself as I am with all my problems and limitations	0.73			
	SII-7	Know how to be myself when interacting with others	0.74			
	SII-8	Help and support others	0.69			*
	SII-9	Respect and trust others	0.66			*
	SII-10	Being open and honest with others	0.64			*
	SII-11	Put one's consciousness in a positive direction	0.72			
	SII-12	Live one's values in relationship with others	0.72			

	SII-13	Act with honesty and truthfulness	0.67			*
	SII-14	Keep working diligently even when no one is watching	0.68			
SIC	SIC-1	In solving problems, I draw on my own ability to hold, accept and go beyond paradoxes	0.73	0.47	0.91	
	SIC-2	Can integrate the seemingly contradictory points of view	0.76			
	SIC-3	Find it challenging to find out what the truth is	0.70			
	SIC-4	Can think of an answer to a problem, even though at first apparently no solution	0.70			
	SIC-5	Can offer new ways of viewing to a problem	0.71			
	SIC-6	Can find an unusual way to solve a problem	0.62			*
	SIC-7	Eager to know many things	0.69			
	SIC-8	Willing to change my mind and try something else	0.65			*
	SIC-9	Willing to admit when I made a wrong decision	0.66			*
	SIC-10	Can improve the original idea	0.64			*
	SIC-11	Can express one's ideas well	0.67			*

* Remove to refine the construct's AVE

/ Remove to refine the construct's internal consistency reliability

TABLE V. RELIABILITY AND CONVERGENT VALIDITY RESULTS AFTER ITEMS REMOVAL

Construct	**Remaining Items**	**Loading**	**AVE**	**CR**
AR	3	0.72 – 0.74	0.53	0.77
AC	2	0.72 – 0.75	0.54	0.70
ATS	2	0.68 – 0.74	0.50	0.67
ATPr	3	0.76 – 0.78	0.59	0.81
ATC	3	0.70 – 0.78	0.54	0.78
ATR	2	0.69 – 0.72	0.50	0.67
ATPa	2	0.75 – 0.86	0.65	0.79
ATL	6	0.72 – 0.76	0.54	0.87
DD	3	0.71 – 0.74	0.53	0.77
DM	2	0.77 – 0.79	0.61	0.76
DE	2	0.77 – 0.78	0.60	0.75
EP	4	0.72 – 0.76	0.54	0.83
EI	3	0.74 – 0.79	0.58	0.80
EO	3	0.78 – 0.81	0.64	0.84
GU	3	0.71 – 0.77	0.56	0.79
GM	4	0.71 – 0.75	0.54	0.82
GP	2	0.71 – 0.77	0.54	0.70
PSC	6	0.73 – 0.80	0.59	0.90
PSP	5	0.72 – 0.79	0.58	0.87
PSA	6	0.70 – 0.77	0.53	0.87
PSW	3	0.79 – 0.84	0.66	0.85
TCp	9	0.71 – 0.79	0.57	0.90
TCd	3	0.75 – 0.78	0.59	0.81
TP	5	0.72 – 0.8	0.58	0.87
TCM	5	0.69 – 0.73	0.51	0.84
COM	5	0.68 – 0.74	0.5	0.83
SIS	4	0.70 – 0.75	0.54	0.82
SII	8	0.68 – 0.76	0.53	0.90
SIC	6	0.69 – 0.77	0.52	0.87

VII. Discussion

This pilot study yielded preliminary proof of Hi-ACT's potential psychometric properties, a scales aimed at assessing undergraduate CT skills more comprehensively, by incorporating both skills and attitudes. A total of ten constructs and 114 items were extracted for Hi-ACT. Within this frame, the factor loadings for all items ranged from 0.68 to 0.86. These values indicating that the items of each specific construct have much in common, and they are contributing to measuring each associated sub-construct. The convergent validity of the construct level (AVE) was confirmed with values were ranged from 0.5 to 0.66, satisfying the required threshold of 0.5. The internal consistency reliability was maintained in an acceptable range. Within the range from 0.67 to 0.9, the CR values exceeding 0.7 were obtained for most of the sub-constructs. Thus, indicating that high internal consistency was achieved.

Six sub-constructs, i.e., ATS, ATR, TCM, COM, SII, and SIC, have items with factor loadings less than the threshold value of 0.7. Low factor loadings might contribute to low CR and AVE. Particularly for ATS and ATR sub-constructs, the CR values (0.67) were slightly lower than other sub-constructs. The CR value of 0.67, indicating that the items only have shared common variance of 45%, which implies that the items in each construct are slightly weak to measure the construct. It could be that these two sub-constructs have very few items compared to other sub-constructs. Each sub-constructs has two items, and one of them has factor loading less than 0.7; ATC3 (0.68) and ATR4 (0.69), which leads to slightly low item reliability. Nevertheless, the CR value above 0.6 is considered acceptable in an exploratory study [51]. Moreover, the AVE of both sub-constructs achieved the value of 0.5. This indicates that, on average, each sub-construct accounts for a minimum of 50% of the variance of its items. Thus, the validity of the items and the sub-construct is indisputable. The COM sub-construct also holds two items with factor loadings lower than 0.7, i.e. COM-3 (0.68) and COM-5 (0.67). However, removing one of these items led to a fall in sub-construct's convergent validity (decreasing the AVE). Also, COM has other strength statistics, i.e. CR value of 0.83. Accordingly, the items were retained. For the same reason, the items with factor loadings below 0.7 in TCM, SII, and SIC sub-constructs were retained.

VIII. Conclusion

The Hi-ACT which that evaluates undergraduates' perceptions of their CT competency was developed. A pilot study was carried out to refine the initial instrument. Based on the responses of 548 university undergraduates to 155 items, an instrument comprising 114 items was established. The findings of statistical test of internal consistency and convergent validity reveal that the Hi-ACT in its pilot form is valid and reliable to measure university undergraduates' CT competency. In future studies, we plan to proceed with further instrument evaluation to provide further evidence of construct validity and discriminant validity.

Furthermore, the Hi-ACT makes a notable contribution to CT literature. It extends the CT assessment study by verifying ten primary constructs and 29 sub-constructs, which delineate the skills and attitudes applicable in CT-based problem solving process. These CT concepts did not comprehensively address in most previous CT assessment studies. Accordingly, findings of this work bring forth comprehensiveness to CT theoretical work, specifically in undergraduate context. It also results in a set of indicators that useful in measuring CT competency holistically.

References

[1] J. M. Wing, "Computational thinking," Commun. ACM, vol. 49, no. 3, pp. 33–35, 2006.

[2] D. Barr, J. Harrison, and L. Conery, "Computational thinking: A digital age skill for everyone," Learn. Lead. with Technol., vol. 38, no. 6, pp. 20–23, 2011.

[3] M. G. Voskoglou and S. Buckley, "Problem solving and computers in a learning environment," Egypt. Comput. Sci. J., vol. 36, no. 4, pp. 28–46, 2012.

[4] K. Yevseyeva and M. Towhidnejad, "Work in progress: Teaching computational thinking in middle and high school," in 2012 Frontiers in Education Conference Proceedings, 2012, pp. 1–2.

[5] World Economic Forum, "The future of jobs report 2018," Geneva, Switzerland, 2018.

[6] N. Talib, S. F. M. Yasin, and K. M. Mohd, "Teaching and Learning Computer Programming Using Gamification and Observation through Action Research," Int. J. Acad. Res. Progress. Educ. Dev., vol. 6, no. 3, pp. 1–11, 2017.

[7] F. Kalelioğlu, Y. Gülbahar, and V. Kukul, "A Framework for Computational Thinking Based on a Systematic Research Review," Balt. J . Mod. Comput., vol. 4, no. 3, pp. 583–596, 2016.

[8] L. Shanmugam, S. F. Yassin, and F. Khalid, "Enhancing Students ' Motivation to Learn Computational Thinking through Mobile Application Development Module (M-CT)," Int. J. Eng. Adv. Technol., vol. 8, no. 5, pp. 1293–1303, 2019.

[9] C. K. Chang, Y. T. Tsai, and Y. L. Chin, "A Visualization Tool to Support Analyzing and Evaluating Scratch Projects," in Proceedings - 6th IIAI International Congress on Advanced Applied Informatics 2017, 2017, pp. 498–502.

[10] Ö. Korkmaz, R. Cakir, and M. Y. Ozden, "A validity and reliability study of the computational thinking scales (CTS)," Comput. Human Behav., vol. 72, pp. 558–569, 2017.

[11] M. Yagci, "A valid and reliable tool for examining computational thinking skills," Educ. Inf. Technol., vol. 24, no. 1, pp. 929–951, 2018.

[12] M. Román-González, J. C. Perez-Gonzelez, and C. Jimenez-Fernandez, "Which cognitive abilities underlie computational thinking? Criterion validity of the Computational Thinking Test," Comput. Human Behav., vol. 72, no. July 2017, pp. 678–691, 2016.

[13] M. Bower, L. N. Wood, J. W. M. Lai, C. Howe, and R. Lister, "Improving the computational thinking pedagogical capabilities of school teachers," Aust. J. Teach. Educ., vol. 42, no. 3, pp. 53–72, 2017.

[14] S. Gretter and A. Yadav, "Computational thinking and media & information literacy: An integrated approach to teaching twenty-first century skills," TechTrends, vol. 60, no. 5, pp. 510–516, 2016.

[15] T. Rahayu and K. Osman, "Knowledge Level and Self-confidence On The Computational Thinking Skills Among Science Teacher Candidates," J. Ilm. Pendidik. Fis., vol. 08, no. April, pp. 117–126, 2019.

[16] S. Senin and N. M. Nasri, "Teachers ' Concern towards Applying Computational Thinking Skills in Teaching and Learning Teachers ' Concern towards Applying Computational Thinking Skills in Teaching and Learning," Int. J. Acad. Res. Bus. Soc. Sci., vol. 9, no. 1, pp. 296–310, 2019.

[17] J. Walden, M. Doyle, R. Garns, and Z. Hart, "An informatics perspective on computational thinking," in Proceedings of the 18th ACM Conference on Innovation and Technology in Computer Science Education., 2013, pp. 4–9.

[18] L. Gouws, K. Bradshaw, and P. Wentworth, "First year student performance in a test for computational thinking," in Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference, 2013, pp. 271–277.

[19] M. Csernoch, P. Biró, J. Máth, and K. Abari, “Testing algorithmic skills in traditional and non-traditional programming environments,” Informatics Educ., vol. 14, no. 2, pp. 175–197, 2015.

[20] D. E. Sondakh, K. Osman, and S. Zainudin, “Holistic Assessment of Computational Thinking for Undergraduate : Reliability and Convergent Validity,” in Proceedings of The 11th International Conference on Education Technology and Computers, 2019, pp. 241–245.

[21] M. Guzdial, “Education paving the way for computational thinking,” Commun. ACM, vol. 51, no. 8, p. 25, 2008.

[22] P. J. Denning, “Computing is a natural science,” Commun. ACM, vol. 50, no. 7, p. 13, 2007.

[23] B. Czerkawski and E. Lyman III, “Exploring issues about computational thinking in higher education,” TechTrends, vol. 59, no. 2, pp. 57–65, 2015.

[24] A. Yadav, J. Good, J. Voogt, and P. Fisser, “Computational thinking as an emerging competence domain,” in Competence-based Vocational and Professional Education Technical and Vocational Education and Training: Issues, Concerns and Prospects, M. Mulder, Ed. Springer, 2017, pp. 1051–1067.

[25] A. V. Aho, “Computation and computational thinking,” Comput. J., vol. 55, no. 7, pp. 833–835, 2012.

[26] P. J. Denning, “The profession of IT beyond computational thinking,” Commun. ACM, vol. 52, no. 6, p. 28, 2009.

[27] S. Grover and R. Pea, “Computational thinking in K-12: A review of the state of the field,” Educ. Res., vol. 42, no. 1, pp. 38–43, 2013.

[28] J. M. Wing, “Research notebook: Computational thinking—what and why?,” The Link Magazine, 2011.

[29] J. M. Wing, “Computational thinking and thinking about computing,” Philos. Trans. R. Soc., vol. 366, pp. 3717–3725, 2008.

[30] C. C. Selby and J. Woollard, “Computational thinking: the developing definition,” 2013.

[31] C. Angeli et al., “A K-6 computational thinking curriculum framework: implications for teacher knowledge,” Educ. Technol. Soc., vol. 19, no. 3, pp. 47–57, 2016.

[32] S. Bocconi, A. Chioccariello, G. Dettori, A. Ferrari, and K. Engelhardt, “Developing Computational Thinking in Compulsory Education-Implications for policy and practice,” 2016.

[33] J. Mueller, D. Beckett, E. Hennessey, and H. Shodiev, “Assessing computational thinking across the curriculum,” in Emerging Research, Practice, and Policy on Computational Thinking, P. J. Rich and C. B. Hodges, Eds. Cham, Switzerland: Springer International Publishing, 2017, pp. 251–267.

[34] S. Hambrusch, C. Hoffmann, J. T. Korb, M. Haugan, and A. L. Hosking, “A multidisciplinary approach towards computational thinking for science majors,” in Proceedings of the 40th ACM Technical Symposium on Computer Science Education, 2009, vol. 41, pp. 183–187.

[35] C. Kazimoglu, M. Kiernan, L. Bacon, and L. MacKinnon, “Learning programming at the computational thinking level via digital game-play,” Procedia Comput. Sci., vol. 9, pp. 522–531, 2012.

[36] C. E. Vergara et al., “Towards a framework for assessing computational competencies for engineering undergraduate students,” in 122nd ASEE Annual Conference & Exposition, 2015, pp. 26.1589.1-26.1589.14.

[37] Y. Allsop, “Assessing computational thinking process using a multiple evaluation approach,” Int. J. Child-Computer Interact., vol. 19, pp. 30–55, 2018.

[38] C. Varela, C. Rebollar, O. García, E. Bravo, and J. Bilbao, “Skills in computational thinking of engineering students of the first school year,” Heliyon, vol. 5, no. 11, 2019.

[39] P. Curzon, M. Dorling, T. Ng, C. Selby, and J. Woollard, “Developing computational thinking in the classroom: A framework.” p. 5, 2014.

[40] F. Kalelioglu, Y. Gulbahar, and V. Kukul, “A framework for computational thinking based on a systematic research review,” Balt. J. Mod. Comput., vol. 4, no. 3, pp. 583–596, 2016.

[41] M. M. Robles, “Executive perceptions of the top 10 soft skills needed in today’s workplace,” Bus. Commun. Q., vol. 75, no. 4, pp. 453–465, 2012.

[42] ACM & IEEE Computer Society, “Computer Science Curricula 2013 Curriculum guidelines for undergraduate degree programs in Computer Science,” 2013.

[43] D. A. Sisk, “Spiritual intelligence: developing higher consciousness revisited,” Gift. Educ. Int., vol. 32, no. 3, pp. 194–208, 2016.

[44] M. Kadkhoda and H. Jahanic, “Problem-solving capacities of spiritual intelligence for artificial intelligence,” Procedia - Soc. Behav. Sci., vol. 32, no. 2012, pp. 170–175, 2012.

[45] J. L’Heureux, D. Boisvert, R. Cohen, and K. Sanghera, “IT problem solving: an implementation of computational thinking in information technology,” in Proceedings of the 13th Annual Conference on Information Technology Education, 2012, pp. 183–188.

[46] N. Kourmousi, V. Xythali, M. Theologitou, and V. Koutras, “Validity and reliability of the problem solving inventory (PSI) in a nationwide sample of Greek educators,” Soc. Sci., vol. 5, no. 2, p. 25, 2016.

[47] E. Britton, N. Simper, A. Leger, and J. Stephenson, “Assessing teamwork in undergraduate education: a measurement tool to evaluate individual teamwork skills,” Assess. Eval. High. Educ., vol. 42, no. 3, pp. 378–397, 2015.

[48] P. S. Strom and R. D. Strom, “Teamwork skills assessment for cooperative learning,” Educ. Res. Eval., vol. 17, no. 4, pp. 233–251, 2011.

[49] D. E. Sondakh, K. Osman, and S. Zainudin, “A Proposal for Holistic Assessment of Computational Thinking for Undergraduate: Content Validity,” Eur. J. Educ. Res., vol. 9, no. 1, pp. 33–50, 2020.

[50] M. Norris and L. Lecavalier, “Evaluating the use of exploratory factor analysis in developmental disability psychological research,” J. Autism Dev. Disord., vol. 40, no. 1, pp. 8–20, 2010.

[51] J. F. Hair, G. T. M. Hult, C. M. Ringle, and M. Sarstedt, A primer on partial least squares structural equation modeling (PLS-SEM), 2nd ed. Sage Publications, Inc., 2017.

[52] N. K. Avkiran, “An in-depth discussion and illustration of partial least squares structural equation modeling in health care,” Health Care Manag. Sci., vol. 21, no. 3, pp. 401–408, 2018.

Optimize the Combination of Categorical Variable Encoding and Deep Learning Technique for the Problem of Prediction of Vietnamese Student Academic Performance

Do Thi Thu Hien[1], Cu Thi Thu Thuy[2], Tran Kim Anh[3], Dao The Son[4], Cu Nguyen Giap[5]*
Department of Information Technology, Thuongmai University, Hanoi, Vietnam[1]
Department of Econometrics, Academy of Finnance Institute, Hanoi, Vietnam[2]
Department of Economics, Thuongmai University, Hanoi, Vietnam[3, 4]
Department of Informatics, Thuongmai University, Hanoi, Vietnam[5]

***Abstract*—Deep learning techniques have been successfully applied in many technical fields such as computer vision and natural language processing, and recently researchers have paid much attention to the application of this technology in socio-economic problems including the student academic performance prediction (SAPP) problem. In this specialization, this study focusses on both designing an appropriate Deep learning model and handling categorical input variables. In fact, categorical data variables are quite popular in student academic performance prediction problem, and deep learning technique in particular or artificial neural network in general only work well with numerical data variables. Therefore, this study investigates the performance of the combination categorical encoding methods including label encoding, one-hot encoding and "learned" embedding encoding with deep learning techniques including Deep Dense neural network and Long short-term memory neural network for SAPP problem. In experiment, this study compared these proposed models with each other and with some prediction methods based on other machine learning algorithms at the same time. The results showed that the categorical data transformation method using the "learned" embedding encoding improved performance of the deep learning models, and its combination with long short-term memory network gave an outstanding result for the researched problem.**

Keywords*—Deep learning techinique; categorical data type; "learned" embedding encoding; student academic performance prediction

I. Introduction

Modern education system has been changed with a new facet distance education, also called distance learning [1]. Educational technology facilitates distance learning and creates the shift from traditional education model to new model that concerns to virtual community of learners [2]. The recent global disease pandemic also encourages the researchers investigate in design and implement new higher education model [3]. In educational technologies, information communication technology (ICT) is a fundamental infrastructure to supply a massive online courses to learners [1]. However, in this trend, the student need smart systems that support them during learning process. The student academic performance prediction (SAPP) plays the key role in such smart systems. Base on academic performance prediction, the systems give learners suitable advices, early warning and many useful instructions.

Utilization of deep learning techniques on student academic performance prediction becomes a desirable research with many achievements recently [4, 5]. Education management information system (EMIS) is popular and it supplies available data resource for researches on educational data mining using deep learning techniques. However, one challenge when applying deep learning techniques for SAPP problem is that the input data of this problem often contains many categorical variables that the deep learning techniques or artificial neural networks in general do not work well directly [6, 7]. Therefore, research on how to convert categorical data to numerical data for construction and training deep learning models in SAPP problem is necessary.

This study focuses on the analysis of categorical variable transformation methods and its compatibility with deep learning models simultaneously. More specifically, the methods of transforming the categorical data is not studied separately from the classification models. Instead, each categorical data encoding method can be adapted to a classification model using some design of deep learning network. Therefore, this study investigates how categorical data conversion is associated with the corresponding deep learning model. There are several deep learning network architectures that can be used to develop a classification model, however, in the scope of this study the focused models include the long short-term memory recurrent network and the Deep Dense network. These deep learning network models were evaluated and considered as great solutions for SAPP data sets in a number of studies [5, 8]. Besides, the conversion of categorical variable to numerical variable in classification problems take an important position, especially classifiers are built based on artificial neural network or deep learning techniques [9, 10]. The common methods such as label encoding, one-hot encoding and its modifications and new "learned" embedding encoding [11, 12] are interest and they are going to be estimated carefully. Analysis experimental

*Corresponding Author

result of the compounds of categorical data transforms method with deep learning models and compare to result of other machine learning algorithms gives important conclusions for solving SAPP problem. The main analyzed indicator is the accuracy of the prediction. Although this index does not reflect all facets of a predictive classifier, it is still a most popular indicator used in this research area.

This article is divided into five main sections. Besides the introduction, the remainder includes Section 2 which presents related researches. Section 3 describes the research methodology and design of the proposal deep learning models for the SAPP problem. Section 4 presents the results and the last part is conclusion and some future research recommendations.

II. Related Works

Utilization of deep learning techniques and machine learning techniques for SAPP problem has been concerned in many researches [4, 5, 13, 14]. Using these techniques can improve the prediction quality [5, 13, 14] and opens many applications in reality [4]. In which, the application of deep learning techniques is a research direction that has received great attention in recent times [5]. For example, in the study [15], the authors proved that the Deep Dense neural network improve the accuracy of failure-prone student prediction. The convolutional neural network was investigated for the same researching area in [16]. These deep learning techniques were utilized for predicting student final performance in [8] and for predicting students' future development in [17]. The results showed that proposed deep learning models worked well in many interesting cases of educational data analysis.

More specifically, the deep learning model built on Deep Dense network, a multilayer perceptron network architecture, was proposed for the SAPP problem [4, 15, 18, 19]. The authors announced the good performance of Deep Dense network and also compared the proposed models with other algorithms such as the decision tree algorithm C4.5, random forest, logistic regression and support vector machine. Deep learning models for SAPP problem based on the Long short-term memory network and convolutional neural network were introduced in [5, 20, 21]. The results showed that the proposed deep learning models have superior results compared to other tested algorithms.

When applying neural networks or deep learning techniques to the problem of classification, one of the problems need to be solved is transforming the categorical data type into numerical data type [22]. There are different transformation methods that are commonly applied, in [22] the authors divided these data transformation methods into three groups: predetermined transformation methods, algorithmic methods and automatic transformation methods. These methods have a certain interference, but the first method focuses on the laws of clear change and often has very low complexity. The second method may give predetermined results, but the algorithmic method is geared towards more complex computational and processing methods. A third method uses machine learning techniques and one of them is neural networks to automatically mine data. Choosing the right method of transforming classified data is one of the challenges that need to be addressed when building a model for SAPP problem and classification problem in general. In the study [23], the researchers used the one-hot encoding conversion method. Although [23] did not evaluate the effectiveness of this transformation method, the results of the accuracy of the tested classification methods showed that this transformation method has good applicability. Besides, the "learned" embedding encoding method introduced in [24, 25] can also be applied. Research [26] was an example showing the suitability when applying the "learned" embedding encoding in categorical variable conversion.

Although there are many studies related to the application of deep learning techniques in the SAPP problem and processing of classification data in neural network, but these two groups of studies are quite independent with each other. The researching approach that directly combines two issues in one optimization process for the problem of SAPP will help improve the quality of prediction. Naturally, this study will focus on this approach.

III. Research Methodology

A. Categorical Variable Encoding

When processing input data of categorical data type for an artificial neural network, it is likely to convert a categorical variable to a numerical variable or a vector that its elements are numerical data type. Three common methods used are: 1) Label Encoding; 2) One-hot Encoding and its modification; 3) "Learned" Embedding encoding. In the Label encoding method each label of a categorical data variable is assigned to a most suitable integer number. This method is very simple, however, finding the right assignment for a specific problem is relatively difficult (especially with categorical variables representing unordered data). The second method, One-hot encoding, turns a categorical variable's value into a binary vector where one element takes value 1 presents the appearance and remained elements get value 0 presents the absence. Element takes value 1 in the position equivalent to each classifier label specified in the encryption method. Both the processing methods mentioned above are relatively easy to implement, and the data transformation process is independent of the data processing model, classification model in this researching situation.

In the third processing method, the "learned" embedding encoding method, a classification data will be transformed into the distribution vector, which is associated with the training and optimization of an artificial neural network. A well-trained vector space will provide a projection in which labels in the classification data are represented by naturally close clusters. This encoding method is first proposed for the word processing problem in natural language processing. Later, this method was used in the analysis of classification data, especially when used with traditional artificial neural network models or deep learning models.

The "learned" embedding encoding method is a learning method, therefore it requires a suitable training data. Besides, this method is an algorithm with several parameters, so the study of parameter optimization is also a problem to solve. The most important of these parameters is the size of the output-

dimension vector. The studies related to parameter optimization for "learned" embedding encoding are mainly done with word processing problem. In this study, the output-dimension vector is calculated by a recommended formula in many studies, as follows:

$$Emb\_size = \min(50, (n\_cat/2)+1) \quad (1)$$

Emb_size: is size of output dimension of an embedding layer.

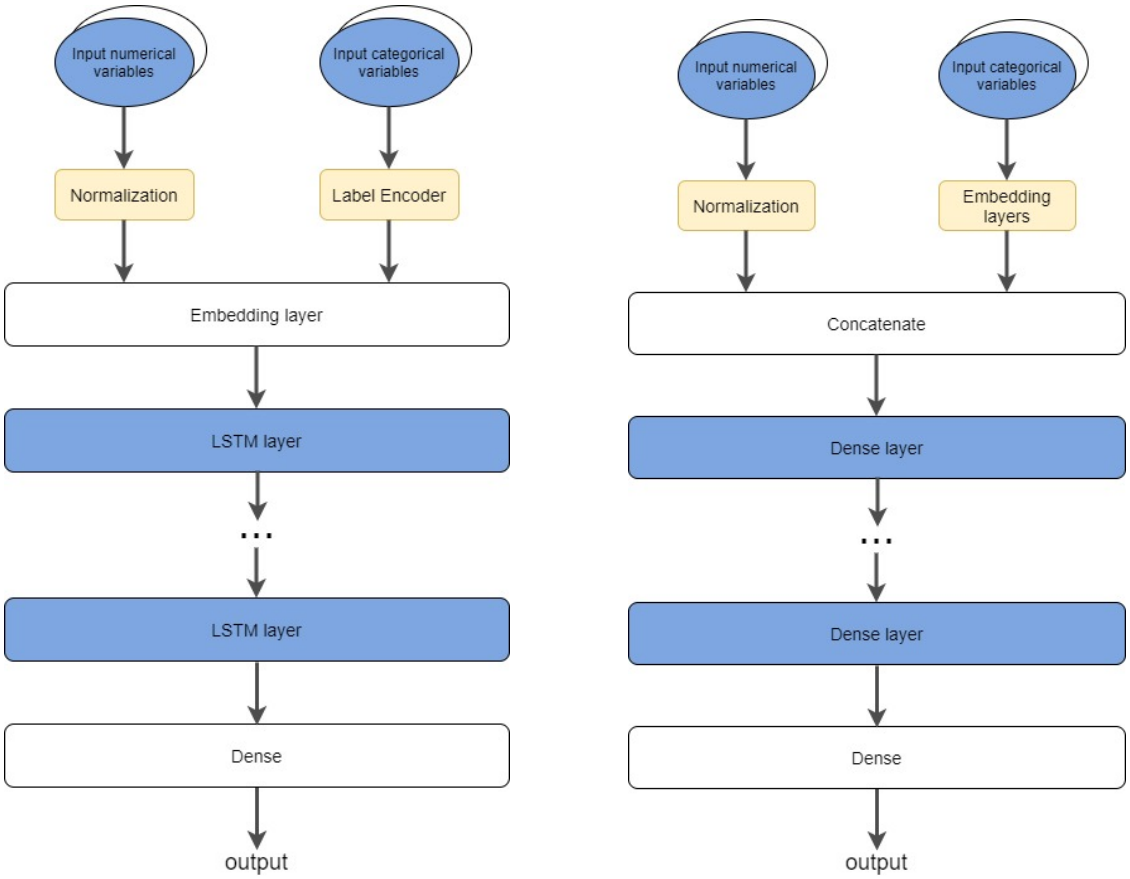
n_cat: is the force of the categorical variable.

B. Proposal Deep Learning Models

Combining methods of processing variables of classification data and deep learning techniques, two deep learning models are proposed for the SAPP problem. In which, a model uses Deep Dense network architecture (Fig. 1a), and one uses a Long short-term memory recurrent network architecture (Fig. 1b).

In particular, with LSTM network model, input categorical variables are converted by label encoding first. This data is combined with input numerical variables that are standardized to make input data for the LSTM network. The requirement of input data of a LSTM network is a 3-dimensional data type, so the input data is processed by an embedding layer first. The number of hidden LSTM layers added follow embedding layer and the number of hidden nodes of each layer can vary by designer. Each LSTM layer followed by a dropout layer to improve network performance based on reducing effect of the overfit problem. The original output variable is a categorical data type, and it is converted to a set of binary variables by One-hot encoding. Therefore, the output layer is a dense layer with the number of nodes equals to the number of output variables.

In the second model, each categorical variable is handled by a separate embedding layer. Then, all output of these embedding classes is aggregated with the normalized numerical variables by a concatenate layer. The output of the concatenate layer is the input to the dense network with the number of hidden layers depending on the design. Each Dense layer followed by a dropout layer too. The output of the problem is a set of binary variables, so the second model uses a dense layer as an output layer similar to the first model. Besides embedding encoding approach, two other encoding methods including label encoding and one-hot encoding are also experimented with Deep Dense network in the same structure.



(a) Long Short-Term Memory Network. (b) Deep Dense Network.

Fig. 1. Proposal Deep Learning Models.

IV. Experimental Result

A. Data Collection

Data of student academic performance was collected from Vietnamese universities, focused on students who followed disciplines related to economics. These data are extracted under the admission of grant number B2019-TMA-2 supported by Vietnamese Ministry of Education and Training. The data included information about student's learning capability at the entry time and two first years at university and final grade. Homology related information was collected but it was removed due to privacy concerns.

The experimental data includes 41 input attributes and one classification output, these attributes are described in Table I. There are 36 categorical input attributes, it takes 87.80% of input data. It means that categorical variable encoding is an important task for this problem. The remaining input attributes are 5 numerical variables. The data includes 524 observations, in which the ratio between normal classes such as pass and good grades and rare classes such as fail and distinction grades is quite large. This data is imbalanced data and it makes predictive classification problem become harder.

B. Experimental Result

The deep learning models proposed in Section III was implemented by Python programming language and this program used the Tensor Flow and Keras libraries. The experiment was run on a computer has following configuration: Core i7 2.0GHz, 16GB RAM, and 2GB GPU. Other machine learning algorithms its results were used to compare with the proposed models were implemented in Weka, an instance machine learning tool.

The method of evaluating the forecasting models is based on accuracy indicator. According to the accuracy measurement, is calculated by the formula: number of correct classified observations/ total number of observations, and, average accuracy is assessed by 10 times running on 10 testing data sets randomly spliced from collected data.

In the scope of this study, there are several parameters of deep learning models fixed. That are the number of hidden layers of the neural network is ignited at 2. The training process uses cell training with batch size = 100, Adam optimization method with learning rate parameter= 0.001. The fraction of the input units to drop is set at 0.15 at each dropout layer. Training accuracy of proposed models with number of interactions epochs=1000, is shown in the following figure.

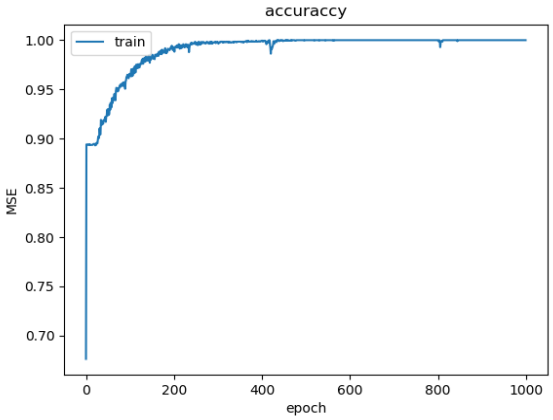
In Fig. 2, it can be seen that the training processes of all proposed deep learning models have a stable convergence and its reach expected error within 1000 epochs. For example, in Fig. 1, the Deep Dense networks converged to highest training accuracy at 100%. Besides, the Deep Dense network combines with "learned" embedding encoding for categorical variables converged faster than other testing deep learning models.

Base on convergence of the accuracy during training process, the proposed Deep learning models are continuously experimented with the same setting above, but the number of epochs is settled at 500. This value ensures that the experiment can reach expected correctness and it saves time also. The experimental result is calculated on the test data, which were generated by randomly selecting 25% of the observations from the original data set.

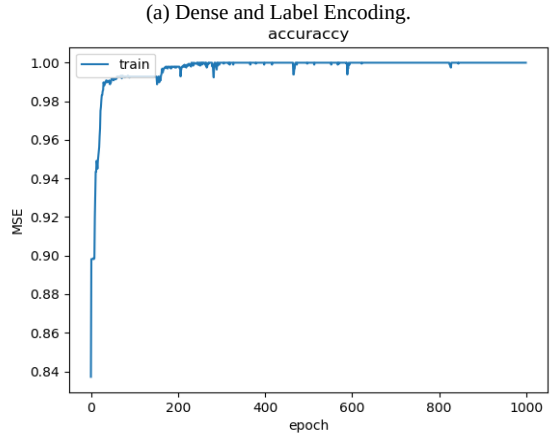
Optimization of number of hidden layers and number of hidden layers nodes are necessary for deep learning model design. However, there are not a thumb rule for setting these values in all situations. The popular approach lays down in experimentation. This study used this approach, specifically, the number of candidates for good setting was retrieved by literature review, and then these candidates were investigated by experiment to seek the optimum solution among candidates.

TABLE I. Description of Input Variables

Variable Name	Range	Description
gap_year	[0-3]	Elapsed time between high school graduation and university entrance.
Mark1, Mark2, Mark3	[0-10]	Individual marks of entrance exam
Total_mark	[0-30]	Total mark of admission
Group_Sub	1-3	Student's group
Gen_Sub_i, i=[1-22]	A, B, C, D, F	Scores of general subjects
$Spec\_Ground\_Sub_i$, i= [1-7]	A, B, C, D, F	Scores of specialized round subjects/modules
$Spec\_Sub_i$, i= [1-6]	A, B, C, D, F	Scores of other specialized subjects/modules



(a) Dense and Label Encoding.



(b) Dense and "Learned" Embedding Encoding.

Fig. 2. Training Convergence.

Based on the literature review and experiment the optimum number of hidden layers in all proposed deep learning models is 2. The average accuracy was not improved when the number of hidden layers was increased to 3 and 4, but the training time increased. Therefore, all deep learning models had 2 hidden layers in on going test.

Fig. 3 presents the performance of Deep Dense network combines with label encoding method or categorical data type. It can be seen that two values of number of hidden nodes give better performances than others. In this case, the best number of hidden nodes is chosen at 100, even though it give average accuracy 83.359% that is very close to result of the selection of 300 hidden nodes. Because the smaller number of hidden nodes leads to faster training process.

According to deep experimentation, the optimum number of hidden layer nodes in Deep Dense network with both Label encoding and One-hot encoding methods is 100, while the optimum setting for Deep Dense network with "learned" embedding encoding is 150, while the optimum number of hidden layer nodes in LSTM Network with "learned" embedding encoding is 100.

The performances of deep learning models with the optimum parameters are depicted in Fig. 4. The accuracy is measured by the average results of 15 run times with the experimental appropriate set of parameters mentioned above. It can be seen in Fig. 4, the box plot chart showed the LSTM-Network with "learned" embedding encoding method had the best performance. This network has highest average accuracy that was 86.26%, and this network was also more stable presented by median, lower quartile and upper quartile lines. Moreover, the box plot of LSTM model showed that most of running test often gave testing accuracy closes to maximum value.

In the same point of view, the embedding encoding method is also good to combine with Deep Dense network, it helps this network architecture works better than other encoding method. In tested SAPP problem, Deep Dense network with Label encoding method was better than using this network architecture with One-hot encoding method. The reason seems to be that most categorical variables in input data are ordinal data type.

In general, the performance of all testing algorithms is showed in Fig. 5. Experimentation proves that for student performance prediction the deep learning models combined with the embedding encoding method called "learned embedding" encoding for categorical variables gave good results. Deep Dense models with label encoding and one-hot encoding methods have competitive results with the best testing machine learning algorithms (SMO and random forest). However, Deep Dense network and LSTM network with "learned" embedding encoding have out performance.

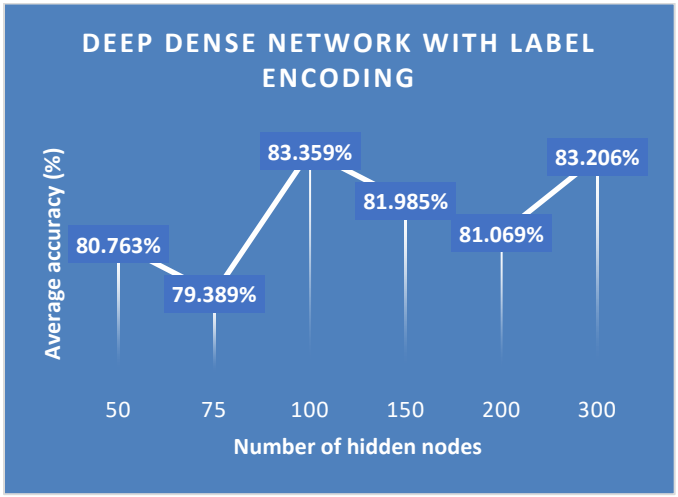


Fig. 3. The Performance of Deep Dense Network with Label Encoding.

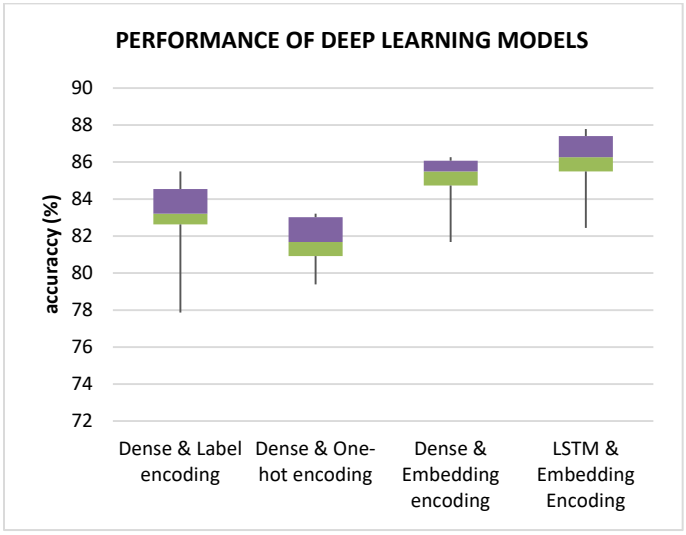


Fig. 4. Performances of Proposed Deep Learning Models.

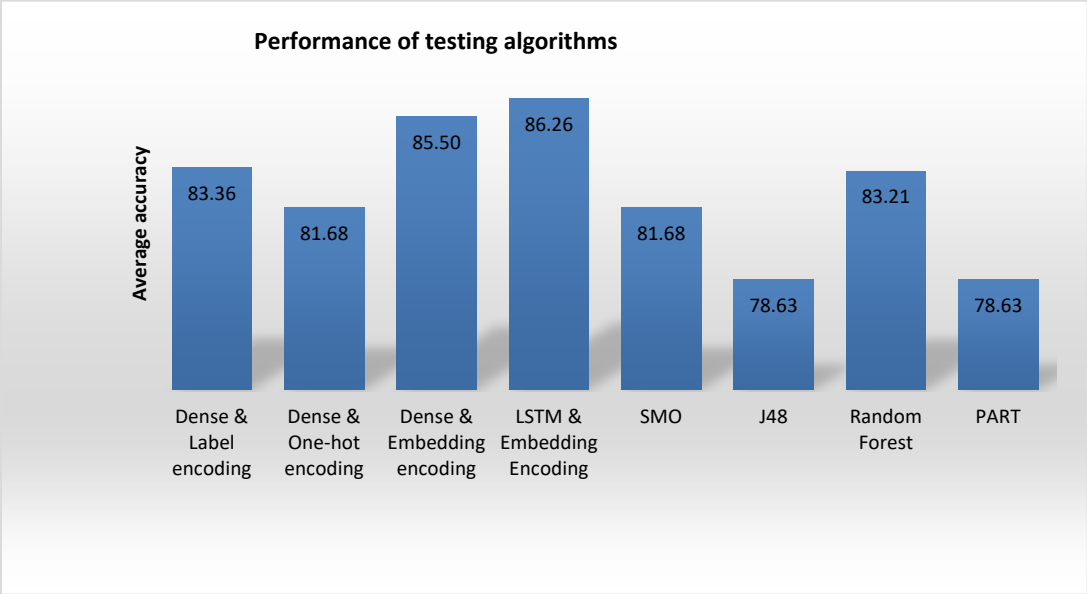


Fig. 5. Performance of All Testing Algorithms.

V. Conclusion

This study is an empirical study to solve the problem of student performance prediction with deep learning techniques, in which the construction of a deep learning model is studied coincidentally with categorical variables processing. Three methods of converting categorical variables to numeric variables including label encoding, One-hot encoding, and the embedding encoding methods were studied. Meanwhile, Deep Dense network and Long-short term memory network architectures designed in accordance with these encoding methods.

The experimental results give good insights and demonstrated the effectiveness of the "learned" embedding encoding method, because this encoding method has ability to learning simultaneously with the neural network in training process. In which using this data transform method with LSTM deep learning architecture gave the best result between other testing methods. This model has average accuracy at 86.26%. Embedding encoding method also improves performance of Deep Dense network also, and it also point out that to solve the prediction problems, which have categorical input variables, a deep learning model need to be designed with categorical variable encoding simultaneously.

The study also has a limitation that it takes time for experiment to find the optimal set of parameters in deep learning model design. Therefore, in this study, some parameters are selected according to recommendations. For example, the commonly used activate function is the "sigmoid" function, the recommended number of hidden layers is two. In the future, this study is going to be extended to solve these limitation.

Acknowledgment

This study was supported by Vietnamese Ministry of Education and Training and Thuongmai University under grant number B2019-TMA-02.

References

[1] Kaplan, A. M., & Haenlein, M. (2016). Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster. Business Horizons, 59(4), 441-450.

[2] Garrison, D. R. (2011, 20 May). E-learning in the 21stcentury: A framework for research and practice. New York: Taylor & Francis. ISBN:0-203-83876-9.

[3] Ilmiyah, S., & Setiawan, A. R. (2020). Students' Worksheet for Distance Learning Based on Scientific Literacy in the Topic Coronavirus Disease 2019 (COVID-19). EdArXiv, 7 Apr. 2020. Web

[4] Muniasamy, A., & Alasiry, A. (2020). Deep Learning: The Impact on Future eLearning. International Journal of Emerging Technologies in Learning (iJET), 15(01), 188-199.

[5] Doleck, T., Lemay, D. J., Basnet, R. B., & Bazelais, P. (2020). Predictive analytics in education: A comparison of deep learning frameworks. Education and Information Technologies, 25(3), 1951-1963.

[6] Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. Journal of Big Data, 7, 1-41.

[7] Zhang W., Du T., Wang J. (2016) Deep Learning over Multi-field Categorical Data. In: Ferro N. et al. (eds) Advances in Information Retrieval. ECIR 2016. Lecture Notes in Computer Science, vol 9626. Springer, Cham.

[8] Aydoğdu, Ş. (2020). Predicting student final performance using artificial neural networks in online learning environments. Education and Information Technologies, 25(3), 1913-1927.

[9] Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. Journal of Big Data, 7, 1-41.

[10] Zhang W., Du T., Wang J. (2016) Deep Learning over Multi-field Categorical Data. In: Ferro N. et al. (eds) Advances in Information Retrieval. ECIR 2016. Lecture Notes in Computer Science, vol 9626. Springer, Cham.

[11] Chen T, Tang L-A, Sun Y, Chen Z, Zhang K. Entity embedding-based anomaly detection for heterogeneous categorical events. 2016. arXiv:1608.07502.

[12] Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: a survey. Knowl Based Syst. 2018;151:78–94.

[13] Manjarres, A. V., Sandoval, L. G. M., & Suárez, M. S. (2018). Data mining techniques applied in educational environments: Literature review. Digital Education Review, (33), 235-266.

[14] Tran, T. O., Dang, H. T., Dinh, V. T., & Phan, X. H. (2017). Performance prediction for students: a multi-strategy approach. Cybernetics and Information Technologies, 17(2), 164-182.

[15] Kostopoulos, G., Tsiakmaki, M., Kotsiantis, S., & Ragos, O. (2020). Deep Dense Neural Network for Early Prediction of Failure-Prone Students. In Machine Learning Paradigms (pp. 291-306). Springer, Cham.

[16] Akour, M., Al, S. H., & Al Qasem, O. (2020). The effectiveness of using deep learning algorithms in predicting students achievements. Indonesian J. Elect. Eng. Comput. Sci., 19(1), 387-393.

[17] Fok, W. W., He, Y. S., Yeung, H. A., Law, K. Y., Cheung, K. H., Ai, Y. Y., & Ho, P. (2018, May). Prediction model for students' future development by deep learning and tensorflow artificial intelligence engine. In 2018 4th international conference on information management (ICIM) (pp. 103-106). IEEE.

[18] Ha, D. T., Loan, P. T. T., Giap, C. N., & Huong, N. T. L. (2020). An Empirical Study for Student Academic Performance Prediction Using Machine Learning Techniques. International Journal of Computer Science and Information Security (IJCSIS), 18(3).

[19] Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior, 104, 106189.

[20] Kim, B. H., Vizitei, E., & Ganapathi, V. (2018). GritNet: Student performance prediction with deep learning. arXiv preprint arXiv:1804.07405.

[21] Akour, M., Al, S. H., & Al Qasem, O. (2020). The effectiveness of using deep learning algorithms in predicting students achievements. Indonesian J. Elect. Eng. Comput. Sci., 19(1), 387-393.

[22] Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. Journal of Big Data, 7, 1-41.

[23] Wen, H., & Huang, F. (2020, May). Personal Loan Fraud Detection Based on Hybrid Supervised and Unsupervised Learning. In 2020 5th IEEE International Conference on Big Data Analytics (ICBDA) (pp. 339-343). IEEE.

[24] Cheng G, Berkhahn F. Entity embeddings of categorical variables. CoRR. 2016. arXiv:1604.06737.

[25] Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: a survey. Knowl Based Syst. 2018;151:78–94.

[26] Chen T, Tang L-A, Sun Y, Chen Z, Zhang K. Entity embedding-based anomaly detection for heterogeneous categori-cal events. 2016. arXiv :1608.07502

Conceptual Model for Connected Vehicles Safety and Security using Big Data Analytics

Noor Afiza Mat Razali[1]
Nuraini Shamsaimon[2], Muslihah Wook[3]
Faculty of Defence Science and Technology
National Defense University of Malaysia
Kuala Lumpur, Malaysia

Khairul Khalil Ishak[4]
Center of Cybersecurity and Big Data
Management and Science University
Shah Alam, Selangor

***Abstract*—The capability of Connected Vehicles (CVs) connecting to a nearby vehicle, surrounding infrastructure and cyberspace presents a high risk in the aspect of safety and security of the CV and others. Data volume generated from the sensors and infrastructure in CVs environment are enormous. Thus, CVs implementations require a real-time big data processing and analytics to detect any anomaly in the CVs's environment which are physical layer, network layer and application layer. CVs are exposed to various vulnerabilities associated with exploitations or malfunctions of the components in each layer that could result in various safety and security event such as congestion and collision. The safety and security risks added an extra layer of required protection for the CVs implementation that need to be studied and refined. To address this gap, this research aims to determine the basic components of safety and security for CVs implementation and propose a conceptual model for safety and security in CVs by applying the machine learning and deep learning techniques. The proposed model is highly correlated to safety and security and could be applied in congestion and collision prediction.**

***Keywords*—Connected vehicles; safety and security monitoring; collision prediction; congestion prediction; machine learning; deep learning**

I. Introduction

Connected Vehicles (CVs) is becoming more relevant in recent years after the realization of Industrial Revolution 4.0 (IR4.0), especially for the implementations in the smart cities and the Intelligent Transportation System (ITS) [1], [2]. CVs introduces a new concept of Vehicle to Vehicle (V2V), Vehicle to Infrastructure (V2I) and Vehicle to Everything (V2X) concepts that has high potential to become a destructive technology that could change on how communities commute and will impact the economic landscape of logistic and transportation industries [3], [4]. Safety and security of CVs are one of the hot topics in research that had gained research attention. The basic characteristic for CVs is its ability to connect to other vehicles, to surrounding infrastructure and the internet through sensors such as Laser Detection and Ranging (LiDAR), Radio Detection and Ranging (Radar), Global Positioning System (GPS), Dedicated Short Range Communication (DSRC), Radio Frequency Identification (RFID), Advance Driver Assistance System (ADAS) and sensors that are embedded in the vehicles itself [5], [6]. Communications devices are either embedded in the vehicle or connected to the vehicle's power socket. Due to the capability of CVs connecting to a nearby vehicle, infrastructure and cyberspace, it presents a high risk in the aspect of safety and security of the CVs and others, particularly if any vulnerability is exploited or malfunction occurred in any of the sensors.

CVs implementations will also require real-time big data processing and analytics, for instance, to detect any anomaly for the CV's network communication or employing CV sensors for collision prediction. This will add an extra layer of required protection for the safety and security of CV implementation. Acknowledging the importance of CV's safety and security, countries, such as United States of America (USA) had introduced the framework for CVs deployment in the country. The framework focuses on the deployment of CV in regards to privacy, cybersecurity, safety regulation, ethical issues and more. Information-Technology Promotion Agency, Japan, has published a Vehicle Information Security Guide [7]. This guide presents potential threats faced by automotive systems and security measures against those threats, aiming at helping automotive system developers improve their security design.

Various studies [4], [6], [8] has discussed regarding the emergence and implementation of connected and autonomous vehicles, to improve the driving experience and reduce the risk of a crash, improve traffic control and provide real-time interactive communications between other vehicles as well as roadside infrastructure in a network.

This proves beneficial, as the future implementation of smart cities requires Intelligent Transportation Systems (ITS) to promote smart mobility in a city. But the implementation of a CV environment contracts several implications and challenges. Researchers also are discussing the improvement of traffic control to reduce traffic congestion.

Hence in this paper, the focus would be on discussing the safety and security of CVs based on physical, network and application layers. This paper proposes a conceptual model for safety and security in CVs by applying the machine learning and deep learning techniques. The proposed model is highly correlated to safety and security and could be applied in congestion and collision prediction.

II. Background

Many studies are being done to investigates the implementations, implications and challenges of the CVs in the

This work was supported under the National Defense University Malaysia Short Grants UPNM/2020/GPJP/ICT/4 and UPNM/2018/GPJP/2/TK/5

Smart Cities or as the main component in the Intelligent Transportation System (ITS) [6], [9], [10]. However, many studies are focusing more on the utilization of individual or combinations of technologies in CVs applications. The ecosystem as a whole or the development of a model for the implementation of CV technologies in a particular targeted environment that take safety and security into deep consideration is still a new field that needed to be explored [4], [6].

A. Safety and Security for Connected Vehicle

The main challenges of CV implementation would be safety and security. In a CV environment, where communications occur between vehicles and infrastructure in a network could open up various possibilities and threats for attacks and misdemeanor. Moreover, there is a lack of clear guidelines and requirements for the usage of CV. Safety focuses on the physical aspects of an accident failure, while security focuses on the failures caused by malicious attackers [11].

1) Safety: Safety focuses on the physical aspect of the CV implementations, that includes several main aspects such as Driver Safety, Vehicle Safety, Road Safety, and Traffic Safety based on studies in various literatures [12]–[15].

Based on Fig. 1, we are focusing on these four aspects as the main pillars of safety in CV, which includes (i) Driver safety; where the condition of the driver maneuvering and handling the car contributes to the care of their safety. (ii) Road safety; the condition of the road infrastructure that can contribute to the occurrence of accidents. (iii) Vehicle safety; the condition of the vehicle, through its ability to operate in a well-mannered condition, and low risk of vehicle failure from occurring. (iv) Traffic safety; where the condition of traffic signs, stoplights can contribute to the lack of information for the driver, regarding the possible warnings and hazards on the road.

Studies by [16]–[22] discussed that safety components in CV environments include road friction, accident and collision prediction, road user detection, cluster identification, route planning and image signing.

Fig. 2 shows the components of safety in the CV. Road friction and conditions are one of the factors that can lead to accidents or collisions from happening. For example, when the condition of the road is slippery due to weather condition, the risk for collision is slightly increased due to the reduced amount of friction from vehicle tires and the road. Accident or collision prediction is a feature where if the proposed model can estimate that an accident would occur, an alert or notification can be sent to alert and notify the driver for action to be taken. Road User Detection is to provide a wider range of information regarding the surrounding of CV. Image Sign Board is a feature to detect road signs along the driveway. As an example, a speed bump is detected, the CV can be alerted to slow down their vehicle. Cluster identification is a feature to determine the cluster head or leader of a certain cluster of CVs in a driveway. Cluster heads are responsible to efficiently communicate road information for their clusters, alert for any situations and are chosen based on their CV capability. Route Planning feature is for CVs to calculate all available routes which are safe and avoid any unwanted situation.

2) Security: As mentioned in [11], security is more leaning towards malicious attacks on the networks where it can be from external or internal, either with intentions or not. This can affect and interrupt the transmission of data and data analysis process. Fig. 3 shows the security principles in Information Security, where each of the principles are important aspects to ensure the security of operations in CV environment. (i) Availability; where only authorized user can reliably access to information they need. For example, a car owner is right to be assigned as an authorized user and respond to the CV information. (ii) Confidentiality; is to prevent unauthorized users from the disclosure of information or data for their benefit by limiting their access[23]. (iii) Integrity; means that data cannot be deleted, modified or added by an unauthorized user. (iv) Authentication; is about data or information that proves who you are. It is about username and password to identified user as a legal user. This authentication in CV is a key component to allow you to gain or responsible for all the action. (v) Authorization; when system approved a legal user that allows them to do the next action.

Studies by [23]–[25] discussed that security aspects in CV environment include cyberattack, intrusion detection and prevention and attack classification. Fig. 4 shows the aspects of safety in CV implementations. The list of security aspects where every of it has their task and operation to be executed in the CV environment to ensure that communication transmission in the network involving CV can be protected from any unwanted occurrences. Intrusion detection and prevention in a network are the possible types of intrusion that can be detected and prevented through the implementation of security measures. Meanwhile, attack classification is the method to classify types of attack which are possible to occur in a CV networking environment. Cyberattack is another unwanted type of attack that can breakdown networking infrastructures.

Fig. 1. Safety Aspects in CV.

Fig. 2. Safety in CV Implementations.

Fig. 3. Security Aspects in CV.

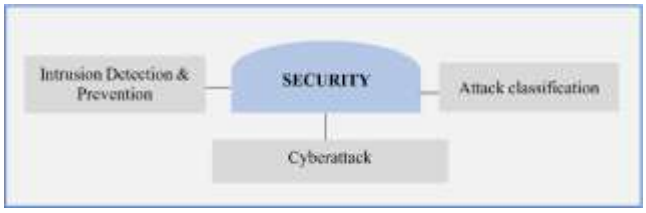


Fig. 4. Security in CV Implementations.

III. Components of Safety and Security in Connected Vehicles

In the CV environment, safety and security need to be considered in three layers perspective which consists (i) Physical Layer; (ii) Network Layer; (iii) Application Layer.

A. Physical Layer

The physical layer is including:

1) Road Side Infrastructure (RSI) and Road Side Units (RSU): an infrastructure-based communication meeting-point, where vehicles can communicate within the proximity of RSU installed in the nearby infrastructure. The higher number of RSU deployed in the infrastructure, the higher the capacity of vehicles able to communicate within the network. In a connected vehicle environment, RSU plays an important element in ensuring the reliability of both V2V and V2I communications [26], [27].

2) On-Board Unit (OBU): OBU and Onboard Sensor (OBS) is a fixed networking device in a vehicle, usually connected to the wireless network. The components of OBU typically consists of the GPS unit, human-machine interface unit, a wireless communication unit, and the central control unit. The central control unit is the main unit, consisting of data transceiver, serial port for information processing, memory, and decision and judgement making [28].

3) RSU and OBU Technologies: In recent years, there are rapid developments of the next-generation vehicle system that are based on connected vehicles (CVs) platform. CVs utilizes state of the art technologies such as Light Detection and Ranging (LiDAR), Global Positioning System (GPS), Dedicated Short Range Communication (DSRC), Advance Driver Assistance System (ADAS) and many more [5].

a) Dedicated Short-Range Communication (DSRC): The usage of DSRC protocol in the vehicular and automobiles industry has introduced innovations and technologies that utilize the DSRC communications service in vehicles to provide traffic safety and enhancement of mobility. DSRC provides dependable and quick information and data interchange for vehicle-based communications. Applications of DSRC help for the implementation of secure communication between RSU and OBU of a CV environment using V2I and V2V [29].

b) Advanced Driver Assistant (ADAS): The implementation of ADAS technologies is based on a vision/camera system, with several of researches have discussed regarding the implementation of camera-based ADAS [30], [31]. Combination of cameras with sensors for surround-view of safety vision detection applications such as to detect pedestrians and automatic braking functionalities have improved in recent years. The camera can be considered an essential part of the CV network to provide vision and imaging for the system.

c) Light Detection and Ranging (LiDAR): The usage of LiDAR is a remote sensing application able to create a three-dimensional (3D) representation of characteristics of vehicles in a CV environment. LiDAR technology uses light to estimate the parameters of a surface, in this case, vehicles. LiDAR is chosen compared to other mainstream sensor technologies due to its ability to acquire accurate calculations and measurements of the vehicle for its speed, type, and position. The data generated by LiDAR is highly accurate, due to its ability to cover an area with a view of 360 degrees, without depending on the light conditions [1], [32].

d) Global Positioning System (GPS): GPS is an essential unit in a CV network as it acts as one of the main units to accurately determine the position of a device/vehicle. GPS signals received from the satellite are interpreted and filtered by the GPS unit in the OBU. The recorded data are then taken to calculate the location, speed and the rate of change of speed of vehicles [28].

e) Traffic Light Controller (TLC): A TLC in a CV environment acts as a device that dictates instructions and provides a set of rules for drivers to rely on, for the main purpose of avoiding collisions, giving directions, and giving warnings. Several studies have proposed solutions on implementing fuzzy logic in TLCs, mainly to control traffic volume, hence reducing delays and increase data interchange. This proves to be beneficial in a CV environment, where maximum data throughput is recommended to maximize productivity [33]–[35].

f) Radio Frequency Identification (RFID): The implementation of RFID in a CV environment is mainly used for vehicle positioning purposes besides GPS. The OBU will be attached with an RFID tag, and the reader placed at the RSU. This creates an intra-vehicles sensor network. The usage of RFID in critical locations acts as a replacement solution, where GPS could not be utilized and used for positioning purposes [36], [37].

B. Network Layer

Communication in CV is one of important element and component to be implemented. CV technology can also further

increase the efficiency and reliability of autonomous vehicles, though these vehicles could be operated solely with their on-board sensors, without communication [38].

There are several types of communications in the connected vehicle, that are mainly split into three parts; vehicle-to-vehicle communication (V2V) and vehicle-to-infrastructure communications (V2I) and vehicle-to-anything communication (V2X). The capability of a connected vehicle, to communicate with other vehicles, and the infrastructure (RSU) reliably open up various possibilities, including to enhance the safety and security of operating vehicles.

V2V communications in connected vehicle environment have contributed to providing important reassurance on the improvement of operational safety in vehicles such as collision warning, as vehicles can communicate actively within the area through message exchange, primarily for accident prevention and warnings [28], [39]. V2I is also an essential part of the CV communication for the road-side units (RSU) to provide continuous connection and communication in the network [40]. V2X is an entire network communication where all infrastructures and vehicles are internetworking amongst each other to communicate and transfer information and data. Vehicular cybersecurity attacks include the shutdown of engines, tampering and disabling of brakes is an example of how an attack on the security of CV implementations could prove dangerous. Various articles have discussed regarding this issue and proposed various methods to detect and prevent these attacks [23], [24], [41], [42].

Fig. 3 shows the components which are involved in the security aspect. Security focuses on network security connection, between road users and the RSU/OBU in the CV environment. A study [23] states that the cybersecurity has its concerns to protect CVs, especially in network communication, to avoid threats or attack that can compromise CV functions. The threats can be done remotely, where data communication can be stolen, altered, and destroyed. Studies have found out the attributes of CV's cybersecurity as follows; (i) it is difficult to estimate all potential attack before it occurs. Attackers would only need to determine a vulnerability gap for them to infiltrate, while defenders are required to ensure all vulnerability gap is secured from potential attacks. Network attack and prevention have its challenges to ensure full security. (ii) there are a variety of connection medium in the network, that includes DSRC such as Wi-Fi, Bluetooth and others. (iii) CVs have different sensors and technologies such as LiDAR, where each sensor has its capabilities, functionalities and types of data which is compatible with CV. (iv) CV environment consists of various components and functions, where there is a possibility if one component dysfunctions, would affect the performance of the whole system. If the system is being attacked it can give bad impact and consequences to CV road users.

C. Application Layer

CV is an application that lies within the concept of Smart City in IoT, that consists of technologies such as smart transportation, smart parking, smart building and others where everything can communicate amongst each other in a network [43]. The smart city is technically an urban high-tech city, that enables people to improve the quality of life alongside technology. People would be able to utilize technology resources to further improve their daily life and expands the growth of an urban city in their country [44]. In a CV environment, varieties of transportation-related fields can be expanded, such as daily traffic monitoring, smart parking that enables users to conveniently locate the nearest parking spot available for their vehicle that will be using technologies and sensors such CCTV, LiDAR, mobile devices, GPS, accelerometers, gyroscope-based applications, weather sensor, ADAS, DSRC, TLC and RFID. All these technologies and devices for applications in CV contribute to the growth of IoT around the world.

D. Connected Road User

In a CV environment, the users of the network would be any pedestrians, cyclists, motorcyclists and other vehicle drivers that possess a personal mobile device such as tablets and smartphones with portable DSRC units. For a vehicle such as cars, buses, and lorries to be connected, they would require an On-board unit, such as a DSRC unit to receive real-time DSRC messages, as well as broadcasts their data and information to other vehicles and connected infrastructures in the CV environment. Through the connection of DSRC Unit to the personal device such as tablets is represented through an application, whereby the personal device would require connection to the DSRC Unit through Bluetooth. Once connected, the drivers would receive real-time traffic data, alert messages, collision warning and communication with other CVs in the network [4], [6], [39].

E. Big Data Analytics

Big data is a concept of data in which it is very massive, unorganized, unstructured data which could not be processed and analyzed by a traditional IT hardware and software in a considerable and tolerable amount of time [45], [46]. This is why techniques for data analytics that includes ML and DL have been introduced to provide a better solution for big data management. This opens up various opportunities in the advancement and development growth of technologies such as IoT in businesses and organizations. Big data analytics are also involved in the implementations of CV [47], [48], where a massive data collection and management occurs in a CV environment as a massive amount of vehicles are involved.

F. Data Analysis: Machine Learning and Deep Learning Techniques in Traffic Data Centre

The traffic data centre is a process of data analysis and determination of results for a certain condition based on collected data through the Machine Learning and Deep Learning techniques such as Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), Linear Regression (LR), Clustering, Artificial Neural Network (ANN), Convolutional Neural Network (CNN) and Fuzzy Logic (FL).

1) Decision Tree (DT): DT is known as a prediction tree that uses a tree structure for sequences and consequences decision specification. Constructing the DT with test points and branches can achieve the prediction. At each test point, a decision is made to choose a specific branch and cross down

the tree. Test points require testing for particular input variables and each of the branches represents the decision being made. In [23], the research has implemented DT in CV, to classify and detect connected autonomous vehicles in cyber-attack. DT is one of the most-used classification models with good readability. In [24], DT was implemented as an intrusion detection for CV in smart cities. The usage of DT is a feature selection and attack classification purposes. This paper concludes that the accuracy for detection rate is effective after considering the false positive and false negative rate generated using this method. From these two authors, it can be seen that DT can make a good classification for prediction, especially on detecting which nodes are malicious and non-malicious, to prevent cyber-attack, mainly in a CV environment.

2) Random Forest (RF): RF is a machine learning method that operates by constructing a multitude of decision trees by using classification, regressing and other tasks. This shows that RF is simplicity and diversity which can be used for more than 1 task. RF operating at training time and outputting the classes into classification or mean prediction into the regression of individual trees. RF is categorized in supervised learning algorithms. Several studies [49], [50] discuss using RF implemented in CV environments to classify supervised time-series based on driver behaviour and to classify vehicle recognition to differentiate among road users such as pedestrians, bicycles, cars and others. As a result, these two studies show that RF is effective to prevent overfitting for driver behaviour and successfully integrate the data and processing to identify and differentiate road user.

3) Naïve Bayes (NB): Classification method based on Bayes theorem in which it has the capabilities to provide the relationship between two event probability and their conditional probabilities. NB makes strong assumptions for presence or absence features of a class that are not related to other features independently. The NB is easily implemented and can be executed efficiently without prior knowledge of the data. In CV environments, [17] used NB to predict accidents and congestion before it happens. The author experimented by completing 10-fold cross-validation on their dataset. All accident severity types including minor, intermediate, major and NULL has experimented. As a result, NB is not better compared to Distributed Random Forest (DRF) but NB has fewer features to collect and make the decision quickly. In [51], NB was used to identify a driver's identity. The classification was made through a voting mechanism (VM) to analyze data in details of automotive characteristics based on a large number of sample data. Henceforth, it has successfully recognized 10 driver identification. In comparison, we can see that NB provides high accuracy for making the right decision and can be calculated with fewer data or samples quickly.

4) Support Vector Machine (SVM): SVM is a classification or regression algorithm that is categorized as a supervised machine learning algorithm and it is mostly used in problem classification. In SVM, each data item has placed as in n-dimensional space (N is number features) with the value of each feature. Specifically, it coordinates to find a hyper-plane where it can be separated into 2 classes. SVM is an extended version of a linear regression which can give high accuracy and give a simple decision boundary. SVM has limitations where it can separate only 2 classes. Usually, implementation of SVM is suitable for text classification, spam detection and computer version identification. The author in [16] researched about the implementation of a CV environment that can predict the friction class for specific road segments by using SVM, logistic regression (LG) and artificial neural networks (ANN) method. Another research in [22], used SVM with the Distance to Border features of the segment blobs for classification to detect and recognize traffic signs based on colour information in image sequences. The studies show that SVM gives good segmentation and classification results and can perform well in high dimensional spaces and the algorithm is very versatile and effective in cases where the number of dimensions is greater than the number of samples.

5) Linear Regression (LR): An analytical technique which is used in model relationships for 2 variables by setting a linear equation to observe the data. X represents an independent variable (the variable wants to explain or forecast) while Y is a dependent variable (explaining the other variables). The leverage of LR works with almost any kind of dataset and gives good information about the features. The opposite of LR is quite some assumptions than accurate decisions. Experimental use of LR in connected vehicles has been made by authors in [17] where they measure the accurate Estimated Time of Arrival (ETA). The reason researchers use this technique is to predict the clearance time after an accident occurred. The purpose of LR calculating the clearance time is to update ETA for indexed trips and giving the most accurate time. As a result, this novel predicts that using one of these methods will decrease the accident rate and give high accuracy and latency results.

6) Clustering and K-Mean Clustering: Clustering is an example of unsupervised classification technique in ML, where several similar data points, are divided into a group different from other groups of similar data points. In a CV environment, an example of the usage of clustering in [20] shows a weight-based clustering algorithm of vehicles in the same road segment to determine the primary cluster head (PCH) and secondary cluster head (SeCH). Another type of clustering is K-Mean Clustering which is a repetitive type of clustering, In [52], K-Mean Clustering was used to collect the journey time and volume data of several clusters, to identify the boundary value of each cluster. Clustering offers several solutions to CV environment implementations but would require a combination with other techniques such as Fuzzy Logic (FL) and DT to produce a reasonable outcome.

7) Artificial Neural Network (ANN): ANN is a collection or group of multiple neurons or perceptrons, where inputs are processed only in a forward-facing direction, that performs well when precise knowledge of a relationship requires some functional approximation. In a CV environment, authors in

[16], [53] implemented ANN in their experiment of studying the effect of the number of hidden layers in the performance of the wireless network. An interesting study in [54] has also implemented ANN to predict the level of severity of a vehicle's driver, at signalized intersections where an accident has occurred. ANN models and implementation have seen an uprise in the field of transportation, due to their adaptive capability and nature. Most implementations of ANN in research are to investigate its capabilities to enhance the wireless network performance, as well as other aspects such as the condition of the vehicle driver. This would prove beneficial for CV infrastructure safety applications.

8) Convolutional Neural Network (CNN): CNN is a neural network which is used productively for the classification and recognition. CNN is highly adept in areas like identification of objects and traffic signs, besides being able to generate vision on self-driving cars. In CV, application and usage of CNN can be seen in the literature [18], [19], [21], [55], where most CNN models are used for accident analysis and prevention, that are applied in some research to efficiently map crash risk, traffic conflicts and perception models for network traffic control. A software-defined network (SDN) model, SeDaTive [21] implements CNN model to provide data input to the model, where the CNN model studies the hidden patterns in data nodes, to plan the most optimal route for the model. This illustrates that using CNN for data classification helps to ensure effective network traffic control. This deep learning method can be considered for a data classifier in collision prevention model.

9) Fuzzy Logic (FL): FL is an AI method that bears a resemblance to human reasoning. The technique and process of FL emulate the way humans make decisions, involving all possibilities between the values of YES and NO. FL can be applied and used in automotive systems, including CV environment. Several studies [25], [56] discuss the use of FL techniques and algorithms, that includes, (i) using FL as a base algorithm for vehicle's decision making systems that can make decisions based on reasoning similar to human reasoning, and (ii) detection of attacks such as message injection through the classification and differentiation of injected malicious, fabricated and normal packets for the vehicle network. Based on these two authors and several other studies, it can be seen that fuzzy logic is an ML algorithm that fits well in-vehicle environment control, which would benefit the CV environment implementations.

10)ML and DL Techniques in CV implementation: Table I illustrates a summary of all works of literature with the techniques or methods of ML and DL related to the CV environment that we surveyed. Some examples of studies include; calculating the severity of injury based on accident impact, road friction estimation, collision prediction and avoidance and cluster recognition. From these studies, can be used as a reference for this study to determine the method or technique which is suitable for our model.

TABLE I. MACHINE LEARNING AND DEEP LEARNING TECHNIQUES USED IN CONNECTED VECHILE

Studied	***Techniques***	***Parameters***
[17]	LR, NB and Distributed RF	Distance and speed
[18]	Multi-layer perceptron (MLP), CNN and DT	Speed, acceleration, deceleration, wait time and wait ratio
[19]	Regional–Convolution Neural Network (R-CNN) and Deep Neural Network (DNN)	Speed ratio, driver characteristic and road/environmental condition
[21]	CNN	Bluetooth, Zigbee, DSRC, Wave, DSA, Wi-Fi, WiMAX and Cellular
[16]	Logistic Regression, SVM and ANN	Level friction segment (error rate, sensitivity and specificity)
[22]	SVM	Image Processing (color, shape, pattern recognition)
[51]	K-Nearest Neighbor and NB	Accelerator pedal, brake pedal, engine speed, steering wheel and vehicle speed
[49]	RF and RNN	Accelerations, lane changes, braking and aggressive and normal left and right turns
[50]	RF	Background filtering, object clustering, vehicle recognition, lane identification and vehicle tracking
[20]	Clustering	Number of neighbouring node, position, speed, road id, direction,
[52]	Fuzzy Logic	Volume, journey time, speed, and weather conditions
[56]	Fuzzy Logic	Speed and acceleration
[54]	ANN	Driver info, seat belt, vehicle type, speed, area type and point of impact

Table II and Table III illustrate the allotment of the ML and DL techniques which is supervised and unsupervised according to the review conducted in Table I. The classifications are also based on a study conducted by [43]. Some examples of supervised ML and DL methods are DT, RF, CNN and RNN. Supervised learning is where the techniques can perform learning on a dataset that is labelled in which the accuracy of the training data can be evaluated provided by an answer key for the algorithm to use. Meanwhile, the unsupervised model algorithm needs to process unlabeled data and learn to understand through features and pattern extraction. Some of the ML and DL techniques which are unsupervised are Clustering and FL.

TABLE II. APPROACH FOR ML METHODS

Machine Learning	
Supervised	**Unsupervised**
DT	Clustering
RF	FL
NB	Classification
SVM	Classification

TABLE III. APPROACH FOR DL METHODS

Deep Learning
Supervised
CNN
RNN

IV. CONCEPTUAL MODEL FOR SAFETY AND SECURITY IN CONNECTED VEHICLE USING BIG DATA ANALYTICS

Based on the literature that we surveyed in this study, we proposed the conceptual model for safety and security illustrated in Fig. 5. Each aspect has its variable to ensure each component functions correctly and efficiently. We propose a safety and security concept to ensure both aspects can give virtuous impact on functionality for CVs implementation. The variables and parameters listed are leaning more towards safety, which is considered the main focus. This is due to being cautious regarding the physical part of a CV, that includes the vehicle's and driver's safety, including the driver's condition, the level of driver's injury if accidents occur, as well as the driving style. Meanwhile, for security, existing technologies such as LiDAR, ADAS and others are used to secure the virtual part of the CV implementation, such as networking, CV communication and attack prevention. Parameters data and existing technologies will be analyzed using several ML and DL methods that include CNN, NB, SVM and others.

A. Conceptual Model for Safety and Security in Connected Vehicle

To address the safety and security for CVs environment, data analysis which is implemented in the application layer would be conducted in the other two layers which are both Physical and Network Layer need to be considered.

Fig. 5. Conceptual Model for Safety and Security in Connected Vehicle.

For security, data analysis in network layer determines to be the method in the communication transmission process. In CV, communication is important for the interaction between RSU and OBU(CV) for data collection and process to produce results of the possibility of congestion, which will be sent to other CVs for action. The transmission process is essential to maintain network security from attacks. If communication is interrupted, hence the possibility of obtaining results would be hindered, that can cause chaos due to inaccurate information.

For safety, data analysis in physical layer determines to be the method of accident and collision prevention, where the model would utilize the ML methods to predict any possible collision from occurring during the congestion. Based on the ML techniques, an alert notification message would be able to be transmitted for the CVs to take action and be prepared for the incoming congestion to avoid any possible collision.

1) Safety: Data Analysis in Physical Layer using Conventional Neural Network and Naïve Bayes: For safety aspect, CNN and Naïve Bayes are considered as two main selected techniques in this paper, based on [17]–[19]. NB, are used in [17]as they are reliable, and fast for collision prediction through the sending of alerts and notifications. In the CV environment, for real-time and minimum data collection, NB is a technique that suits well as it has a fast computation time and produces the right decision quickly. CNN is also selected, as it can analyze collision risks at intersections [18], as well as to detect vehicles and lane through image processing obtained from a single front-facing camera [19]. It is believed that the combinations of NB and CNN can provide an optimum alert or notification system for collision prediction in a CV environment.

2) Security: Data Analysis in Network Layer using Decision Tree and Fuzzy Logic Decision Tree and Fuzzy Logic would be considered as the two main techniques for security, based on [23], [24], [56] as it can classify and detect connected and autonomous vehicles in cyber-attack. DT is selected as a feature selection and attack classification purposes. DT can make an efficient classification for prediction, especially on detecting which nodes are malicious and non-malicious, to prevent cyber-attack, mainly in a CV environment. Fuzzy Logic algorithms help in detecting network attacks such as message injection through the classification and differentiation of injected malicious, fabricated and normal CAN packets. Based on these two techniques, the security of the communication in CV based on the proposed model can be improved. This is important in the model, where the RSU/OBU needs to ensure constant connection and communication with all the CVs in the environment. If the network is attacked, or infiltrated with unwanted and malicious nodes, the alert or notifications of collision would not be able to be transmitted, hence increasing the risk of collision and accidents from occurring.

B. Data Analysis using Machine and Deep Learning

In this study, in regards to safety and security for CV implementation, we are proposing the application of the model

in Collision Prediction Model and Congestion Prediction Model. The scenario in a CV environment using ML and DL methods to provide an overview of how to predict collision from occurring and how traffic congestion can be estimated. Also communicated between CVs so that a notification can be sent amongst CVs to alert drivers of the upcoming traffic in front of them. By applying the model in this application, both safety and security aspects will be addressed. The application for collision and congestion prediction also alert notification, mainly to alert the driver of the upcoming or potential collision and congestion so that they can be alert and provide necessary action for their vehicle. Fig. 6 illustrates the structure for the implementation of the proposed model in a CV environment. The application approach is divided into three main sections which are Infrastructure, variables and parameters, and the Collision and Congestion Prediction Model. The infrastructure focuses on network functions so that the network is safe from any internal or external threats. Infrastructure depends on the main parameters which are; (i) Speed, (ii) Distance, (iii) Time, (iv) Position. These are the basic elements in a study of CV environment [55]. Each parameter will communicate and updated in this model, to provide real-time alert notification if a collision is potentially occurring and if traffic road is congested. This can help drivers to prevent from an accident by taking necessary action and to reduce traffic congestion in which CV can re-plan route for involving traffic jam. Safety and Security aspects are illustrated in Fig. 5. Each aspect has its parameters to ensure each component functions correctly and efficiently.

C. Applicable Simulation Scenario for Collision Prediction and Congestion Prediction

Fig. 7 and 8, is a simulated scenario of CV infrastructure in a smart city. Essential components including RSU, OBU and CV. RSU and OBU use technological sensors that are relevant for CV implementation, such as GPS, LiDAR, ADAS, TLC, DSRC and RFID. These components will collect variables data such as Time, Speed, Distance and Position of the CV to calculate and provide a prediction of the upcoming potential collision.

In this scenario, several types of communication occur, that includes V2V, V2I and V2X. All the communication would be a pathway to transfer alerts and notifications regarding collision potentially occur. Through the sent alerts and notifications, drivers would get information regarding the collision earlier and enable them to slow down the CV and give a clear passage for authorities such as ambulances, police and fire squad to arrive at the location. V2V communications are focusing on communications between vehicles, while V2I is the communication between vehicles to infrastructures. V2X is where all components interact and communicate with each other. The notifications are an outcome of the collected and analyzed variable data that will be available and sent for all types of road users.

1) Collision prediction: In Fig. 7, the prediction of collision can be performed when CV A have lost control of driving. The analysis is based on the parameters which are collected by the sensors from RSU to CV(OBU) and OBU to RSU, then the data will be processed. All parameters must functional to be calculated are; (i) Time: To get the current time, (ii) Position: to get the current position of CV from time to time (iii) Distance: distance between CV positions, (iv) speed: to get the speed of CV. If a CV is driving at fast speed with a short distance between other CVs (based on the position and time), an alert notification will be sent to alert the CVs of the upcoming possible collision.

If a collision occurs in RSU A area, hence information in RSU A will be analyzed to calculate the level of the collision to the level set in the system. If the level of collision is similar and accurate to the level set in the system, an alert notification will be sent to RSU B and the information will be communicated to all CV in the area. This process would continue for all the available RSUs. By the alert notification sent, the other incoming CVs can take action and slow down their vehicle when there are getting near and arriving at the location of the collision. If no alert notification is being sent, the possibility for a larger scale of collision can occur. Through the alert notification, CVs can provide passage for emergency authorities to attend to the collision location and the other CVs can decide to re-route and take another road to avoid the area.

2) Congestion prediction: In Fig. 8, the prediction of congestion can be performed when RSU detecting potential congestion by the amount of CV in a one RSU area. The analyzed is based on the parameters that be taken by the sensors from RSU to OBU(CV) and OBU to RSU then the data will be processed. The parameters to be calculated are: (i) Time: To get the current time, (ii) Position: to get current position of CVs from time to time (iii) Distance: distance between CV positions, (iv) speed: to get the speed of CV. CVs that are moving slowly in a short distance between one another within an RSU area indicates that a congestion might have occurred, as the time for CV to change position is longer. When a congestion has been detected, an alert notification will be sent to CVs for the upcoming congestion for them to take necessary action such as slowing down, or reroute.

In Fig. 8, there are several components which is RSU and CV (OBU) that are important for communication. RSU A will connect with CV for their data parameters and calculates the possibility of congestion. If traffic congestion occurs in the range of RSU A, the information would be sent to RSU B, for RSU B to send notifications and alert to all CVs within its area. With the communicated information, the rate of traffic congestion can be reduced if the CVs can prepare for the congestion or change its route to another less congested route.

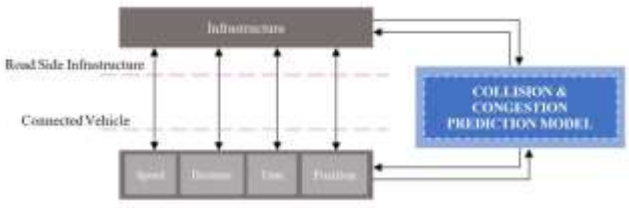


Fig. 6. Collision and Congestion Prediction Model for Safety and Security in Connected Vehicle.

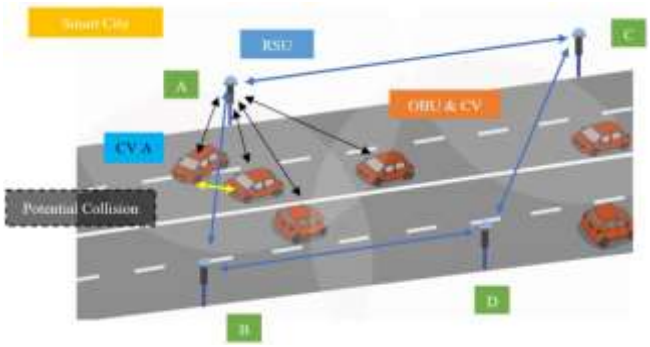


Fig. 7. Simulation Scenario for Collision Prediction.

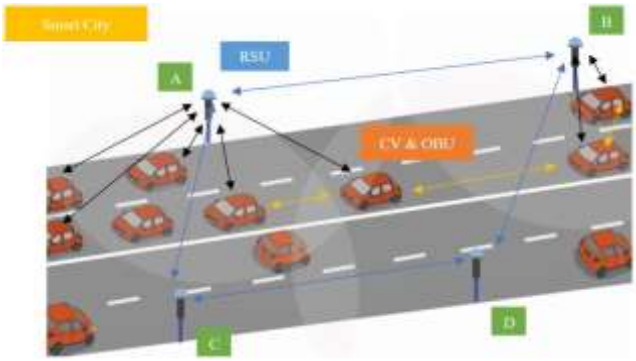


Fig. 8. Simulation Scenario for Congestion Prediction.

V. Conclusion

In this paper, thorough research has been conducted to study the safety and security in connected vehicles through the implementation of big data analytics technology which is machine learning and deep learning. The study also includes a discussion of CV's related technologies and the techniques of machine learning and deep learning that have been applied by other researchers for various implementations. In this study, a conceptual model for both safety and security of CV has been proposed, which includes an application for collision and congestion prediction, by implementing several machine learning and deep learning techniques. The proposed model concerns all layers of CV implementation in IoT which is an application, physical and network layer. A simulation scenario has also been proposed and discussed theoretically, in which for future work, a real simulation, which collects real data as well as data analysis using necessary devices would be conducted.

Acknowledgment

The authors would like to thank all experts especially Associate Prof Dr. Suzaimah Ramli and Prof Ir. Dr. Norazman Mohamad Nor for precious contribution in provided their insight and expertise that greatly assisted towards the whole research activities. This work was supported under the National Defense University Malaysia Short Grants UPNM/2020/GPJP/ICT/4 and UPNM/2018/GPJP/2/TK/5.

References

[1] B. Lv et al., "LiDAR-Enhanced Connected Infrastructures Sensing and Broadcasting High-Resolution Traffic Information Serving Smart Cities," IEEE Access, vol. 7, pp. 79895–79907, 2019.

[2] M. Soyturk, K. N. Muhammad, M. N. Avcil, B. Kantarci, and J. Matthews, From vehicular networks to vehicular clouds in smart cities. Elsevier Inc., 2016.

[3] U. Z. A. Hamid, S. Z. Ishak, and F. Imaduddin, "Current Landscape of the Automotive Field in the ASEAN Region: Case Study of Singapore, Malaysia and Indonesia- A Brief Overview," Asean J. Automot. Technol., vol. 1, no. 1, pp. 21–28, 2019.

[4] A. Sumalee and H. W. Ho, "Smarter and more connected: Future intelligent transportation system," IATSS Res., vol. 42, no. 2, pp. 67–71, 2018.

[5] H. A. Ameen, A. K. Mahamad, S. Saon, D. M. Nor, and K. Ghazi, "A review on vehicle to vehicle communication system applications," Indones. J. Electr. Eng. Comput. Sci., vol. 18, no. 1, pp. 188–198, 2019.

[6] Y. J. S. Yorgos, M. Golias, G. Dedes, C. Douligeris, and S. Mishra, "Challenges, Risks and Opportunities for Connected Vehicle Services in Smart Cities and Communities," IFAC-PapersOnLine, vol. 51, no. 34, pp. 139–144, 2019.

[7] H. Kobayashi, C. Konno, M. Kayashima, and M. Nakano, "Approaches for Vehicle Information Security," Ipa, 2013.

[8] E. Uhlemann, "Introducing connected vehicles [Connected vehicles]," IEEE Veh. Technol. Mag., vol. 10, no. 1, pp. 23–28, 2015.

[9] A. F. Hasan, M. F. Che Husin, K. A. Rosli, M. N. Hashim, and A. F. Zainal Abidin, "Multiple Vehicle Detection and Segmentation in Malaysia Traffic Flow," IOP Conf. Ser. Mater. Sci. Eng., vol. 318, no. 1, 2018.

[10] D. Tokody, A. Albini, L. Ady, Z. Raynai, and F. Pongrácz, "Safety and Security through the Design of Autonomous Intelligent Vehicle Systems and Intelligent Infrastructure in the Smart City," Interdiscip. Descr. Complex Syst., vol. 16, no. 3, pp. 384–396, 2018.

[11] L. Ben Othmane, A. Al-Fuqaha, E. Ben Hamida, and M. Van Den Brand, "Towards extended safety in connected vehicles," IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC, no. Itsc, pp. 652–657, 2013.

[12] S. Doecke, A. Grant, and R. W. G. Anderson, "The Real-World Safety Potential of Connected Vehicle Technology," Traffic Inj. Prev., vol. 16, pp. S31–S35, 2015.

[13] W. G. Najm et al., "Description of Light-Vehicle Pre-Crash Scenarios for Safety Applications Based on Vehicle-to-Vehicle Communications," no. May, p. 109p, 2013.

[14] X. Zeng, K. Balke, and P. Songchitruksa, "Potential Connected Vehicle Applications to Enhance Mobility, Safety, and Environmental Security," Southwest Reg. Univ. Transp. Cent., vol. 7, no. 2, p. No. SWUTC/12/161103-1, 2012.

[15] C. Sun et al., "An active safety control method of collision avoidance for intelligent connected vehicle based on driving risk perception," J. Intell. Manuf., 2020.

[16] G. Panahandeh, E. Ek, and N. Mohammadiha, "Road friction estimation for connected vehicles using supervised machine learning," IEEE Intell. Veh. Symp. Proc., no. Iv, pp. 1262–1267, 2017.

[17] H. Al Najada and I. Mahgoub, "Anticipation and alert system of congestion and accidents in VANET using Big Data analysis for Intelligent Transportation Systems," 2016 IEEE Symp. Ser. Comput. Intell. SSCI 2016, no. December, 2017.

[18] J. Hu, M. Huang, and X. Yu, "E ffi cient mapping of crash risk at intersections with connected vehicle data and deep learning models," Accid. Anal. Prev., vol. 144, no. June, p. 105665, 2020.

[19] N. Formosa, M. Quddus, S. Ison, M. Abdel-Aty, and J. Yuan, "Predicting real-time traffic conflicts using deep learning," Accid. Anal. Prev., vol. 136, no. December 2019, 2020.

[20] "(T) abubakar_2019_Enhanced weight-based clustering algorithm to provide reliable delivery for VANET safety applications.pdf."

[21] A. Jindal, G. S. Aujla, N. Kumar, R. Chaudhary, M. S. Obaidat, and I. You, "SeDaTiVe: SDN-Enabled Deep Learning Architecture for Network Traffic Control in Vehicular Cyber-Physical Systems," IEEE Netw., vol. 32, no. 6, pp. 66–73, 2018.

[22] C. G. Kiran, L. V. Prabhu, A. Rahiman V., and K. Rajeev, "Traffic sign detection and pattern recognition using support vector machine," Proc. 7th Int. Conf. Adv. Pattern Recognition, ICAPR 2009, pp. 87–90, 2009.

[23] Q. He, X. Meng, R. Qu, and R. Xi, "Machine Learning-Based Detection for Cyber Security Attacks on Connected and Autonomous Vehicles," pp. 1–19, 2020.

[24] M. Aloqaily, S. Otoum, I. Al Ridhawi, and Y. Jararweh, "An intrusion detection system for connected vehicles in smart cities," Ad Hoc Networks, vol. 90, no. May, 2019.

[25] F. Martinelli, F. Mercaldo, V. Nardone, and A. Santone, "Car hacking identification through fuzzy logic algorithms," IEEE Int. Conf. Fuzzy Syst., pp. 0–6, 2017.

[26] J. Barrachina, P. Garrido, M. Fogue, and F. J. Martinez, "Road Side Unit Deployment : A Density-Based Approach," vol. 5, pp. 30–39, 2013.

[27] T. J. Wu, W. Liao, and C. J. Chang, "A cost-effective strategy for road-side unit placement in vehicular networks," IEEE Trans. Commun., vol. 60, no. 8, pp. 2295–2303, 2012.

[28] Q. Yang, L. Wang, W. Xia, Y. Wu, and L. Shen, "Development of on-board unit in vehicular ad-hoc network for highways," 2014 Int. Conf. Connect. Veh. Expo, ICCVE 2014 - Proc., pp. 457–462, 2014.

[29] E. Paikari, S. Tahmasseby, and B. Far, "A simulation-based benefit analysis of deploying connected vehicles using dedicated short range communication," IEEE Intell. Veh. Symp. Proc., no. Iv, pp. 980–985, 2014.

[30] S. Dabral, S. Kamath, V. Appia, M. Mody, B. Zhang, and U. Batur, "Trends in camera based Automotive Driver Assistance Systems (ADAS)," Midwest Symp. Circuits Syst., pp. 1110–1115, 2014.

[31] Z. Zhong, S. Liu, M. Mathew, and A. Dubey, "Camera radar fusion for increased reliability in ADAS applications," IS T Int. Symp. Electron. Imaging Sci. Technol., pp. 1–4, 2018.

[32] J. Chen, S. Tian, H. Xu, R. Yue, Y. Sun, and Y. Cui, "Architecture of Vehicle Trajectories Extraction With Roadside LiDAR Serving Connected Vehicles," IEEE Access, vol. 7, pp. 100406–100415, 2019.

[33] G. H. Kulkarni and P. G. Waingankar, "Fuzzy logic based traffic light controller," ICIIS 2007 - 2nd Int. Conf. Ind. Inf. Syst. 2007, Conf. Proc., no. August, pp. 107–110, 2007.

[34] W. M. El-Medany and M. R. Hussain, "FPGA-based advanced real traffic light controller system design," 2007 4th IEEE Work. Intell. Data Acquis. Adv. Comput. Syst. Technol. Appl. IDAACS, no. September, pp. 100–105, 2007.

[35] S. Mehan, "Introduction of Traffic Light Controller with Fuzzy Control System," Int. J. Electron. Commun. Technol., vol. 2, no. 3, pp. 119–122, 2011.

[36] J. Wang, D. Ni, and K. Li, "RFID-based vehicle positioning and its applications in connected vehicles," Sensors (Switzerland), vol. 14, no. 3, pp. 4225–4238, 2014.

[37] N. Lu, S. Member, N. Cheng, S. Member, N. Zhang, and S. Member, "Connected Vehicles : Solutions and Challenges," vol. 1, no. 4, pp. 289–299, 2014.

[38] A. Talebpour and H. S. Mahmassani, "Influence of connected and autonomous vehicles on traffic flow stability and throughput," Transp. Res. Part C Emerg. Technol., vol. 71, pp. 143–163, 2016.

[39] B. Y. R. Khatoun and S. Zeadally, "Smart Cities : Consepts."

[40] X. Yang, J. Liu, F. Zhao, and N. H. Vaidya, "A vehicle-to-vehicle communication protocol for cooperative collision warning," Proc. MOBIQUITOUS 2004 - 1st Annu. Int. Conf. Mob. Ubiquitous Syst. Netw. Serv., pp. 114–123, 2004.

[41] "Challenges Facing Using AI in Cybersecurity | OpenMind." [Online]. Available: https://www.bbvaopenmind.com/en/technology/artificial-intelligence/challenges-facing-using-ai-in-cybersecurity/. [Accessed: 05-Jul-2020].

[42] M. H. Eiza and Q. Ni, "Driving with Sharks: Rethinking Connected Vehicles with Vehicle Cybersecurity," IEEE Veh. Technol. Mag., vol. 12, no. 2, pp. 45–51, 2017.

[43] M. A. Al-garadi, A. Mohamed, A. Al-ali, X. Du, and M. Guizani, "Surveys," Underst. Commun. Res. Methods A Theor. Pract. Approach, pp. 222–237, 2014.

[44] V. Albino, U. Berardi, and R. M. Dangelico, "Smart cities: Definitions, dimensions, performance, and initiatives," J. Urban Technol., vol. 22, no. 1, pp. 3–21, 2015.

[45] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," J. Big Data, vol. 2, no. 1, p. 21, Dec. 2015.

[46] P. Russom, "BIG DATA ANALYTICS - TDWI BEST PRACTICES REPORT Introduction to Big Data Analytics," TDWI best Pract. report, fourth Quart., vol. 19, no. 4, pp. 1–34, 2011.

[47] J. Narula, "Are we up to speed?: from big data to rich insights in CV imaging for a hyperconnected world.," JACC. Cardiovasc. Imaging, vol. 6, no. 11, pp. 1222–4, Nov. 2013.

[48] S. Amini, I. Gerostathopoulos, and C. Prehofer, "Big data analytics architecture for real-time traffic control," in 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2017, no. Tum Llcm, pp. 710–715.

[49] D. Alvarez-Coello, B. Klotz, D. Wilms, S. Fejji, J. M. Gomez, and R. Troncy, "Modeling dangerous driving events based on in-vehicle data using random forest and recurrent neural network," IEEE Intell. Veh. Symp. Proc., vol. 2019-June, no. Iv, pp. 165–170, 2019.

[50] Y. Cui, H. Xu, J. Wu, Y. Sun, and J. Zhao, "Automatic Vehicle Tracking with Roadside LiDAR Data for the Connected-Vehicles System," IEEE Intell. Syst., pp. 44–51, 2019.

[51] Y. Xun, Y. Sun, and J. Liu, "An Experimental Study Towards Driver Identification for Intelligent and Connected Vehicles," IEEE Int. Conf. Commun., vol. 2019-May, 2019.

[52] L. Abberley, K. Crockett, and J. Cheng, "Modelling Road Congestion Using a Fuzzy System and Real-World Data for Connected and Autonomous Vehicles," IFIP Wirel. Days, vol. 2019-April, pp. 1–8, 2019.

[53] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial Neural Networks-Based Machine Learning for Wireless Networks: A Tutorial," IEEE Commun. Surv. Tutorials, vol. 21, no. 4, pp. 3039–3071, 2019.

[54] H. T. Abdelwahab and M. A. Abdel-Aty, "Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections," Transp. Res. Rec., no. 1746, pp. 6–13, 2001.

[55] S. Aoki, T. Higuchi, and O. Altintas, "Cooperative Perception with Deep Reinforcement Learning for Connected Vehicles," Apr. 2020.

[56] J. Rios-Torres, A. Malikopoulos, and P. Pisu, "Online Optimal Control of Connected Vehicles for Efficient Traffic Flow at Merging Roads," IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC, vol. 2015-Octob, pp. 2432–2437, 2015.

A New Hybrid KNN Classification Approach based on Particle Swarm Optimization

Reem Kadry[1]
Faculty of Computer Science
October University of Modern Science and Arts, Egypt

Osama Ismael[2]
Faculty of Computers and Artificial Intelligence
Cairo University, Egypt

***Abstract*—K-Nearest Neighbour algorithm is widely used as a classification technique due to its simplicity to be applied on different types of data. The presence of multidimensional and outliers data have a great effect on the accuracy of the K-Nearest Neighbour algorithm. In this paper, a new hybrid approach called Particle Optimized Scored K-Nearest Neighbour was proposed in order to improve the performance of K-Nearest Neighbour. The new approach is implemented in two phases; the first phase help to solve the multidimensional data by making feature selection using Particle Swarm Optimization algorithm, the second phase help to solve the presence of outliers by taking the result of the first phase and apply on it a new proposed scored K-Nearest Neighbour technique. This approach was applied on Soybean dataset, using 10 fold cross validation. The experiment results shows that the proposed approach achieves better results than the K-Nearest Neighbour algorithm and it's modified.**

***Keywords*—K-nearest neighbour; outlier; multidimensional; particle swarm optimization; scored k-nearest neighbour**

I. Introduction

The rapid development in data collection techniques and storage technologies enables organizations to reserve large amount of data. With the help of data mining algorithms the quality of decision making may be supported and human error is avoided. One of the most popular algorithms of data mining is K Nearest Neighbour (KNN). KNN is a simple effective supervised classification algorithm that is easily to be implemented on different kinds of data. Also, KNN is called lazy learning algorithm while rest of classification algorithms are called eager ones. Precisely speaking, there is no explicit training phase in KNN. It starts working only when it gets an unseen tuple for classification [1].

Multidimensional data and outliers are two of the main problems in KNN algorithm. KNN algorithm does not work accurately with multidimensional data as the distance in each dimension is hardly calculated. Feature reduction and optimization of multidimensional data will help to increase the effectiveness of KNN. Moreover, the presence of outliers in the data to be mined by KNN affects the accuracy of the result. Outlier reduction in KNN is a common challenge in many researches in order to reduce the unwanted data set which is not relevant to the pattern.

Motivated by these facts, in the current work, an effective hybrid approach for pattern detection is devised which uses the output of Particle Swarm Optimization as an input to proposed new scored KNN algorithm. We make two modifications on PSO, the first modification is proposing a new way to calculate the learning rate that was inspired from the neural networks, the second one is a threshold accuracy percentage as a termination condition. We utilize the PSO to solve the multidimensional data and then using this optimized data as initialization to our new proposed scored KNN algorithm. The idea of the proposed scored KNN technique is near to weighted KNN, taking into consideration the calculation of mean and maximum distance for each k-nearest neighbor, then give score for every distance. The aim is to be close to the mean distance. This overcomes the outliers problem in KNN, because KNN equalize all neighbours regardless different distances. So that it will be helpful for achieving maximum accuracy in minimum computation time.

In order to employ formal and systematic evaluation of the work in this research, a set of experiments has been designed on Soybean data sets [18] using 10 fold cross validation.

II. Related Work

KNN is one of the algorithms which performs classification regardless a past knowledge of the data set to be classified. Classification is done based on similarity. Despite the popularity and simplicity of KNN, it has some shortcomings affecting its performance. Many researches are done to enhance KNN performance, within multiple improvement directions; outliers reduction in [2][3][4], appropriate K value selection [5], integration of multiple algorithms [6] [7] [8] and more other improvements [9] [10][11].

Divya and Senthil [2] introduced a new outlier reduction approach based on distance learning for categorical attributes and introduced distance learning framework. The main problem in the introduced approach was the need to remove redundant or irrelevant features to improve classification performance and decrease cost of classification. So, Selahaddin and Ahmet [3] introduced a new density weighted KNN to reduce the effect of irrelevant data. They obtained the coefficient of density of each element of the training data set. Then, determined the relation of each test element based on the total of density coefficients of neighbour that belong to the same class. The problem of this approach was the amount of resources needed to run it, and so it needs more time and memory requirements. Also, Guo-Feng et al [4] proposed a new short term load forecasting mode based on weighted K nearest neighbour algorithm to achieve more satisfied accuracy. The problem with approach is the need to use an

optimization technique in order to improve the forecasting accuracy.

Another enhancement of KNN algorithm is introduced by Natalia et al [5] to select adaptive number of nearest neighbour; for each test point in N, the method looks from 1 to M nearest neighbour at the same time, and finds the value k. This algorithm was applied on diabetes data and it is needed to be generally applicable to classification problem of different data sets to ensure its performance.

More enhancements were done by integration of multiple algorithms, Bahramian and Nikravanshalmani [6] proposed a new classification algorithm based on feature selection with genetic algorithm and combination of k nearest neighbour and Adaboost (a practical boosting algorithm works on classification problems in order to change a group of weak classifiers into a strong one) algorithms to increase the accuracy of classifying diabetes dataset. Also Chen and Hao [7] proposed a hybrid framework of weighted support vector machine and weighted KNN. First, establish detailed feature weighted SVM for data classification giving different weights for different features. Then, estimate the importance of each feature to get weights, this by computing the information gain. Finally, weighted KNN by computing k weighted nearest neighbours from the historical dataset. Reyhaneh Sadat Moayeri et al [8] proposed a hybrid predictive model to evaluate dental implants success using SVM, Neural Networks and KNN. The combined classifier aimed to be achieved a higher accuracy than using one classification algorithm, but the main issue that the data used was only for the medical field with specific attributes.

More enhancements done by Li Yu et al [9], they discussed the effect of distance function on KNN performance. They used four different functions; Euclidean, Cosine, Chi square and Minkowsky and compared the performance results on different medical datasets. Also, Yanpeng et al [10] introduced a multifunction nearest neighbour approach, by combing fuzzy similarity relations and class memberships. This approach gives an adaptively to deal with KNN classification problems.

Shubham Pandey et al [11], proposed a new technique called Modified K Nearest Neighbour (MKNN) which was based on assigning the class label of the required instance into K validated training data points. Then, compute the validation of the data sample in the train set. After that, perform a weighted KNN on any test sample. When they implemented their experiment on Soybean dataset, the reached average accuracy of the MKNN was 85.56 %. The issue was that they compared their results with the classic KNN only. In our research we will compare the results of the proposed approach; Particle Optimized K Nearest Neighbour, with the classic KNN and also with this Modified KNN.

Our proposed approach focuses on solving two main problems affecting the accuracy of KNN; which are multidimensional data and outlier reduction. This was achieved through the integration between Particle Swarm Optimization technique and the new Scored KNN approach, in order to introduce an effective hybrid approach for pattern detection called Particle Optimized Scored KNN (POSKNN).

III. Particle Optimized Scored KNN Approach (POSKNN)

KNN is considered to be simple, effective, intuitive and competitive classification algorithm in several domains. Despite KNN advantages, it has some limitations that can affect its performance. It is very sensitive to irrelevant or redundant features because all features contribute to the similarity and thus to the classification. By careful feature selection or feature weighting, this can be avoided [12]. Also, the presence of outliers will have an effect on the accuracy of classification. As the distance between an object and its neighbour increases, the more it will be considered as an outlier. Outlier reduction is considered as a common problem in KNN algorithm as it affects its accuracy.

Motivated by the above mentioned problems, we propose a novel hybrid approach called Partial Optimized Scored KNN. To build and utilize the model, two phases are undergone; first multi-dimensional feature selection phase. In which, we aim to select the most important features, the result of this phase is a reduced set of features. Second, outlier reduction phase, the input of it is the set of selected feature from phase one, and introducing a new scored KNN technique, aiming to give a score for each distance in order to determine the extreme distance. The next two subsections we will give a detailed description for these two phases.

A. Mulit-Dimensional Feature Reduction Phase

Feature reduction is a critical step in data pre-processing and important research content in data mining tasks such as classification. Feature selection is to effectively reduce feature dimension and improve classification accuracy and efficiency by deleting irrelevant and redundant features in data sets [13]. KNN algorithm has some problems to deal with multi-dimensional datasets. As the number of dimensions increases, the calculated distances are less considerable, so the performance of KNN results decreases.

The main goal of the multi-dimensional feature selection phase is to get rid of the redundant and irrelevant data. This will result in improving the classifier effectiveness and the classifier accuracy through increasing the true positive predictive values. In this research, multi-dimensional feature reduction phase is done through utilizing the Particle Swarm Optimization (PSO) technique. PSO has powerful convergence ability to the optimization value and it can be hybridized with other algorithms easily. PSO is a natural optimization technique based on the synchronization of the movement mechanism of swarms. PSO algorithm simulates animal's social behavior including insects, herd, birds and fishes. These swarms conforms a cooperative way to find food, and each member in the swarms keeps changing the search pattern according to the learning experiences of its own and other members [14].

PSO is a computational searching algorithm which optimizes a problem by trying to improve different solutions on the basis of a specific quality measure. It solves a problem by having a population of different solutions (called particles). Each particle moves around in the search area according to a simple mathematical equation, with specific position and

velocity for each particle. In each move a particle is affected by its local position and also is affected by the updated better position founded by other particles. The process is repeated till a stopping condition is founded which is the number of iterations. At each iteration of PSO, position and velocity of every particle is updated according to (1) and (2):

$$v_i(t+1) = v_i + C_1 \times rand_1 \times (pBest_i - x_i) + C_2 \times rand_2 \times (gBest - x_i) \quad (1)$$

$$x_i(t+1) = x_i + v_i(t+1) \quad (2)$$

Where, the initial position and velocity of particles are generated randomly within the search space. $Rand_1$ and $rand_2$ are generated random numbers ϵ [0, 1]. C_1 is the learning rate of personal experience and C_2 is the learning rate of global experience.

Some basic parameters may affect PSO performance, as the number of particles, number of iterations and learning rate. The number of particles used in PSO ranges from 10 to 100. There is no exact rule in literature for selection of swarm size. But normally, when the dimension of problem at hand increases, the swarm size should also be increased [15]. Too few particles prompt the algorithm to get trapped in local optima, while too many particles slow down the algorithm. Also, the number of particles has an effect on the computational complexity; if its value increase the time consumed to reach good optimization results increases. Based on the done experiments– as illustrated in the next section- we set the number of particles to 20.

The number of iterations in PSO represents the stopping condition of the algorithm. The termination condition may be maximum number of iterations, or termination when finding an acceptable solution to a given problem. We proposed a termination condition depending on the achieved accuracy, in order to decrease the time complexity. Several test cases were implemented in order to illustrate the effectiveness of the proposed termination condition. The algorithm is supposed to terminate when reaching threshold minimum accuracy equals to 94 %. In our case based on the proposed termination condition, the algorithm was terminated after a range from 32 to 37 iterations.

Also, the learning factor represented by the coefficients C1 and C2. To adjust the weight of empirical information of each particle, the coefficient C1 is used, and to adjust the weight of integrated information of the optimal particle in the current population, the coefficient C2 is used. If the value of C1 is too large, it will be easy to fall into local convergence. Also, if the value of C2 is too large, the algorithm will cost an expensive computation overhead and easily fall into iterative divergence. Conventional PSO algorithms usually sets C1 = C2 = 2 or other fixed constants [16]. In our research, we proposed a new way to calculate the learning rate inspired from the calculation of learning rate in neural networks, which depending on the number of iterations as illustrated in (3).

$$Learning\ rate = \frac{Number\ of\ particles}{Number\ of\ iterations} \quad (3)$$

As the number of iterations increases the learning factor decreases for every particle, that's why we divided by the number of iterations. When applying PSO to the used dataset with the proposed termination condition and the proposed learning rate, it gave better performance than applying it with constant values for iteration number and learning rate.

B. KNN Outliers Reduction Phase

Outlier reduction is considered to be an important problem in the field of data mining. In general, outlier reduction is the concept of searching for instances in a dataset which are inconsistent with the remainder of that dataset [17]. KNN accuracy is greatly affected by the presence of outliers. In our research, we aim to identify the unusual records in the used dataset by introducing the new scored KNN approach.

In the second phase; KNN Outliers Reduction Phase, we took the output of the optimized selected features from phase 1, as an input. First, we applied KNN algorithm by initializing the value of k equals to 2 then adjusting the value of k till k equals to 20. And we chose the value of k with respect to experienced test results to get optimized classification results. Then, calculate the distance using standard Euclidean distance. The result of this step will detect the outliers, as the distance to its k^{th} nearest neighbour is considered as the outlying score. Secondly, to solve the problem of outliers, new scored KNN is proposed by sorting the calculated distance in descending order. Thirdly, the mean distance is calculated and the maximum distance is obtained, to be used in the new scored KNN function as follows in (4):

$$New\ Scored\ KNN = 1 - \frac{|distance - mean|}{|maximum - mean|} \quad (4)$$

Where:

- Distance is calculated by the Euclidean distance: $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)}^{\ 2}$
- Mean: is the mean distance based on **K** parameter.
- Maximum: is the maximum distance based on **K** parameter.

Finally, get the sum of the scored distance for each class and the class with maximum score will be the predicted class. The main aim of introducing this function is to be close to the mean distance in order to give the outliers low score.

Our novel hybrid approach; Particle Optimized Score K-Nearest Neighbour is the result of combining 1st phase and 2nd phase. The experiment of the proposed approach will be discussed in details in the experiment section.

IV. Experimental Evaluation

In our experiments the new proposed scored KNN algorithm was run based on the output of applying the particle swarm optimization algorithm and compared the experiment results with the traditional and MKNN algorithm [11]. The used datasets is: "Soybean large dataset"; it is composed of 36 attributes, 15 classes and 684 instances [18].

All the experiments are executed using Matlab 2014. We applied Particle Swarm Optimization on Soybean dataset with the following inputs; number of particles N, number of iteration T and the acceleration coefficients c1 and c2. After

making multiple experiments with different range of number of particles, we found that the best accuracy reached with N = 20. T is the maximum number of iterations, but we proposed using threshold accuracy equals to 94%, which can be achieved after around 30 iterations. But it was concluded that when the value of k is small the threshold accuracy could not be reached, so the algorithm takes more iterations and terminate with different accuracy values less than 94 %. Regarding the acceleration coefficients c1 and c2, also a new method for calculation was proposed, in which c1 and c2 will be equal to the number of particles divided by the number of iterations and it is updated every iteration.

PSO's parameters have a great effect to the convergence rate and the time for reaching the optimal solution. Fig. 1 shows the convergence curve of PSO algorithm, with x-axis representing the number of iterations and with y-axis representing the fitness value that minimize our optimization function (minimize number of features).

After the 1st iteration, the best fitness value of 0.383, was reached by the 7th particle. After the 5th iteration, the other particles were guided by the 7th particle to better best position, and achieved other best fitness value of 0.0315. PSO continued its iterations till reaching the threshold accuracy of 94 %, and converged to the best optimal solution at iteration number 30 with minimum fitness value of 0.243.

Fig. 2 shows that when reaching accuracy of 74 % the number of features were 29 with k value equals to 3, then when the accuracy was 80 %, the number of features was 26 with k value equals to 5. Then the number of features decreased gradually till 18 features, and reached the threshold accuracy of 94 % when k value equals to 14.

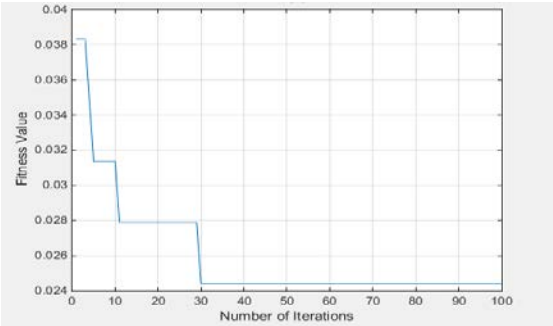


Fig. 1. PSO Convergence Curve.

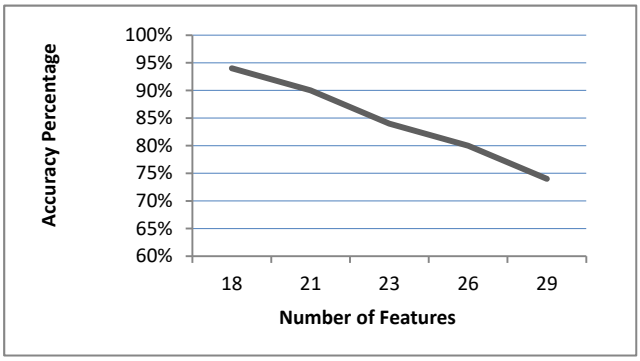


Fig. 2. Number of Iterations based on Percentage.

V. Evaluating Classifier Performance

An evaluation of the classifier performance was needed to know how accurate the classifier is, to predict the class label of tuples. In our experiments the following evaluation measures were used:

- The accuracy of a classifier.
- Error rate.
- Sensitivity and specificity.
- Precision and recall.

The percentage of test set tuples that are correctly classified by the classifier, shows the accuracy of a classifier on a given test set. Also, the error rate or misclassification rate of a classifier is simply 1 – accuracy. When the main class of interest is rare (e.g. in fraud reduction application), the class of interest (or positive class) is "fraud", which occurs much less frequently than the negative "non-fraudulent" class. In medical, there may be a rare class, such as "cancer". Therefore, other measures are needed, that recognize how the classifier can correctly predict the positive tuples (cancer = yes)) and how it can correctly recognize the negative tuples (cancer = no). To apply this, the sensitivity and specificity measures can be used sequentially. The true positive rate is defined as Sensitivity (i.e., the proportion of positive tuples that are correctly identified), while the true negative rate is defined as specificity (i.e., the proportion of negative tuples that are correctly identified). Finally, Precision is a measure of exactness, whereas recall is a measure of completeness. The Recall is also most likely known as sensitivity. These measures are defined as:

$$Accuracy = \frac{TP + TN}{P + N} \qquad Error\ rate = \frac{FP + FN}{P + N}$$

$$Sensitivity = \frac{TP}{P} \qquad Specificity = \frac{TN}{N}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} = \frac{TP}{P}$$

A comparison between the performance evaluation measures of the classical KNN with k values ranges from 3 to 14, MKNN and the new proposed POSKNN approach were conducted through a different set of experiments on the Soybean dataset. The experiments were applied to a different number of nearest neighbors (K) to verify the effect of varying number of K on the experimental results. Table I shows the average of evaluation of classical KNN, MKNN and POSKNN algorithm.

As shown in Table I, when applying POSKNN with different values of k equals 3, 7, 9 and 14 it gave highest average accuracy which shows how our classification is close to real value. It was concluded that when the k value is small the classifier could not reach the threshold accuracy of 94%, so it continued until the 100 iteration and then terminated with accuracy values less than 94%, and when the k value reaches 14, it terminated when reaching the threshold accuracy. POSKNN gave lowest error rate; this shows that the proposed approach gives minimum misclassification rate. Also, it gave

highest recall, indicating that the number of correct results of classification is high. Moreover, POSKNN gave highest specificity showing that the number of actual negatives that is correctly classified is high.

Fig. 3 shows the accuracy evaluation of classical KNN, MKNN and POSKNN algorithm. When K was set to 3, the traditional KNN algorithm performs better than MKNN and POSKNN algorithm. As K increases PSOKNN becomes better than the KNN, and MKNN algorithms.

Based on the empirical studies, after applying different K values of 3, 7, 9 and 14, POSKNN reached an accuracy of 80.86 %, 91.8 %, 92.15 % and 94.21% respectively. When setting K to 14 the best accuracy was reached. When K is too small, the POSKNN classifier may be misleading because of noise in the data. On the other hand, when k is increased more than 14, the POSKNN classifier may misclassify the test instance because its list of nearest neighbours may include data points that are located far away from its neighbourhood, which may include points from other classes. On the other hand, traditional KNN and MKNN reach the best accuracy when K is set to 9, and as the value of K increases, the accuracy decreases.

Fig. 4 shows the average accuracy of KNN, MKNN and POSKNN with K value ranges from 3 to 14. It is observed that our proposed approach (POSKNN) outperform the other algorithms.

From the experimental results, it is noticed that the PSOKNN outperforms traditional KNN and MKNN in all cases, except when K is extremely small, the classifier may be misleading because of noise in the data.

TABLE I. AVERAGE PERFORMANCE EVALUATION MEASURES

Point of Comparison	KNN	MKNN	POSKNN
Accuracy	83.94 %	85.56 %	88.59 %
Error Rate	0.0662	0.0578	0.04854
Recall	75.48 %	77.5 %	79.98 %
Precision	88.4 %	89.3 %	91.2 %
Specificity	89.23 %	90.1 %	91.5 %
Sensitivity	75.48 %	77.5 %	79.98 %

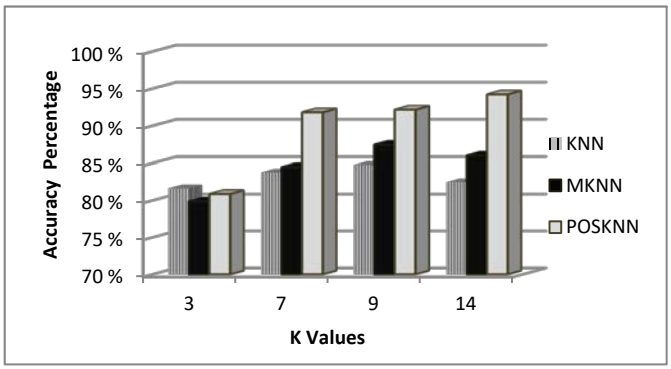


Fig. 3. The Accuracy Evaluation of Classical KNN, MKNN and POSKNN Algorithm.

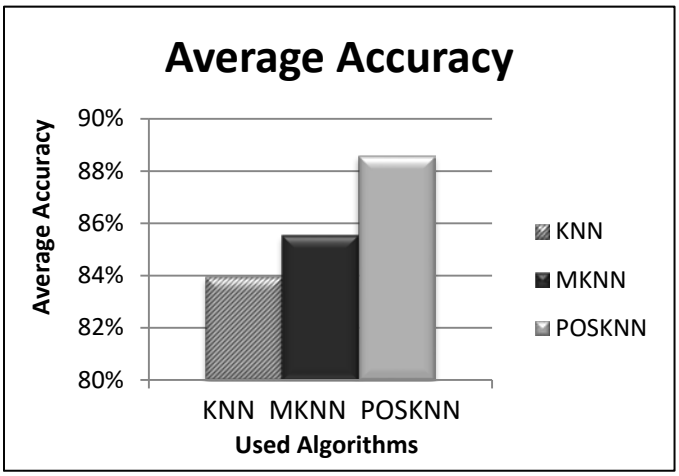


Fig. 4. Average Accuracy of the Three Algorithms.

Hypothesis testing was also used to ensure with a predefined level if the proposed approach has higher performance than the others [19]. T-test was used to check if the differences between the averages of the resulting values for KNN, MKNN and PSOKNN are statistically significant [20]. The results of carrying out the T-test, with using a confidence level equal 95% are illustrated in Table II. GraphPad software (http://www.graphpad.com/) was used for calculating T-test.

TABLE II. T-TEST RESULTS

KNN (average)	MKNN (average)	PSOKNN (average)	K Value	T Value	Two-tailed P value	Statistical significant
81.62	79.87	80.86	3	0.7020	0.5554	Not Statistically Significant
83.75	84.46	91.8	7	0.7788	0.5176	Statistically Significant
82.45	90.08	94.21	14	6.2054	0.0250	Extremely statistically significant.

T-test results demonstrate that when values of k were large (k>=7), the results were statistically significant, while for small values of k (k=3), the results were not statistically significant. When K is set to 14, the two-tailed P value is 0.0250. Therefore, the difference between the means is extremely statistically significant. In the case of K equals 3, the two-tailed P value is 0.5554, it states that the difference between means is not statistically significant. Finally, when K is set to 7, the two tailed P value is 0.5176; thus the difference between means is statistically significant.

VI. CONCLUSION AND FUTURE WORK

This paper proposes a new hybrid approach to improve the performance of KNN classifier, which is called Particle Optimized Scored K-Nearest Neighbour. The proposed approach applies Particle Swarm Optimization to solve the problem of multidimensional data in KNN. PSO is applied with two modifications; first, introducing threshold accuracy

as a termination condition for the number of iterations, second, proposed a new technique to calculate the learning factor inspired from neural networks and customized on PSO algorithm.

In order to get the best performance, we applied a new Scored KNN on the optimized output of PSO. The new Scored KNN takes into account the value of the mean distance and this helped to solve the problem of outliers' problem in KNN. The new proposed hybrid approach was evaluated on Soybean dataset. Results showed that the proposed approach; Particle Optimized Scored K Nearest Neighbour (POSKNN), gave better accuracy than the classic KNN and the MKNN.

As a future work, an implementation of the proposed algorithm will be done on different dataset we need to investigate other optimization algorithms to be applied with the new proposed Scored KNN on the same dataset and compare results. Furthermore, we need to see the effect of applying the same proposed system using different distance measure such as Manhattan, Minkowski and cityblock. Moreover, we will calculate the execution time of the 3 methodologies; classical KNN, MKNN and POSKK, to ensure that the accuracy was increased without increasing the execution time.

REFERENCES

[1] A. Lamba and D. Kumar "Survey on KNN and its variants", International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, Issue 5, May 2016.

[2] K. Divya and N. Kumaran "Improved outlier detection using classic KNN algorithm", International Research Journal of Engineering and Technology, vol. 3, Issue 12, December 2016.

[3] S.Akben and A. Alkan "Density weighted K-nearest neighbors algorithm for outliers in the training set are so close to the test element", Journal of Electrical Engineering, vol. 3, Issue 3, pp. 150 – 161, 2015.

[4] G. Fan, Y. Guo, J. Zheng and W. Hong "Application of weighted K nearest neighbour algorithm for short term load forecasting ", Energies Journal, vol. 5, Issure 12, pp. 916, March 2019.

[5] N. Labuda, J. Seeliger., T. Gedrande and K. Kozak, "Selecting adaptive number of nearest neighbors in k-nearest neighbor classifier apply diabetes data", Journal of Mathematics and statistical Science, Issue 1, 2017.

[6] S. Bahramian and A. Nikravanshalmani, "Hybrid algorithm based on K-nearest algorithm and Adaboost with selection of feature by genetic algorithms for the diagnosis of diabetes", International Journal of Mechatronics, Electrical and Computer Technology (IJMEC), pp. 2977-2986, 2016.

[7] Y. Chen and Y. Hao, "A feature weighted support vector machine and K-nearest neighbour algorithm for stock market indices prediction", Expert System Applications, vol. 80, pp. 340–355, March 2017.

[8] R. Sadat, M. Khalili and M. Nazari "A hybrid method to predict success of dental implants", International Journal of Advanced Computer Science and Applizations, vol. 7, Issue 5, May 2016.

[9] L. Hu, M. Huang,, S. Ke and C. Tsai "The distance function effect on k‐nearest neighbor classification for medical datasets", SpringerPlus, vol. 1, Issue 5, August 2016.

[10] Y. Qu, C. Shang, N. Parthalain, W. Wu. and Q. Shen "Multi- functional nearest-neighbour classification", Soft Computing Journal, vol. 22, Issue 8, pp. 2717 – 2730, March 2017.

[11] S. Pandey, V. Sharma and G. Agrawal,"Modification of KNN algorithm", International Journal of Engineering and Computer Science, vol.8, Issue 11, pp. 24869 – 24877, November 2019.

[12] S. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) approach for predicting economic events: Theoretical background", International journal of engineering research and application, vol. 3, Issue 5, pp. 605 – 610, January 2013.

[13] J. Wang, J, Xu, C. Zhao, Y. Peng and H. Wang, "An ensemble feature selection method for high- dimensional data based on sort aggregation", System science & control engineering journal, vol. 7, Issue 2, pp. 32 – 39, May 2019.

[14] D. Wang, D. Tan and L. Liu, "Particle swarm optimization algorithm: an overview", Soft Computing, vol. 22, pp. 387 – 408, January 2018.

[15] A. Jorhedi and J. Jasni, "Parameter selection in particle swarm optimization: a survey", Journal of experimental & theoretical artificial intelligence, vol. 25, Issue 4, pp. 527-542, June 2013.

[16] M. Ren, X. Huang, X. Zhu and L. Shao, "Optimized PSO algorithm based on the simplicial algorithm of fixed point theory", Applied intelligence, vol. 50, Issue 7, pp. 2009 – 2024, July 2020.

[17] F, Benjelloun, A. Ousso, A. Bennani., S. Belfkih and A. Lachen, A., "Improving outliers detection in data streams using LiCS and voting", Journal of King Saud University- Computer and Information Sciences, pp. 1319 – 1578 August 2019.

[18] D. Grant, Nelson, R. Nelson, S. Cannon and R. Shoemaker, "SoyBase, the USDA-ARS soybean genetics and genomics database", Nucleic Acids Research, vol. 38, Issue suppl_1, pp. D843 – D846, 2009.

[19] T. Kim, "T test as a parametric statistic", Korean Journal of Anesthesiology, vol. 68, Issue 6, December 2015.

[20] F. Emmert and M. Dehmer "Understanding statistical hypothesis Testing: the logic of statistical inference", Machine learning & knowledge extraction, vol. 1, pp. 945-961, August 2019.

An Effective Heuristic Method to Minimize Makespan and Flow Time in a Flow Shop Problem

Miguel Fernández[1], Avid Roman-Gonzalez[2]
Department of Engineering, Pontifical Catholic University of Peru, Lima 32, Peru[1]
Image Processing Research Laboratory (INTI-Lab), Universidad de Ciencias y Humanidades, Lima, Perú[1,2]

***Abstract*—In this paper, it is presented a heuristic method for solving the multi-objective flow shop problem. The work carried out considers the simultaneous optimization of the makespan and the flow time; both objectives are essential in measuring the production system's performance since they aim to reduce the completion time of jobs, increase the efficiency of resources, and reduce waiting time in queue. The proposed method is an adaptation of multi-objective Newton's method, which is applied to problems with functions of continuous variables. In this adaptation, the method seeks to improve a sequence of jobs through local searches recursively. The computational experiments show the potential of the proposed method to solve medium-sized and large instances compared with other existing literature methods.**

***Keywords*—Flow shop problem; multi-objective optimization; non-dominated solution**

I. Introduction

In a flow shop environment, J jobs must be processed on a set of N machines following the same order. The flow shop problem (FSP) consists of determining the sequence of jobs that optimizes one or more performance measures within the $J!$ possible sequences. The FSP is classified as NP-hard for most of the classic problems, for example [1]: $F_2||\sum c_j$, an FSP with two machines and with the aim of minimizing the sum of the completion time of all the jobs (flow time); $F_2||L_M$, an FSP with two machines and with the objective of minimizing the maximum delay; $F_3||c_M$, an FSP with three machines and with the aim of minimizing the completion time of the jobs (makespan). Given the computational complexity that the FSP presents, various heuristics and metaheuristics methods have been proposed in the literature to solve medium-sized and large instances.

Widmer and Hertz (1989) [2] proposed a heuristic method to solve the problem to minimize the makespan. This method consists of two phases: the first phase considers an initial sequence based on a solution to the traveling salesman problem, and the second phase consists of improving this solution using tabu search techniques. Ho (1995) [3] proposed a heuristic to minimize flow time. In this paper, a simulation study was carried out to test the proposed heuristic effectiveness, comparing it with other methods. Murata et al. (1996) [4] proposed a multi-objective genetic algorithm. In this paper, it is considered a weighted sum of multiple objective functions with variable weights. Ponnambalam et al. (2004) [5] proposed a multi-objective evolutionary search algorithm; the authors solve a traveling salesman problem and employ a genetic algorithm to minimize the makespan, flow time, and downtime. Pasupathy et al. (2006) [6] proposed a multi-objective genetic algorithm, using local search techniques and minimizing makespan and flow time. This algorithm makes use of the principle of non-dominance in conjunction with an agglomeration metric. One can mention other works that adopt the generic algorithm for the FSP [7, 8, 9, 10, 11].

II. Problem Formulation

The FSP is a working system of J jobs and N machines in series, where each job must be processed in each of the N machines. All jobs must follow the same processing sequence: first on machine 1, then on machine 2, so on consecutively. The assumptions are as follows:

- Each machine works continuously and without interruptions.
- Each machine can process just one job at a time.
- Each job can be processed by one machine at a time.
- The processing times of the jobs in the machines are deterministic data.
- The setup times of the machines are included within the processing time.

The performance measures or objective functions considered are the makespan (c_M) and the flow time (c_F). The makespan optimization seeks to reduce the completion time of the jobs and aims to efficiently use resources, while the optimization of flow time reduces the average number of jobs in the queue [6]. The following notation is used to formulate the FSP:

Sets

i: Job index, $i = \{1, \dots, J\}$

k: Order index, $k = \{1, \dots, K\}$

m: Machine index, $m = \{1, \dots, N\}$

Parameters

J, K: Numbers of Jobs

N: Numbers of machines

d_{im}: Processing time of job i on the machine m

Variables

R_{ik}: 1, if the job i is executed in the order k; 0, in other cases.

p_{km}: Processing time of the job to be executed in the order k and on the machine m

c_{km}: Completion time of the job to be executed in the order k and on the machine m

The FSP is formulated as follows:

$$Min\, c_M = c_{K,N} \quad (1)$$

$$Min\, c_F = \sum_{k=1}^{K} c_{k,N} \quad (2)$$

Subject to:

$$\sum_{k=1}^{K} R_{ik} = 1 \ \forall i \quad (3)$$

$$\sum_{i=1}^{J} R_{ik} = 1 \ \forall k \quad (4)$$

$$p_{km} = \sum_{i=1}^{J} d_{im} R_{ik} \ \forall k, m \quad (5)$$

$$c_{1,1} = p_{1,1} \quad (6)$$

$$c_{k,1} = c_{k-1,1} + p_{k,1} \ \forall k | k \geq 2 \quad (7)$$

$$c_{1,m} = c_{1,m-1} + p_{1,m} \ \forall m | m \geq 2 \quad (8)$$

$$c_{k,m} \geq c_{k-1,m} + p_{k,m} \ \forall k, m | k \geq 2 \text{ and } m \geq 2 \quad (9)$$

$$c_{k,m} \geq c_{k,m-1} + p_{k,m} \ \forall k, m | k \geq 2 \text{ and } m \geq 2 \quad (10)$$

$$R_{ik} \in \{0,1\} \quad \forall i, k \quad (11)$$

$$c_{km}, p_{km} \geq 0 \quad \forall k, m \quad (12)$$

Objectives (1) and (2) represent the minimization of makespan and flow time, respectively. Constraints (3) and (4) determine the order of execution of the jobs. According to the order of execution, each job's processing time on the machines is defined in restriction (5). Constraints (6) - (10) determine the completion time of the jobs on the machines. Constraints (11) and (12) define the domain of the decision variables.

III. Multi-Objective Optimization

A multi-objective optimization problem is defined as follows [12]:

$$Min\, F(x) = \{F_1(x), \ldots, F_r(x)\}$$

$$s.t \ \ x \in X$$

Where, x is a decision variable of dimension n, $x = \{x_1, \ldots, x_n\}$, and X is the search space contained in $\mathbb{R}^n$. Generally, the search space X is generated by a set of restrictions and ranges of the decision variables. The multi-objective optimization problem consists of finding a solution $x^* \in X$, so that $\nexists\, y \in X$ such that:

$$F_i(y) \leq F_i(x^*) \text{ for all } i = 1, \ldots, r$$

$$F_j(y) < F_j(x^*) \text{ for some } j = 1, \ldots, r$$

Here, x^* is the call of a non-dominated solution. A non-dominated solution cannot be improved relative to any objective function without worsening at least one other objective function. The set of non-dominated solutions is called the Pareto optimal set, and the image of a given Pareto optimal set is called the Pareto frontier.

IV. Newton's Method for Multi-Objective Optimization

Newton's method for solving multi-objective optimization problems was developed by [13]. The method is based on a multi-start descent algorithm, which consists of generating initial solutions, which will be improved recursively, following a search direction (Newton's direction), with the objective functions.

A. Newton's Direction

Given a function $F: U \subset R^n \rightarrow R^m$ twice continuously differentiable and a non-stationary point $x \in X$, Newton's direction in x, denoted by $s(x)$, is obtained by solving the following problem:

$$\min \max_{j=1,\ldots,r} \nabla F_j(x)^T s + \frac{1}{2} s^T \nabla^2 F_j(x) s$$
$$s.t. \ \ s \in \mathbb{R}^n.$$

The optimal value of the problem, denoted by $\theta(x)$, and Newton's direction are determined as:

$$\theta(x) = \inf_{s \in \mathbb{R}^n} \max_{j=1,\ldots,r} \nabla F_j(x)^T s + \frac{1}{2} s^T \nabla^2 F_j(x) s$$

$$s(x) = \operatorname*{argmin}_{s \in \mathbb{R}^n} \max_{j=1,\ldots,r} \nabla F_j(x)^T s + \frac{1}{2} s^T \nabla^2 F_j(x) s$$

This problem is solved recursively, determining in each step t, the values of $s(x_t)$ and $\theta(x_t)$, and then doing $x_{t+1} = x_t + s(x_t)$, until $\theta(x_t) \approx 0$ (with a certain level of tolerance), that is, until it is not possible to continue improving the objective functions simultaneously.

V. Heuristic Method for the Flow Shop Problem

In this article, a heuristic method based on Newton's method is proposed for the FSP. The proposed method adapts Newton's method, considered a discrete search space.

A. Principal Structure

The procedure starts from a randomly generated sequence of s^* jobs (initial solution). This solution is improved recursively by applying local searches in neighborhoods by the insertion method [14] and by the two-job exchange method [2]. If J is the number of jobs, the insertion method consists of removing a job placed in the *i-th* position and inserting it in the *k -th* position (see Fig. 1a), the size of the generated neighborhood is $(J-1)^2$. The two-job exchange method consists of exchanging the job placed in the *i*-th position with the job placed in the *k*-th position (see Fig. 1b), the size of the generated neighborhood is $J(J-1)/2$.

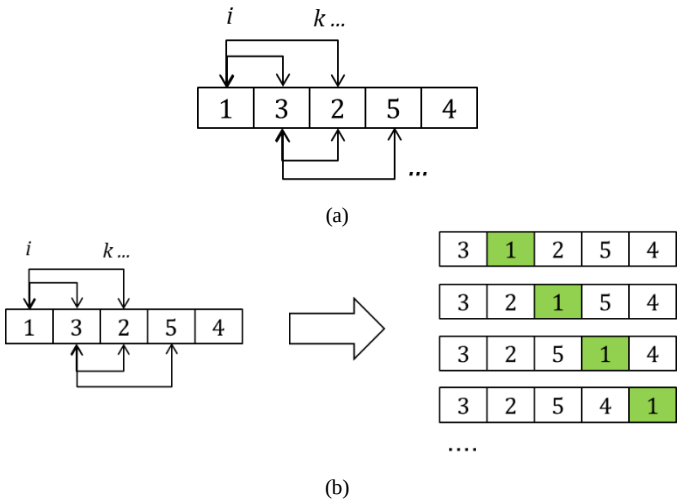


Fig. 1. Neighborhood of Solutions for FSP: (a) Insertion Method; (b) Two-Job Exchange Method.

The pseudo-code of the main structure is presented below:

Main structure (NS)
  $S \leftarrow \emptyset$, $ND \leftarrow \emptyset$;
  for $i = 1, \dots, NS$ **do**
    Generate a initial solution s^*;
    improvement ← TRUE;
    while (improvement = TRUE) **do**
      Improve the jobs sequence s^* **by insertion metho**
      $s' \leftarrow s^*$;
      Improve the jobs sequence s^* **by interchange me**
      if ($s' = s^*$) **then**
        improvement ← FALSE;
      end
    end
    $S \leftarrow S \cup \{s^*\}$;
  end
  $ND \leftarrow$ non-dominated solutions in S;
Return (ND)

Here, NS represents the number of solutions generated initially, S the set of all sequences that have been enhanced, and ND the non-dominated set of S.

B. Improve the Jobs Sequence

The procedure starts from an initial sequence of jobs $s_0 = s^*$, at $t = 0$; its objective is to improve at least one objective function in each iteration. In each iteration *t*, from s_t a neighborhood $N(s_t)$ is generated, and evaluating the parameter θ, the best neighbor (formed solution) of $N(s_t)$ is chosen. The best neighbor of $N(s_t)$ will be s_{t+1}, and the value of t is increased by one. The procedure stops when it is no longer possible to improve a sequence (*descent = FALSE*). Finally, s^* is assigned the best sequence found.

Improve the jobs sequence
  $s_0 \leftarrow s^*$, $t \leftarrow 0$, descent ← TRUE;
  while (descent = TRUE) **do**
    $df_1 \leftarrow \emptyset$, $df_2 \leftarrow \emptyset$;
    $\theta \leftarrow \min_{s \in N(s_t)} \max_{j=1,2}(f_j(s) - f_j(s_t))$;
    if ($\theta < 0$) **then** $s_{t+1} \leftarrow \operatorname{argmin}_{s \in N(s_t)} \max_{j=1,2}(f_j(s) - f_j(s_t))$;
    else if ($\theta = 0$) **then**
      foreach $s \in N(s_t)$ **do**
        if ($\max_{j=1,2}(f_j(s) - f_j(s_t)) = 0$) **then**
          $df_1 \leftarrow df_1 \cup \{f_1(s) - f_1(s_t)\}$,
          $df_2 \leftarrow df_2 \cup \{f_2(s) - f_2(s_t)\}$;
        else
          $df_1 \leftarrow df_1 \cup \{0\}, df_2 \leftarrow df_2 \cup \{0\}$;
        end
      end
      if ($\min_{s \in N(s_t)} \min_{j=1,2}(df_j(s)) = 0$) **then** descent ← FALSE;
      else if ($\min_{s \in N(s_t)} df_1(s) = 0$) **then**
        $s_{t+1} \leftarrow \operatorname{argmin}_{s \in N(s_t)} df_2(s)$;
      else if ($\min_{s \in N(s_t)} df_2(s) = 0$) **then**
        $s_{t+1} \leftarrow \operatorname{argmin}_{s \in N(s_t)} df_1(s)$;
      else $s_{t+1} \leftarrow \operatorname{argmin}_{s \in N(s_t)} df_1(s)$;
      end
    else descent ← FALSE;
    end
    if (descent=TRUE) **then** $t \leftarrow t + 1$;
    end
  end
  $s^* \leftarrow s_t$;
Return (s^*)

VI. Computational Experiments

The computational experiments were carried out in MATLAB and executed on a computer with a 2.4 GHz processor and 2 GB of RAM.

The instances used in the experiments were taken from [15]. Each instance is represented by $J \times N$, where *J* is the number of jobs and *N* is the number of machines. In this study, the instances TA31, TA41, TA61, and TA71 are used. The results obtained by the proposed method are compared with the results of the MOGLS [4], ENGA [8], GPWGA [10], and PG-ALS [6]. The proposed method was applied considering 100 initial solutions with ten replicas for each instance. Tables I to IV show the non-dominated solutions of the cited existing methods and the proposed method's non-dominated solutions. Fig. 2 to 5 illustrate the Pareto frontiers that are obtained by different methods.

TABLE I. COMPUTATIONAL RESULTS OF THE TA31 INSTANCE: 50×5

Existing Algorithms		Proposed Method	
c_M	c_F	c_M	c_F
2724	71531	2724	68516
2729	68036	2729	68139
2731	67028	2733	67883
2752	66061	2734	67826
2757	66052	2735	66222
2758	66047	2743	66158
2763	66032	2746	66024
2765	66024	2748	65977
2770	65979	2752	65717
2799	65963	2757	65531

TABLE II. COMPUTATIONAL RESULTS OF THE TA41 INSTANCE: 50×10

Existing Algorithms				Proposed Method	
c_M	c_F	c_M	c_F	c_M	c_F
3047	93511	3133	90663	3072	92115
3052	93013	3134	90641	3080	91797
3059	92666	3135	90448	3084	91241
3063	92602	3137	90408	3098	91023
3070	92508	3148	90364	3099	90981
3074	92493	3152	90305	3106	90955
3075	92124	3156	90254	3120	90656
3076	91757	3197	90207	3141	90628
3087	91688	3209	90165	3142	90557
3097	91256	3237	90158	3146	90520
3099	91236	3249	90099	3147	90428
3111	91149	3298	90075	3154	89538
3132	90882				

TABLE III. COMPUTATIONAL RESULTS OF THE TA61 INSTANCE: 100×5

Existing Algorithms		Proposed Method	
c_M	c_F	c_M	c_F
5493	287684	5493	261717
5495	262647	5495	259338
5498	262335	5498	259088
5527	261411	5538	258507
5563	261071	5539	258501
5564	260706		

TABLE IV. COMPUTATIONAL RESULTS OF THE TA71 INSTANCE: 100×10

Existing Algorithms				Proposed Method	
c_M	c_F	c_M	c_F	c_M	c_F
5801	325462	5858	314749	5836	318588
5803	324725	5877	312785	5842	313791
5804	318924	5881	312632	5843	313790
5806	318299	5892	312534	5848	313769
5816	318055	5897	312349	5849	313208
5827	316972	5904	312207	5856	312643
5832	316642	5915	310887	5866	311872
5836	316542	5920	310515	5874	309183
5837	316292	5928	310359	5903	308818
5838	316161	5934	310297	5905	308291
5840	315753	5995	310227	5912	307660
5851	315184	6001	310040	5960	307349
5856	314879	6009	310005		

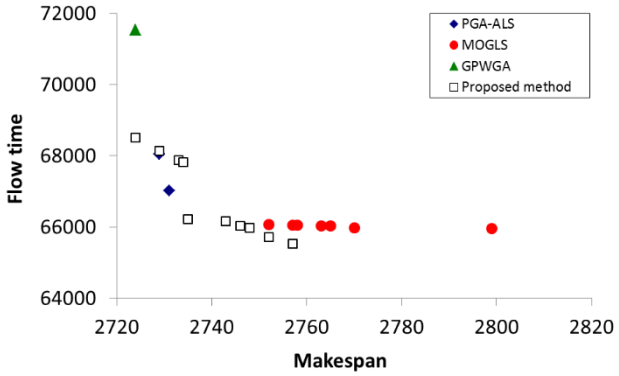


Fig. 2. Approximation of the Pareto Frontier for Instance TA31: 50 × 5.

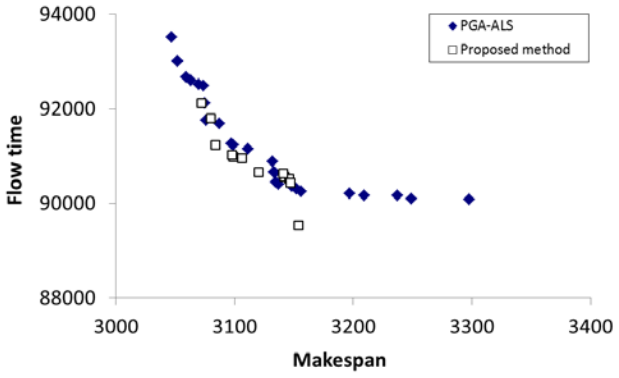


Fig. 3. Approximation of the Pareto Frontier for Instance TA41: 50 × 10.

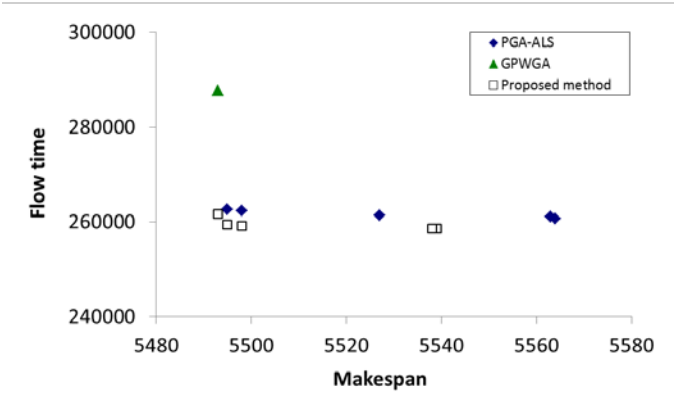


Fig. 4. Approximation of the Pareto Frontier for Instance TA61: 100 × 5

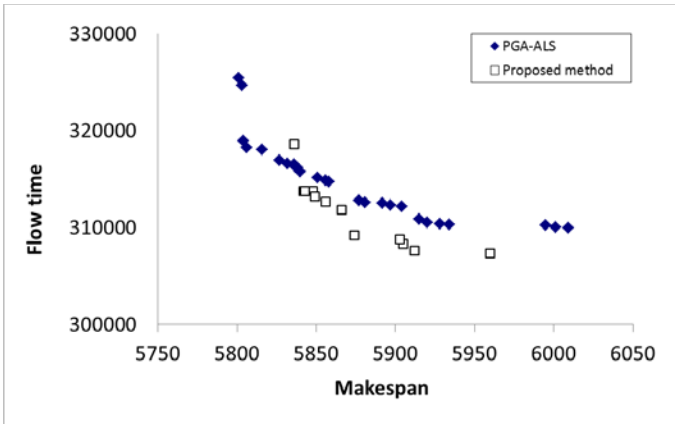


Fig. 5. Approximation of the Pareto Frontier for Instance TA71: 100 × 10.

The results found show that the proposed method has obtained a good approximation of the Pareto frontier and even surpassing the solutions found by the existing methods. Note that problems with 50 or 100 jobs can be considered complex problems. Experiments indicate that generating 100 initial solutions is sufficient to obtain good results in cases with ten or fewer machines.

VII. Conclusions

In this paper, a heuristic method is proposed to solve the flow shop problem, considering the simultaneous optimization of the makespan and the flow time. This method is inspired by multi-objective Newton's method.

In Section 6, the proposed method demonstrated its ability to obtain a set of satisfactory solutions in medium-sized and large instances, generating 100 initial solutions. However, for more extensive cases (concerning the number of jobs or machines), the initial solutions should be increased.

Lastly, unlike other methods, the proposed method has an advantage because it is not necessary to calibrate several parameters by carrying out previous experiments, as happens, for example, with the genetic algorithm.

References

[1] PINEDO, M. Scheduling theory algorithms and system. 3ª Edição. New York: Prentice Hall, 2008.

[2] WIDMER, M.; HERTZ, A. A new heuristic method for the flow shop sequencing problem. European Journal of Operational Research, v. 41, p. 186-193, 1989.

[3] HO, J. Flowshop sequencing with mean flowtime objective. European Journal of Operational Research, v. 81, p. 571-578, 1995.

[4] MURATA, T.; ISHIBUCHI, H; TANAKA, H. Multi-objective genetic algorithm and its applications to flowshop scheduling. Computers Ind. Eng., v. 30, n. 4, p. 957-968, 1996.

[5] PONNAMBALAM, S. G.; JAGANNATHAN, H.; KATARIA, M.; GADICHERLA, A. A TSP-GA multi-objective algorithm for flow-shop scheduling. The International Journal of Advanced Manufacturing Technology, v. 23, p. 909-915, 2004.

[6] PASUPATHY, T.; RAJENDRAN, C.; SURESH, R. K. A multi-objective genetic algorithm for scheduling in flow shops to minimize the makespan and total flow time of jobs. The International Journal of Advanced Manufacturing Technology, v. 27, p. 804-815, 2006.

[7] REEVES, C. R. A genetic algorithm for flowshop sequencing. Computers & Operations Research, v. 22, p. 5-13, 1995.

[8] BAGCHI, T, P. Multiobjective scheduling by genetic algorithms. Boston: Kluwer, 1999.

[9] BUZZO, R. W.; MOCCELLIN, J. V. Programação da produção em sistemas flow shop utilizando um método heurístico híbrido algoritmo genético-simulated annealing. Gestão & Produção, v. 7, p. 364-377, 2000.

[10] CHANG, P. C.; HSIEH, J. C.; LIN, S. G. The development of gradual priority weighting approach for the multi-objective flowshop scheduling problem. International Journal of Production Economics, v. 79, p. 171-183, 2002.

[11] KAMIRI, N.; ZANDIEH, M.; KARAMOOZ, H.R. Bi-objective group scheduling in hybrid flexible flowshop: A multi-phase approach. Expert Systems with Applications, v. 37, p. 4024-4032, 2010.

[12] KONAK, A.; COIT, D.; SMITH, A. Multi-objective optimization using genetic algorithms: a tutorial. Reliability Engineering & System Safety, v. 91, p. 992-1007, 2006.

[13] FLIEGE, J.; DRUMMOND, L. M. G.; SVAITER, B. Newton's method for multiobjective optimization. Optimization Online, 2008.

[14] NAWAZ, M.; ENSCORE, JR., E.; HAM, I. A heuristic algorithm for the m-machine, n-job flow-shop sequencing problem. The International Journal of Management Science, v. 11, n. 1, p. 91-95, 1983.

[15] TAILLARD, E. Benchmarks for basic scheduling problems. European Journal of Operational Research, v. 64, p. 278-285, 1993.

Level of Budget Execution According to the Professional Profile of Regional Governors Applying Machine Learning Models

José Luis Morales Rocha[1], Mario Aurelio Coyla Zela[2], Nakaday Irazema Vargas Torres[3]
Jarol Teófilo Ramos Rojas[4], Daniel Quispe Mamani[5], José Oscar Huanca Frias[6]
Universidad Nacional de Moquegua, Moquegua, Perú[1, 2, 3, 4]
Universidad Nacional de Juliaca, Juliaca, Perú[5, 6]

***Abstract*—Machine Learning is a discipline of artificial intelligence that implements computer systems capable of learning complex patterns automatically and predicting future behaviors. The objective was to implement a Machine Learning model that allows to identify, classify and predict the influence of the professional training of the governors in the execution of the public spending of the regional governments of Peru. Of the 14 indicators of academic training, professional experience and university studies were selected as significant indicators that contribute to the execution of public spending by the 25 governors of Peru. For the prediction of the execution of the public spending of the regional governors, a supervised learning algorithm was implemented. The mean square error for the Machine Learning regression model was 4.20 and the coefficient of determination was 0.726, which indicates that the execution of public spending by regional governments is explained with 72.6% by the professional experience and university studies of the governors. The regional governors of Peru with university studies and professional experience achieve better results in the execution of public spending in the regional governments of Peru.**

***Keywords*—Machine learning; multiple regression; professional experience; university studies; public budget; governor; public spending**

I. Introduction

The execution of public spending is carried out by the institutions of the public sector of a country over a year. It is carried out to acquire goods and services and for the provision of subsidies and transfers, in order to satisfy the needs of its inhabitants, public consumption and contribute to the redistribution of wealth [1]. In [2], the reduction in public expenditure causes the economic growth, and the deterioration of the population's living conditions. Currently, it is observed that public expenditure is insufficient to adequately meet the duties that the State has with its population. Educational training [3] is the key factor for the development of a country.

Through the observation technique, 14 indicators of academic training of the governors of Peru were collected through the electoral platform of the National Elections Jury, the execution of public spending through the Portal of Economic Transparency of the friendly consultation of the Ministry of Economy and Finance. With the multiple regression analysis, professional experience and university studies were selected as significant indicators that contribute to the execution of public spending by the governors of the 25 regions of Peru. The determination coefficient (0.726) of the Machine Learning regression model indicates that the execution of public spending by regional governors is explained in 72.6% by professional experience and university studies.

The purpose of the research is to propose a machine learning model with Machine Learning techniques that allows to identify and predict the influence of the professional training of the governors in the execution of the public spending of the regional governments of Peru.

In this context, [4], [5] the professional profile of regional governors stands out, as they must achieve results with significant coefficients in executing the budget according to a schedule with the objective of reducing the gaps of economic inequality and unsatisfied social needs in an efficient, effective and transparent manner, for the welfare of citizens and thus achieving regional and national development.

This paper is organized as follows. Section II reviews some related works. Section III is made up of the theoretical background; Section IV is the presentation of the obtained results. Section V is titled discussion and contains the analysis of the results. Section VI contains the paper conclusion, followed by the last section, which is Section VII and presents suggestions for future research.

II. Related Works

This section presents the references of different investigations related to Machine Learning, academic training and public spending.

In [6] proposes a machine learning approach to detect and prevent cyberbullying using machine learning techniques. Evaluation of the proposed approach to the cyberbullying dataset shows that the neural network works best achieving 92.8% accuracy and support vector machines reaches 90.3.

In [7] he considers that many Machine Learning approaches are used to generate different models for prediction. However, she claims that the success of Machine Learning-based approaches depends on several factors. Likewise, it considers that no particular Machine Learning technique is effective in all its applications and that the success of the technique depends on the application in the problem to

be solved, so it is important to understand its behavior that guarantees to use the technique. It also uses conventional statistical techniques for bioclimatic modeling.

In [8] he tries to predict the Facebook profile using the Machine Learning technique, stating that the candidates can be chosen. They identified the characteristics that can be extracted from Facebook, through which the personality prediction is viable, the data has been extracted using the Facebook Graph API, which was carried out on a web page. To build the Machine Learning knowledge base, the personality test was implemented for students close to graduating from the University, in order to execute the training and categorization of the machine learning models through the use of the tool for the knowledge analysis, and thus check the degree of accuracy of the algorithms used in predicting the personality of the Facebook user.

In [9] he considers that the promotion of intrusion detection methods in computer networks poses a challenge for researchers, because with the growth of computer networks, new content-based infiltrations constantly appear. The work constitutes different Machine Learning techniques applied to the data processing stages for detection. Taxonomies and connection attribute classification sketches are described. In the detection of anomalies from Machine Learning techniques, it is concluded that it is of great applicability for those who seek areas within the detection of intrusion in computer networks from Machine Learning techniques.

In [10] he mentions that Machine Learning or machine learning is based on the process of systems that learn from historical data to predict future data. Machine Learning in different environments requires a large amount of multivariate and multidimensional environment data, which need to be analyzed to diagnose through statistical procedures, which represents an area of opportunity for machine learning.

In [11], k-means is selected as the base clustering and provides an algorithm for clustering sets of multiple clusters of k-means based on a hypothesis. In addition, they study the extraction of credible local labels from a grouping of bases, the production of different groupings of bases, the construction of the grouping relationship, and the final assignment of each object. In [12] proposes a clustering method based on distance weighted and K-means. In [13] reviewed and applied two known and used clustering methods, k-means and hierarchical clumping, to air pollution studies.

In [14] he considers that applying the technique of documentary review and gathering information for the budget execution of the Portal of the Ministry of Economy and Finance and for the variable of professional training of the mayor of the National Elections Jury Portal. Results were obtained that show that through statistical tests, the influence of the mayor's professional training in the budget execution of the Puno Region is not significant, concluding that the labor practice in the public sector and the mayor's age, more district poverty has a significant influence on budget execution; However, the level of education and the specific profession of the mayor do not significantly influence the budget execution of the Puno Region, years 2015-2018.

As indicated [15] it analyzes public debt in Tamaulipas and compares it with the evolution of public spending, indicating which items and where spending has been directed in the period (2003-2013). First, the situations that affect the public indebtedness of the states are studied; low fiscal pressure, absence of fiscal sovereignty, high state public expenditures of the governing parties, restricted public financial transparency. The state's indebtedness is also analyzed in relation to federal transfers and the application of the resources transferred from the expenditure budgets of that period. It concludes that an assessment of the current situation of Tamaulipas that it presents in terms of indebtedness, verifying that the accelerated growth of debt and public spending in Tamaulipas in the analyzed period, which is reflected in comparable economic growth rates, losing positions in various rankings of competitiveness and safety.

As indicated [16] in their research Influência dos public expenditures do not grow economico dos municípios da Região Sudeste do Brasil. The public expenditures analyzed were related to health, education and culture in 2010, in this investigation it shows a multivariate nonlinear regression mathematical model, whose purpose is to analyze the relationship between economic growth and public spending. The results showed an estimation error of 14.98% on average for the municipalities analyzed. The explanatory power of the model was 97.7% with high reliability. The State of Sao Paulo showed the highest economic growth and the State of Rio de Janeiro among the smallest. The evidence found indicates that the southeast of Brazil there is a positive influence of public spending in relation to economic growth, with the highest spending on education and health. With the application of the model it is inferred that public spending drives the municipal and or state GDP. Concluding that public spending plays an important role for economic growth in the Southeast Region of Brazil.

As pointed out [17] in their article the effect of public spending on the chances of reelection of Spanish local governments during the period 2000-2007, using the logit methodology for panel data. The results of which indicate the increases in municipal public spending have a positive impact on the chances of reelection of local governments, and these increases when the pre-electoral period approaches. We can also observe that there are other variables that positively affect the probability of reelection, such as the volume of income from transfers, which are right-wing parties, or the fact that a municipality has obtained an absolute majority in the previous elections. But we can also observe that the number of years that a mayor is in office has a negative impact on the chances of reelection.

As indicated by [18] in their research Institutional operational plan and the efficiency of public spending in regional governments whose purpose was to determine if the Institutional Operational Plans (POI) affect the efficiency of public spending in the departments from Peru, being an applicative type research with a cross-sectional, analytical, observational and non-experimental design. The approach was quantitative and the level was explanatory correlation, whose purpose was to determine the degree of relationship of the variables, in a population of 25 institutional operational plans

of the Regional Governments of Peru, the sample was non-probabilistic (10 POI of the Regional Governments that had greater budget execution); the instrument that was used was the data record, structured in 8 sentences. Whose result was 0.078, that is, the institutional operational plans do not affect the efficiency of the public spending of the Regional Governments in fiscal year 2018 and this leads us to the following conclusion that the Institutional Operational Plan (POI) did not affect the efficiency of public spending of Regional Governments during 2018 in the health and education sector.

As stated, [19], [20] in their article The impact of public spending on private investment in Mexico (1980-2015). The objective of which was to investigate the relationship between private investment, public spending and public investment in Mexico, for the period 1980-2015. A time series analysis was carried out using an ADL model, which included the variables private investment, primary public spending and GDP. The results obtained that the total net effect of primary public spending and GDP on private investment is positive and considerable. Therefore, in Mexico, between 1981 and 2015, the fall in private investment as a proportion of GDP can be explained by the fall in different types of public spending as a proportion of GDP. Coming to the following conclusion.

Government investment was reduced to such an extent that it acted to the detriment of the country's total investment and by remaining at low levels it cannot explain the behavior of private investment. Regarding the relationship between private and government investment, government investment spending must be very small and that it positively influences private investment must exceed the minimum level. The results of the econometric analysis show that the total net effect of primary public spending on private investment is positive and of considerable magnitude.

III. Theoretical Background

Machine learning is considered a discipline in the field of artificial intelligence, responsible for developing systems that are capable of automatically learning complex patterns with a huge amount of data to predict future behaviors [21].

A. Multiple Regression Model

Machine Learning [22] is considered by many researchers as Artificial Intelligence, one of the most important characteristics of Artificial Intelligence is the ability to learn. Machine learning was designed with the purpose of implementing computer systems that can adapt and benefit from its knowledge. The Machine Learning technique consists of learning the data inputs, evaluating the results of the model and optimizing the output [23].

Machine learning algorithms consist of the following main models:

1) Supervised learning: the algorithm is trained with the inputs and outputs, in order to predict the output of future inputs.

2) Unsupervised learning: the algorithm is presented with inputs without desired outputs.

3) Reinforced learning: the algorithm interacts with the environment and achieve a specific goal without training.

The machine learning algorithm [24] is composed of supervised, semi-supervised and unsupervised learning, it is applied in different areas of knowledge.

Machine learning [25] instead of feeding the data into the program, now uses the data and the output it has collected to derive its program (also known as a model). This model can be used to make predictions.

Machine learning is a collection of algorithms and techniques that are used to design systems that learn from data. Machine learning algorithms have a solid mathematical and statistical foundation, but do not take into account domain knowledge. Machine learning consists of the following disciplines [25]:

1) Scientific computing
2) Mathematics
3) Statistics

B. Types of Machine Learning Algorithms

There are the following types of machine learning algorithms:

1) Supervised learning algorithms: are those algorithms that are trained with labeled data. This means that they are data composed of examples of the desired responses. Most of the machine learning is supervised.

2) Unsupervised learning algorithms: are those algorithms that are used in data without labels, and the objective is to find relationships in the data.

C. Multiple Regression Model

The construction of the multiple regression model is an iterative process, this process of construction of the model is helped by graphs that help to visualize the relationships between the different variables in the data, generate associations between the variables of the problem under study and consider the importance of develop such relationships between variables. This model can be fit and inferred, performing fit diagnostics to verify the assumptions of the model [26]. It is essential to know how the academic training of the governors intervenes in the budget execution.

Estimation and inference in multiple linear regression. The multiple linear regression model is as follows:

$$E\left(Y\middle|X_1 = x_1,.,x_p = x_p\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + . + \beta_p x_p \quad (1)$$

So,

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + e_i \quad (2)$$

Where:

e_i= random fluctuation (or error) in Y_i such that $E(e_i|X) = 0$.. In this case, the response variable Y_i is predicted from p predictor (or explanatory) variables $X_1, X_2, \ldots, X_p,$, and the relationship between Y_i and $X_1, X_2, \ldots, X_p$, is linear in the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

1) Least squares estimates: The least squares estimate of $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the values of $b_0, b_1, b_2, \ldots, b_p$ for the sum of the squared residuals.

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (3)$$

$$RSS = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_p x_{pi})^2 \quad (4)$$

is a minimum. For RSS to be minimal with respect to $b_0, b_1, b_2, \ldots, b_p$ we need

$$\frac{\partial RSS}{\partial b_0} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_p x_{pi}) = 0$$

$$\frac{\partial RSS}{\partial b_1} = -2\sum_{i=1}^{n} x_{1i} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_p x_{pi}) = 0$$

$$\vdots$$

$$\frac{\partial RSS}{\partial b_p} = -2\sum_{i=1}^{n} x_{pi} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_p x_{pi}) = 0$$

This gives a system of (p + 1) equations in (p + 1) unknowns. In practice, a software package is needed to solve these equations and thus obtain the least squares estimates, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$.

2) Least squares matrix formulation: A convenient way to study the properties of the least square estimates, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ is to use matrix and vector notation. Define the vector (n × 1), Y, the matrix n × (p + 1), X, the vector (p + 1) × 1, β of unknown regression parameters and the vector (n × 1), e of random errors by.

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} \ldots x_{1p} \\ 1 & x_{21} \cdots x_{2p} \\ \vdots & \vdots \\ 1 & x_{n1} \cdots x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

We can write the multiple linear regression model in matrix notation as:

$$Y = X\beta + e \quad (5)$$

Also, let x_i denote the ith row of matrix X. Then:

$$x_i' = (1 \;\; x_{i1} \;\; x_{i2} \ldots \; x_{ip}) \quad (6)$$

is a row vector 1 × (p + 1) that allows us to write in the following way:

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e_i = x_i'\beta \quad (7)$$

The residual sum of squares as a function of β can be written in matrix form as:

$$RSS(\beta) = (Y - X\beta)'(Y - X\beta) \quad (8)$$

D. K-Means Algorithms

In the k-means algorithm [13] k points are randomly selected from a data set to be considered as the initial central points for the grouping. The Euclidean distance is then determined by using the distance to determine the distance between the data points and the centroids of the cluster, and the data set is grouped again according to the distance.

Finally, the average distance in each group is calculated, and the new center is adjusted in the data set of each group, finally the final result of the grouping is obtained through multiple iterations.

The k-means clustering algorithm [27] is a fundamental clustering technique in the field of machine learning. Clustering begins by randomly initializing a centroid for each of a total of k groups.

$$c_i = argmin \left| x_i - \mu_j \right|^2 \quad (9)$$

Also, m and k are the number of data and groups, respectively, and c_i is the index of the group to which the ith center is now assigned.

Second, assuming that the group assigned to each x_i is constant, the new centroids are calculated using a new subset of points, where *m* y *k* are the number of data and groups, respectively, and c_i is the index of the group to which *ith* data that is now assigned.

$$\mu_j = \frac{\sum_{i=1}^{m} [c_i = j] x_i}{\sum_{i=1}^{m} [c_i = j]} \quad (10)$$

where X shows 1 if the condition is true and 0 otherwise. This two-step process is repeated iteratively until convergence is achieved, ultimately minimizing the error criterion.

$$\varepsilon = \sum_{j=1}^{k} \sum_{i=1}^{m} [c_i = j] \left\| x_i - \mu_j \right\|^2 \quad (11)$$

The algorithm will converge to a minimum value, so that the result is sensitive to the initial position of the centroids and will generally only converge to a minimum value.

Centroids at different locations are often initialized and error checked to determine the optimal value.

IV. Results

This section shows the results obtained from applying supervised machine learning algorithms.

A. Research Variables

The indicators of academic training of the governors of the regions of Peru were obtained from the National Elections Jury (consultation of jurisdictional files), the execution of public spending of the regional governments from the Portal of Economic Transparency of the friendly consultation of the Ministry of Economy and Finance (MEF).

Of the 14 indicators of academic training, professional experience and university studies were selected as significant indicators that contribute to the execution of public spending by the governors of the 25 regions of Peru.

B. Statistic Analysis

The measures of central tendency on the quantitative characteristics of the research are shown.

In Table I, it is shown that the average age of the governors is 57 years, with 11 years of professional experience on

average, an average income of PEN 91298 and 86% on average of execution of public spending by regional government. Other descriptive measures such as the amount of data, standard deviation, maximum and minimum values and quartiles are observed for each of the quantitative characteristics of the regional governors of Peru.

In Fig 1, the correlations that exist between the quantitative variables are shown. We can observe that only between the execution of public spending and professional experience there is a moderate correlation (r = 0.43), which is not observed with the other variables.

In Fig. 2, a scatter diagram is shown, indicating that the greater the professional experience, the greater the execution of public spending in the regional governments of Peru. It is also observed that governors with university studies show a greater execution of public spending.

C. Regression Analysis

In Table II, the summary of the multiple regression model is shown.

The results show a coefficient of determination of 0.378, which indicates that the execution of public spending is explained in 37.8% by the professional experience and by the university studies of the regional governors of Peru.

TABLE I. DESCRIPTIVE STATISTICS

Statistics	Age	Professional experience	Income	Public spending execution
count	25.00	25.00	25.00	25.00
mean	57.16	11.48	91298.16	86.07
std	10.92	11.39	110761.60	5.91
min	38.00	0.00	0.00	68.65
25%	51.00	4.00	21600.00	83.44
50%	57.00	7.00	63700.00	87.05
75%	67.00	14.00	106200.00	90.64
max	74.00	42.00	412938.00	94.65

Fig. 1. Correlation Matrix.

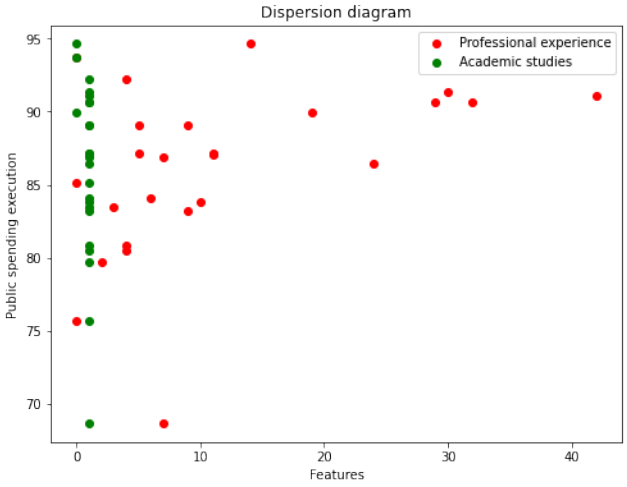


Fig. 2. Dispersion Diagram.

TABLE II. MODEL SUMMARY

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,615[a]	,378	,321	4,86272

a. Predictors: (Constant), Professional experience, University studies

In Table III, the Analysis of Variance (ANOVA) offers information about the adequacy of the regression model to estimate the values of the dependent variable. Through the Snedecor F statistic, it is observed that the Sig. (P-value = 0.005) is less than 0.05 of significance, this means that the regression model is significant.

The regression model (Table IV) shows that the coefficients of the regression are significant, this indicates that professional experience and university studies are determining factors that contribute to the execution of public spending by the regional governors of Peru.

TABLE III. ANOVA[A]

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	315,726	2	157,863	6,676	,005[b]
Residual	520,214	22	23,646		
Total	835,940	24			

a. Dependent Variable: Public spending execution

b. Predictors: (Constant), Professional experience, University studies

TABLE IV. COEFFICIENTS[A]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	90,293	2,967		30,437	,000
University studies	-7,761	2,993	-,436	-2,593	,017
Professional experience	,228	,087	,440	2,616	,016

a. Dependent Variable: Public spending execution

D. Supervised Machine Learning Algorithms

The Machine Learning linear regression model algorithm has the following characteristics:

- Independent variables: 02
- Dependent variables: 01
- Training set: 70%
- Test set: 30%

The coefficients of the multiple linear regression model applying Machine Learning techniques are shown in the table.

In Table V the regression model with the application of Machine Learning techniques is shown:

$$Pse = 90.25 + 0.2299\ Prof.exp. - 8.0086\ Univ.stud \quad (12)$$

The mean square error for the Machine Learning regression model is: MSE = 4.20303.

Table VI shows the prediction values with the Machine Learning model.

E. Evaluation and Validation of the Algorithm

Mean square error is the most widely used endpoint for regression supervised learning problems.

In Fig. 3, it is observed that the prediction values for cases, the results obtained are very similar to the test data set, the low value of the mean square error (4.20) indicates that there is a good fit of the regression model of Machine Learning.

F. Cluster Analysis with K-Means

There are really many alternatives for Cluster formation, but for the present study k-means algorithms were used.

In Fig. 4, a characterization map of the regional governments of Peru is observed, according to some characteristics of the regional governors such as age, academic studies, professional experience and execution of public spending in the regional governments of Peru.

TABLE V. COEFICIENTES DEL MODELO

Model	B
(Constant)	90.2501657
University studies	0.22998494
Professional experience	-8.00861339

TABLE VI. PREDCCIÓN DEL MODELO

Y test	Y predicted
90.64	88.91111546
86.88	83.85144686
91.13	91.90091964
80.81	83.16149205
85.16	82.24155231
86.43	87.76119078
91.33	89.14110040
83.44	82.93150712

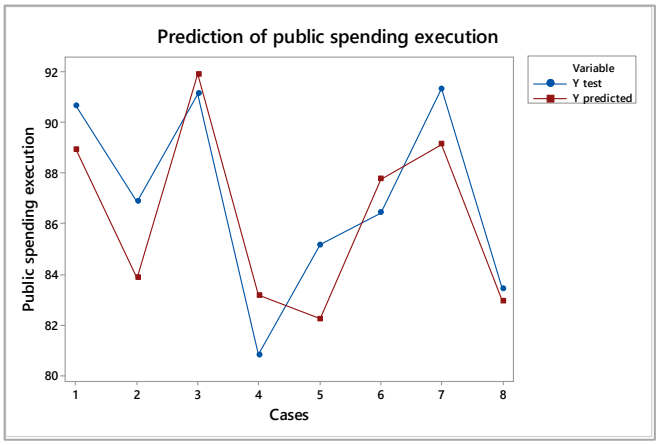


Fig. 3. Prediction of Public Spending Execution.

In Cluster 1, there are 7 departments, Ancash, Arequipa, Cusco, Junín, Pasco, Piura and Tacna.

In Cluster 2, we find 6 departments, Amazonas, La Libertad, Madre de Dios, Tumbes, Lima and the Constitutional Province of Callao.

Cluster 4 considers 7 departments, Apurímac, Ayacucho, Cajamarca, Huancavelica, Huánuco, Puno and Ucayali.

In Cluster 1, there are 5 departments, Ica, Lambayeque, Loreto, Moquegua and San Martin.

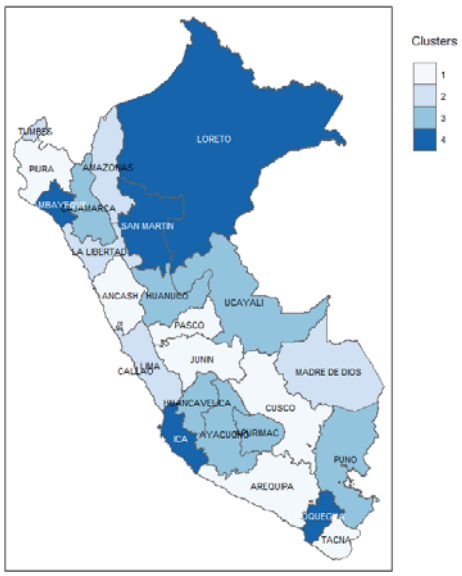


Fig. 4. Mapa De Caracterización De Gobiernos Regioanles Del Perú.

TABLE VII. CLASIFICACIÓN DE LOS GOBIERNOS REGIONALES DEL PERÚ

Cluster	N	Age	Academic studies	Professional experience	Public spending execution
1	7	44	University	5	82%
2	6	68	University	6	83%
3	7	56	No university	9	90%
4	5	64	University	31	90%

According to the results of Table VII, Cluster 1 shows regional governors with an average age of 44 years, with university studies, with an average of 5 years of professional experience and are those who execute public spending on average in a 82%. In Cluster 2, there are those governors who, on average, are 68 years old, with university studies, on average 6 years of professional experience and execute 83% of public spending. In Custer 3, there are governors with an average age of 56 years, they do not have university studies, but they have an average of 9 years of professional experience and they execute 90% of public spending. Finally, in Custer 4, there are governors with 64 years of age on average, with university studies and with approximately 31 years of professional experience and show an execution of public spending of 90%.

V. Discussion

According to the objective, to identify the factors that influence the professional training of governors in the execution of public spending of the regional governments of Peru, the results of Table IV show that professional experience and university studies present significant coefficients, this indicates that are determining factors that contribute to the execution of public spending of the regional governors of Peru, results that when compared with what was found in [28] who indicate that the variables included in the model are significant, this means that significance is necessary for The variable is considered as a predictor variable. These results affirm that professional experience and university studies are more significant factors that contribute to the execution of public spending by the regional governors of Peru. In the regression [26] the information is collected, the dependent and the independent variables are selected, the model is built and finally it is validated. It means that once, at least one of the regressors has been determined, it is important to answer the question Which one(s) is(are) useful? special attention must be taken when including regressors, because only the significant one(s) that have value to explain the answer must be considered.

Regarding the objective: to implement a Machine Learning model that allows making predictions in the execution of public spending based on the professional experience and university studies of the governors of Peru, the results indicate that the governors of the regional governments of Peru with studies University students and professional experience achieve better results in the execution of public spending in the regional governments of Peru. [14] concludes stating that work experience in the public sector has a significant influence on budget execution; However, the level of education and the specific profession of the mayor, do not significantly influence the budget execution of the Puno Region, it also concludes that most of the mayors have a level of professional education, which constitutes 32.11% , of the total and that the level of education and specific professional training of the mayors of the Puno Region are indistinct and varied with respect to the budget execution. These results also affirm that professional experience and higher education influence the execution of public spending by the governors of Peru.

VI. Conclusions

Professional experience and university studies are the most significant factors that contribute to the execution of public spending by the regional governors of Peru.

The governors of the regional governments of Peru with university studies and more professional experience achieve better results in the execution of public spending in the regional governments of Peru.

With the application of k-means algorithms, the regional governors of Peru were classified according to characteristics such as age, academic studies, professional experience and execution of public spending in 4 groups: Cluster 1, there are 7 governors of the departments from Ancash, Arequipa, Cusco, Junín, Pasco, Piura and Tacna, in Cluster 2 the governors of the departments of Amazonas, La Libertad, Madre de Dios, Tumbes, Lima and the Constitutional Province of Callao, in Cluster 3, It considers the governors of the departments of Apurímac, Ayacucho, Cajamarca, Huancavelica, Huánuco, Puno and Ucayali and Cluster 4 includes the governors of the departments of Ica, Lambayeque, Loreto, Moquegua and San Martin. The governors with 64 years of age on average, with university studies and with approximately 31 years of professional experience, show a greater execution of public spending.

VII. Future Work

Future work will focus on proposing Machine Learning models to analyze:

The theory of public choice and orientation of expenditures executed by the Peruvian municipalities.

Incidence of the modalities of labor connection in the financial effectiveness of the Peruvian municipalities

References

[1] R. Durán, A. Barreix, L. Corrales, and A. Rasteletti, “Reengineering the Expenditure Budget of the Federation in Mexico The experience of a modern zero-based budget.” Felipe Herrera Library of the Inter-American Development Bank, 2018, doi: http://dx.doi.org/10.18235/0001225.

[2] C. Tello, “Austerity, public expenditure and economic growth with social justice,” Jorunal Econ. Lit., vol. 16, pp. 54–60, 2019.

[3] O. Holguín and J. R. Dos Santos, “Economy and Education: An Approach from the Perspective of Development and Economic Growth,” Comp. Res. Public Policies, pp. 63–86, 2019.

[4] M. Marcel and M. Guzman, “Budgets for development in Latin America.” Felipe Herrera Library of the Inter-American Development Bank, 2014.

[5] E. Cavallo and A. Powell, “Building opportunities to grow in a challenging world Latin America and the Caribbean Macroeconomic Report 2019.” Felipe Herrera Library of the Inter-American Development Bank, 2019.

[6] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, “Social media cyberbullying detection using machine learning,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 5, pp. 703–707, 2019, doi: 10.14569/ijacsa.2019.0100587.

[7] M. Bhattacharya, “Machine Learning for Bioclimatic Modelling,” Int. J. Adv. Comput. Sci. Appl., vol. 4, no. 2, pp. 1–8, 2013, doi: 10.14569/ijacsa.2013.040201.

[8] E. R. Sarchi Albuja, “Predicting Personality from Facebook Profiles Using Machine Learning to Support the Selection of Human Talent Personnel,” University of the Armed Forces, 2018.

[9] J. L. Rivero, “Machine learning techniques for intrusion detection in computer networks,” Cuba. J. Comput. Sci., vol. 8, no. 4, pp. 52–73, 2014, [Online]. Available: http://repositorio.espe.edu.ec/xmlui/handle/21000/14165.

[10] J. C. González-Islas, “Machine Learning in Physiotherapeutic Applications,” Pädi Sci. Bull. Basic Sci. Eng. ICBI, vol. 7, no. Especial, pp. 104–110, 2019, doi: 10.29057/icbi.v7iespecial.4473.

[11] L. Bai, J. Liang, and F. Cao, “A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters,” Inf. Fusion, vol. 61, no. January, pp. 36–47, 2020, doi: 10.1016/j.inffus.2020.03.009.

[12] W. Yang, H. Long, L. Ma, and H. Sun, “Research on clustering method based on weighted distance density and k-means,” Procedia Comput. Sci., vol. 166, pp. 507–511, 2020, doi: 10.1016/j.procs.2020.02.056.

[13] P. Govender and V. Sivakumar, “Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019),” Atmos. Pollut. Res., vol. 11, no. 1, pp. 40–56, 2020, doi: 10.1016/j.apr.2019.09.009.

[14] Y. M. Poma, “Influence of the mayor’s professional training in the budget execution of the Puno region, years 2015-2018,” National University of the Altiplano, 2019.

[15] F. García Fernández, R. A. Vaquera Salazar, and J. A. Serna Hinojosa, “Tamaulipas: indebtedness and public spending (2003-2013),” Inf. Econ., vol. 403, pp. 70–90, 2017, doi: 10.1016/j.ecin.2017.05.005.

[16] L. Degenhart, M. Vogt, and V. C. da S. Zonatto, “Influência dos gastos públicos no crescimento econômico dos municípios da Região Sudeste do Brasil,” REGE - Rev. Gestão, vol. 23, no. 3, pp. 233–245, 2016, doi: 10.1016/j.rege.2016.06.005.

[17] M. T. Balaguer-Coll and M. I. Brun-Martos, “The effect of public expenditure on the probability of mayors’ reelection,” Account. Mag. Spanish Account. Rev., vol. 16, no. 1, pp. 74–80, 2013, doi: 10.1016/S1138-4891(13)70008-6.

[18] E. S. Chuquiyauri and G. G. Robles, “The institutional operational plan and the efficiency of public spending on regional governments,” no. 2018, pp. 6–16, 2020.

[19] F. J. Santos, M. de la O. Barroso, and C. Guzmán, “The global economy and social enterprises,” World Econ. Mag., vol. 35, pp. 177–196, 2013, [Online]. Available: https://www.redalyc.org/pdf/866/86629567010.pdf.

[20] F. S. Gutiérrez Cruz, “The impact of public spending on private investment in Mexico (1980-2015),” Econ. UNAM, vol. 14, no. 42, pp. 136–149, 2017, doi: 10.1016/j.eunam.2017.09.006.

[21] A. D. J. Bedoya Carrillo, “Development of a web tool for classifying bone health in schoolchildren according to age and sex using machine learning techniques,” Universidad Católica de Santa María, 2019.

[22] A. Raghda Essam, E.-K. Hatem, L. Soha Safwat, and S. Yasmine Ibrahim, “Prediction of employee performance using machine learning techniques,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 8, pp. 80–88, 2019, doi: 10.14569/IJACSA.2019.0100812.

[23] F. K. Alsheref and W. H. Gomaa, “Blood diseases detection using classical machine learning algorithms,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 7, pp. 77–81, 2019, doi: 10.14569/ijacsa.2019.0100712.

[24] N. Nadar and R. Kamatchi, “A novel student risk identification model using machine learning approach,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 11, pp. 305–309, 2018, doi: 10.14569/ijacsa.2018.091142.

[25] L. Wei-Meng, Python Machine Learning, vol. 53, no. 9. Indianapolis,: John Wiley & Sons, Inc, 2019.

[26] D. J. Olive, Linear regression. Switzerland: Springer Nature, 2017.

[27] T. A. Sipkens and S. N. Rogak, “Technical note : Using k -means to identify soot aggregates in transmission electron microscopy images,” J. Aerosol Sci., no. September, p. 105699, 2020, doi: 10.1016/j.jaerosci.2020.105699.

[28] R. Vilà, M. Torrado, and M. Reguant, “Multiple Linear Regression Analysis with SPSS,” REIRE Rev. d ’Innovació i Recer. en Educ., vol. 12 (2), pp. 1–10, 2019, doi: 10.1344/reire2019.12.222704.

Investigating Students' Computational Thinking Skills on Matter Module

Noraini Lapawi[1], Hazrati Husnin[2]
Centre of Teaching and Learning Innovation
National University of Malaysia
Bangi, Malaysia

***Abstract*—The revolution of the fourth industrial has impacted most aspect of our life and demanding a paradigm shift including education. It has become to our attention that there is a need to inculcate complex problem-solving skills among youth to equipped them to face the challenges in the era of digital technology. To fulfill the needs, computational thinking was introduced in school curriculum in Malaysia in 2017. It is still rather new, and this creates opportunity to understand how computational thinking can best be integrated in teaching and learning. In this study, we developed a module for a science topic, Matter and examine its impact on computational thinking skills on 65 students at secondary level. The computational thinking skills integrated in this study were abstraction, decomposition, algorithm, generalization, and evaluation. A quasi-experimental method was employed, and the ANCOVA result showed that there was no significant difference between control and treatment group on computational thinking skills. However, the score means for each of the computational thinking skills for both groups, showed that three skills in the treatment group were higher than the control group. The three computational thinking skills were decomposition, evaluation, and algorithm. This study suggested that CT involved mental process and proper planning is crucial to integrate computational thinking skills as teaching and learning is very contextual in nature.**

Keywords*—Computational thinking skills; problem solving skill; teaching and learning; decomposition; evaluation; algorithm; science module; matter; secondary level students

I. Introduction

The 4th Industrial Revolution has a global impact on human life, economic and the social landscape. The rapid growth of technology requires a paradigm shift in various fields including education, especially in shaping a generation to become competent, resourceful, and competitive to cope with the challenges in the near future. It is crucial to ensure the education given to the young generations is robust and meeting these demands. The World Economic Forum (WEF) stated complex problem-solving skill as one of the highest percentages out of nine skills outlined in the list of skills in demand for the future workplace [1]. Therefore, in most of development countries such as Malaysia, inculcating critical thinking and problem-solving skills is one of the prime agenda in education sector especially for the critical subjects such as science and mathematics [2]. By doing so, students can apply science knowledge to make decisions and solving complex problems in the context of real life creatively and innovatively.

This study is fully funded by Dana Penyelidikan FPEND, GG-2019-014.

Knowing the high demands, problem solving skills has been introduced in most of curriculum system at the early stage of school. One of the approaches used is via computational thinking (CT). CT has been known as an approach that has been proven effective in helping to improve students' problem-solving skills using the concepts of computer science [3]. The common concepts of CT are decomposed, pattern recognition, abstraction, algorithm, logical reasoning, and evaluation. These concepts can be used not just to formulate problem but also to produce an automated solution [4]. Thus, the approach of applying CT skills in teaching and learning is seen to have the potential to produce students who have good problem-solving skills. This is in line with the intention to fulfill the needs for the fourth Industrial Revolution. However, the implementation of CT in Malaysia's education is rather new, and it is still unclear how CT can best be integrated in teaching and learning. Therefore, this study aims to examine the CT approach used in teaching and learning of Matter topic and the impact on students' CT skills. This study hypothesized that there is no significant difference in the mean score of the computational thinking skills test between the treatment and control groups (H_{01}). This paper begins by discussing about CT in teaching and learning in global context and scoping down to Malaysia context. The Matter module and its implementation in this study is also discussed, followed by the methodology employed and discussions on the findings.

II. Literature Review

A. Computational Thinking in Teaching and Learning

Computational thinking involves a systematic thought processes in solving problems based on computer science concept i.e. programming concept such as decomposition, abstraction, and algorithmic thinking [5]. According to [5], CT skill is seen as a necessity for every individual in today's digital era. Although computational thinking is an approach to problem-solving that normally being associated to the use of computers, the process of using mental skills is more prominence rather than solely relying on just computer [6]. Through integrating CT components such as decomposition and abstraction in the ideation process of developing solutions, creativity and innovation can be encouraged [7].

Exam-oriented and teacher-centered has been the practice for quite a while in most context of our education system. Through this practice, the learning approach emphasize memorizing rather than developing full understanding of the topic [8]. In addition, the teacher-centered or some may refer to

as conventional approach is a kind of learning approach that emphasizes on the content and achievement [9]. Thus, the learning itself has become less meaningful. Therefore, the conventional approach needs to be transformed into a student-centered approach by integrating technology in the process of learning [10]. In relation to inculcating computational thinking skills, this could be done through activity that involved plugged-in activity such as using programming [11]. In addition, the skills on coding or programming is also crucial in the future digital workforce [12]. In relation to this study, computational thinking is a form of skill that is suitable to be practiced by all students [13]. Thus, the concepts of computational thinking used in this study is decomposition, algorithm, abstraction, evaluation, and generalization.

B. Computational Thinking in Malaysia

In the context of education in Malaysia, computational thinking began to be offered in the curriculum since 2017 in some subjects such as Information Communication Technology (known as TMK) at primary level and Basic Computer Science (known as ASK) at secondary level. In this subject, students were exposed to the basic components of computational thinking such as decomposition, algorithm, and abstraction [14]. It could be said that the implementation of CT in Malaysia is still new and the dissemination of CT knowledge especially among teacher is still in progress. For example, a study showed that there was a misconception among teachers towards the concept of CT and its integration in teaching and learning [15]. Furthermore, a study also showed the low level of concern among teachers on applying CT in teaching and learning in Malaysia [16] albeit other study showed there were positive attitude among teachers towards the implementation of CT in the curriculum [17]. The inconsistency of findings presented in these studies described there is more need to be done regarding CT research area. On the other hand, this also raised opportunity for further study especially on how CT can best be integrated in teaching and learning. For example, a study by [18] showed that the CT skills among science students in one of district in Malaysia is low.

III. MATTER MODULE

The content covered in the module used in this study is Matter. The module emphasized group activity and creating opportunity for the students to engage in the learning activity and solving problem. The activities require students to solve problems using computational thinking skills represented by a simple programming using Scratch and hands-on activities. During the process, students can learn from their peers as they engage during the group work. As such, cooperative learning was applied, and scaffolding occurred between students and students and teachers with students and this helped those in the zone of proximal development [19].

The theoretical underpinned the design of learning activity in the module was based on Constructionism theory [20]. The theory emphasizes the construction of a new idea can be developed if students are able to produce artifact that is meaningful and shareable among their peers. In this module, the idea of learning by making was instilled and students were challenged to used the knowledge gained to develop a new complex idea, which in this case the programming project that they produced [21]. The activity in the module allowed students to integrate decomposition, algorithm, abstraction, generalization, and evaluation when explaining the changes in states of matter through Scratch program.

IV. METHODOLOGY

This study employed a quasi-experimental method to examine the effectiveness of CT approach integrated in Matter module in developing students' CT skills. There were 65 participants participated in this study. They were divided into two group as described in Table I. The control group received a conventional approach intervention, while treatment group received CT Matter Module as intervention. Before the study takes place, teachers were given three days training on the module as well as training on Scratch application. The intervention for both control and treatment group took place for five weeks. The instrument used to measure CT skills was Computational Thinking Skills Test (UKPK) for both groups. The UKPK test were administered both group before and after the intervention.

The UKPK was developed by adapting Bebras CT Task rubric. There was a total of 15 items in the UKPK in a form of multiple-choice questions ranging from easy, moderate, and difficult question. The UKPK were administered to the participants in control and treatment group during pre-test to identify the level of students' existing CT skills and to ensure the homogeneity of both groups. The UKPK instrument again were administered to both group after they received the intervention. Table II showed the details of the UKPK instrument and the items.

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

TABLE I. THE INSTRUMENT AND THE INTERVENTION

Group			
	Pre-Test	***Intervention***	***Post-Test***
Control (n=31)	Computational Thinking Skills Test (UKPK)	Conventional approach	Computational Thinking Skills Test (UKPK)
Treatment (n=34)	Computational Thinking Skills Test (UKPK)	CT Matter Module	Computational Thinking Skills Test (UKPK)

TABLE II. THE ITEM OF UKPK INSTRUMENT

No	CT Skills	Level of Difficulty		
		Easy	*Moderate*	*Difficult*
1	Abstraction, Algorithm, Decomposition	O1		
2	Algorithm, Decomposition, Evaluation	O2		
3	Abstraction, Algorithm, Evaluation	O3		
4	Abstraction, Algorithm, Evaluation	O4		
5	Algorithm, Evaluation		O5	
6	Algorithm		O6	
7	Generalisation, Evaluation		O7	
8	Abstraction, Algorithm, Evaluation		O8	
9	Abstraction, Algorithm, Evaluation		O9	
10	Algorithm, Decomposition, Evaluation		O10	
11	Abstraction, Algorithm, Evaluation		O11	
12	Evaluation			O12
13	Desomposition, Generalisation, Evaluation			O13
14	Abstraction, Decomposition, Evaluation			O14
15	Abstraction, Generalisation, Evaluation			O15

V. FINDINGS AND DISCUSSION

A. Findings

The descriptive analysis for the pre-test showed that the UKPK score mean for treatment group is higher than the control group. The score means for the control group is M = 3.70 (SD = 0.58) meanwhile score means for the treatment group is M = 3.79 (SD = 0.59). Homogeneity analysis between the control group and the treatment group was also tested to determine whether there were differences in terms of CT skills before the intervention (treatment) was performed. This was done using an independent sample t-test and Levene test with the significance value of 0.05.

Table III shows that the results of the Levene test for homogeneity of variance for UKPK pre-test. Levene test of the significance value for UKPK pre-test is 0.653. Therefore, it can be assumed that the variance between control and treatment group is equal as the value p> 0.05.

Meanwhile the independent sample t-test showed there is no significant difference between the mean score of UKPK pre-test for control and treatment group. The result is presented in Table IV. This also explained that both control and treatment group are homogenic before the intervention take place.

The descriptive analysis for post-test showed that the UKPK score mean for treatment group is higher than the control group. Whereby the score means for control group, M is 3.74 (SD = 0.514). Meanwhile, the score means for treatment group, M is 3.88 (SD = 0.686). As for the score means for each of the CT skills for both groups, the result showed that three skills in the treatment group were higher than the control group. The three CT skills were decomposition, evaluation, and algorithm. The result is presented in Table V.

Levene test was conducted after the intervention take place. Table VI shows the results of the Levene test for homogeneity of variance for UKPK post-test. The significant value for the Levene test is 0.299 (p>0.05). Therefore, it can be assumed that the variance between control and treatment group is equal.

To test the hypothesis H_{01}, ANCOVA test was performed based on significant level of 0.05. The result is presented in Table VII.

TABLE III. LEVENE TEST FOR VARIANCE HOMOGENEITY (UKPK PRE-TEST)

Source	F	Sig.
UKPK (pre-test)	0.204	.653

TABLE IV. INDEPENDENT T-TEST (UKPK PRE-TEST)

Dependable Varaiable	T	Dk	Sig.	Mean difference
UKPK (pre-test)	-.576	63	0.567	-.084

TABLE V. THE SCORE MEANS FOR CT SKILLS

CT Skills	*Intervention*	*Post-Test*
Computational Thinking Skills Test (UKPK)	Mean (M) Control Group	Mean (M) Treatment Group
Decomposition	0.774	0.853
Evaluation	0.516	0.706
Abstraction	0.871	0.765
Algorithm	0.742	0.765
Generalisation	0.839	0.794

TABLE VI. LEVENE TEST FOR VARIANCE HOMOGENEITY (UKPK POST-TEST)

Dependable Varaiable	F	Sig.
UKPK (post-test)	1.097	.299

TABLE VII. RESULT OF ANCOVA TEST FOR UKPK POST-TEST

Dependable Varaiable	Sum of Square	dK	Mean Square	F	Sig.	Partial eta squared
UKPK (post-test)	0.338	1	0.338	0.896	0.347	0.014

Based on Table VII, there was no significant different for the UKPK post-test, F (65) = 0.896, p = 0.347, with a small effect size (partial $\eta 2 = 0.014$). Thus, the study failed to reject H_{01} hypothesis. However, based on the mean score of UKPK post-test, it was found that the treatment group outperformed the control group.

B. Discussion

Based on the findings, it could be said that the process of integrating CT skills does not take immediate effect. Findings from previous studies indicate that the process of applying computational thinking skills takes longer to obtain significant results, especially when using programming approach such as Scratch. This finding corroborates with a study conducted by [22] whereby allocating an hour in a week for CT intervention yields insignificant result. On the contrary, the studies by [23][24] allocated a longer period for CT intervention found to be significant. The effectiveness of a module might also being influence by the approach used by the teacher in carrying out the module. In this study, the approach used to teach CT was via plugged-in i.e. Scratch. The duration of training on the module and Scratch received by teachers in this study was short. Although [25] suggest CT training can be taught to teachers in a short training, it might work differently in this study especially for the teachers without computer science background. Although many studies have shown the effectiveness of using plugged-in approach to integrate CT skills [26], it is challenging for teachers without computer background to comprehend since CT components are closely related to computer science concept [27]. On another note, it could be argued that by delivering CT training via technology medium such as by using computer programming solely to learn CT might hinder one's awareness on the thought process of computational thinking while engaging with the programming activity. This could be one of the factors that contributed to the insignificant result. Scratch is just a tool to teach CT and not being able to realize that CT is a mental process and differentiate the affordances of Scratch in teaching CT will hinder the process of understanding the real concept of computational thinking [28]. In the end, students will end up creating Scratch projects but not being able to transfer those skills into different subject effectively. It is the role of a teacher to ensure students are aware of the thinking process involved while dealing with the activities in the module.

Although the finding of the study is insignificant, the findings of descriptive data show that the Matter Module is effective compared to conventional approach. This effectiveness can be seen through three of the five components of computational thinking integrated in the Matter Module have higher mean score. The findings of this study are in line with the study conducted by [29] for two CT skills, namely decomposition and algorithm. Whereas in this study, the decomposition, algorithms, and evaluations were achieved after using the Matter Module. The effectiveness of instilling CT skills in teaching and learning depending largely on the approach used by the teachers and the activity planned in associated to the CT skills. In this study, the decomposition skill was applied at the early stage by the participants in solving large problems to smaller problems. While algorithm skill was used in Scratch programming steps and evaluations were widely used throughout the Scratch project developed by the participants. It could be said that it is not a one size fits all as integrating CT skills in a learning activity is contextual. Although there are many CT concepts [5], not all can fit in with the content and learning activities. If the content is related to computer science or technical subjects, there are variety of CT concept can be integrated [30]. In the context of this study, the learning activities designed in this Matter module was more relevant to decompose, evaluation dan algorithm concept. This explained the higher means score of these three CT concepts as opposed to the other CT concepts integrated in the module, abstraction, and generalization.

VI. Conclusion

It can be concluded that integrating CT requires a thorough planning albeit the opportunity it can brings specially to inculcate problem solving skills among students to fulfill the demand for the future workforce. This study suggest that instilling computational thinking skills require longer time to demonstrate meaningful findings in most cases. It is crucial to properly consider which CT skills are suitable to the context of the learning environment. Contextual here might be referred to the type of learning activity, the nature of the subject matter, and the facilities to support the delivery of the teaching and learning. On an important note, integrating CT in teaching and learning to foster problem solving skills need to be viewed as a mental process rather than the use of computer solely. The future extension of this study shall be considering the allocation of intervention time to see how it will affect the effectiveness and fostering CT skills among students.

Acknowledgment

The authors extend their appreciation to the Dana Penyelidikan FPEND at National University of Malaysia, GG-2019-014 for funding this study.

References

[1] World Economic Forum, The Future of Jobs Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution, Growth Strategies, 2016.

[2] Kementerian Pendidikan Malaysia, Kurikulum Standard Sekolah Menengah, Dokumen Standard Kurikulum dan Pentaksiran Sains Tingkatan 3, Bahagian Pembangunan Kurikulum, Putrajaya: 2015.

[3] S. Kassan, K. F. Looi, and Y. M. Tham, Asas Sains Komputer Tingkatan 1, Kementerian Pendidikan Malaysia, Kuala Lumpur: 2016.

[4] Y. B. Kafai, and Q. Burke, "Computer Programming Goes Back to School," Phi Delta Kappan, vol. 95, pp. 61–65, September 2013.

[5] J. M. Wing, "Computational Thinking," Communications of the Association for Computing Machinery (ACM), vol. 49(3), pp. 33–35, March 2006.

[6] E. N. Caeli, and A. Yadav, "Unplugged Approaches to Computational Thinking: a Historical Perspective," TechTrends, vol. 64, pp. 29–36, July 2019.

[7] S. M. S. Azman, M. Arsat, and H. Mohamed, "The framework for the integration of computational thinking in ideation process," 2017 IEEE 6th International Conference on Teaching, Assessment and Learning for Engineering, Hong Kong, pp.61–65, December, 2017.

[8] F. Rani, M. Mohammad, and S. Suslistyo, "Pengembangan multimedia simulatif kimia berbasis inkuiri terbimbing pada materi analisis kualitatif kation golongan 1," Jurnal Inkuiri, vol. 4(3), pp. 120–126, 2015.

[9] U. A. Syed Noor, "An effective use of ICT for education and learning by drawing on worldwide knowledge, research and experience: ICT as a change agent for education (A Literature review)," Scholarly Journal of Education, vol. (2)4, pp. 38-45, April 2013.

[10] P. A. Ertmer, and A. T. Ottenbreit-Leftwich, "Teacher Technology Change: How Knowledge, Confidence, Beliefs, and Culture Intersect," Journal of Research on Technology in Education, vol. 42(3), pp. 255-284, September 2018.

[11] F. Kaleliog, "A new way of teaching programming skills to K-12 students : Code .org.," Computer in Human Behavior, vol. 52, pp. 200–210, November 2015.

[12] M. Maavak, and A. S. Ariffin, "Is Malaysia Ready for the Fourth Industrial Revolution?: The Automotive Sector as an i4.0 Springboard," in Brunet-Thornton, R., & Martinez, F. (Ed.), Analyzing the Impacts of Industry 4.0 in Modern Business Environments: IGI Global, 2018, pp. 41-64.

[13] A. P. Rehmat, H. Ehsan, and M. E. Cardella "Instructional strategies to promote computational thinking for young learners," Journal of Digital Learning in Teacher Education, vol. 36(1), pp. 46-62, February 2020.

[14] K. Samudin, K. F. Looi, and Y. M. Tham, "Asas Sains Komputer Tingkatan 1," Percetakan Rina Sdn. Bhd, Kuala Lumpur: 2016.

[15] U. L. Ling, T. C. Saibin, J. Labadin, and N. A. Aziz, "Preliminary Investigation: Teachers' Perception on Computational Thinking Concepts," Journal of Telecommunication, Electronic and Computer Engineering, vol. 9(2), pp. 2289–8131, May 2018.

[16] S. Senin, and N. M. Nasri, "Teachers' Concern towards Applying Computational Thinking Skills in Teaching and Learning," International Journal of Academic Research in Business and Social Sciences, vol. 9(1), pp. 296–310, February 2019.

[17] U. L. Ling, C.S Tammie, N. Nasrah, L. Jane, and A. A. Norazilah, "An Evaluation Tool To Measure Computational Thinking Skills: Pilot Investigation," Herald NAMSCA, vol. 1, pp. 606–614, September 2018.

[18] C. Samri, O. Kamisah, and A. N. Nazrul, "Level of computational thinking skills among secondary science student : Variation across gender and mathematics achievement," International Council of Association for Science Education, vol. 31(2), pp. 159–163, June 2020.

[19] R. E. Slavin, Educational Psychology : Theory And Practice, 8th ed.,. Boston: Pearson, 2006.

[20] S. Papert, Situating constructionism, Norwood, NJ: Ablex Publishing Corporation, 1991, pp.1-28.

[21] Y. Kafai, and M. Resnick, Constructionism in practice: Designing, thinking, and learning in a digital world, Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 1996.

[22] F. Kalelioğlu, and Y. Gülbahar, "The Effects of Teaching Programming via Scratch on Problem Solving Skills: A Discussion from Learners' Perspective," Informatics in Education, vol. 13(1), pp. 33–50, January 2014.

[23] Ö. Korkmaz, and A.Oluk, "Comparing students' Scratch skills with their computational thinking skills in terms of different variables," I.J. Modern Education and Computer Science, vol. 11, pp. 1–7, November 2016.

[24] J. M. Sáez-López, M. Román-González, and E. Vázquez-Cano, "Visual programming languages integrated across the curriculum in elementary school: A two year case study using "scratch" in five schools,", Computers & Education, vol. 97, pp. 129-141, June 2016.

[25] M. Bower, L. N. Wood, J. W. M. Lai, C. Howe, R. Lister, R. Mason, K. Highfield, and J. Veal "Improving the computational thinking pedagogical capabilities of school teachers," Australian Journal of Teacher Education, vol. 42(3), April 2017.

[26] X. Basogain, M. A. Olabe, J. C. Olabe and M. J. Rico, "Computational Thinking in pre-university Blended Learning classrooms," Computers in Human Behavior, vol. 80, pp. 412-419, March 2018.

[27] A. Alfayez, and J. Lambert, "Exploring Saudi Computer Science Teachers' Conceptual Mastery Level of Computational Thinking Skills," Computers in the Schools, vol. 36(3), pp. 143-166, August 2019.

[28] C. Kale, M. Akcaoglu, T. Cullen, D. Goh, L. Devine, N. Calvert, and K. Grise, "Computational What? Relating Computational Thinking to Teaching," TechTrends, vol. 62 (6), pp. 574–584, November 2018.

[29] S. Atmatzidou, and S. Demetriadis, "Advancing students' computational thinking skills through educational robotics: A study on age and gender relevant differences," Robotics and Autonomous Systems, vol. 75, pp. 661–670, January 2016.

[30] L. Shanmugam, S. F. Yassin, and F. Khalid, "Incorporating the elements of computational thinking into the Mobile Application Development Life Cycle (MADLC) model," International Journal of Engineering and Advanced Technology, vol. 8(5), pp. 815-824, June 2019.

Analysis of Steganographic on Digital Evidence using General Computer Forensic Investigation Model Framework

Muh. Hajar Akbar[1]
Master Program of Informatics
Universitas Ahmad Dahlan
Yogyakarta, Indonesia

Sunardi[2]*
Electrical Engineering Department
Universitas Ahmad Dahlan
Yogyakarta, Indonesia

Imam Riadi[3]
Information System Department
Universitas Ahmad Dahlan
Yogyakarta, Indonesia

***Abstract*—Steganography is one of the anti-forensic techniques used by criminals to hide information in other messages which can cause problems in the investigation process and difficulties in obtaining original information evidence on the digital crime. Digital forensic analysts are required ability to find and extract the messages that have been inserted by using proper tools. The purpose of this research is to analyze the hidden digital evidence using steganography techniques. This research uses the static forensics method by applying five stages in the Generic Forensics Investigation Model framework, namely pre-process, acquisition & preservation, analysis, presentation, and post-process as well as extracting files that have been infiltrated based on case scenarios involving digital crime. The tools used are FTK Imager, Autopsy, WinHex, Hiderman, and StegSpy. The results on the steganographic file insertion experiment of 20 files indicate that StegSpy and Hiderman are effective on the steganographic analysis of digital evidence. StegSpy can detect the presence of secret messages with 85% success rate. The extraction process using Hiderman for 18 files with containing steganographic messages had 100% successful.**

Keywords—Steganography; anti forensics; general computer forensic investigation model; hiderman

I. INTRODUCTION

Various kinds of crimes and criminal acts currently involve information and communication technology [1] [2]. The widespread of computers and other digital devices usage without security can lead various parties to crimes [3]. Perpetrators of crimes can be subject to punishment based on the evidence [4]. Digital criminals usually use anti-forensic techniques thus causing difficulties to find the digital evidence [5]. One of the anti-forensic techniques is steganography [6]. Steganography is an interesting science to study and research today [7]. Confidentiality, security, or integrity of the information to be conveyed are the main factors in steganography [8] [9]. This technique allows the perpetrator to hide information by inserting the information into other messages in the form of digital media such as text, images, audio or video without arousing suspicion. [10] [11]. Computer crimes related to the misuse of steganographic techniques have been reported through the mass media, including a report from Trend Micro November 2017 with the title "REDBALDKNIGHT's Daserf Backdoor Now Using Steganography". It has been reported that the Bronze Butler or Tick type malware was spread by the creator through a steganography technique by inserting it into an image with the extension jpg to spy on Japanese, South Korean, Russian, Singaporean and Chinese companies. Reported by Kompas.com December 9th, 2017 entitled "16 Years of 9/11 Attack: WTC Collapsed not because of a Plane Collision?". At that time, terrorists hide their terror activities in various digital media such as images, audios, and videos. The maps and photos of targets as well as orders for terrorist activity in sport chat rooms, porn bulletin boards, and other websites. The existence of cases reported by the mass media regarding crimes using steganography techniques inserted in electronic storage media. It's becomes a challenge that must be resolved by investigators and law enforcers in order to reveal the mode, objective, and perpetrators of crimes related to evidence obtained. Therefore, the process of steganography detection is very important for digital forensic investigators [12].

Digital forensics is a applied science to identify, extract, analyze, and present the evidence that has been stored on digital devices [13] [14], or help prevent illegal acts in the process of operating activities carried out [15] use generally accepted methods to make the evidence acceptable in court [16]. Forensic techniques and forensic analysis based on correct methods will have almost 100% success in collecting forensic data [17]. The process of digital forensic investigations on computers or similar devices can be carried out using live forensics or static forensics methods [18]. In this study, static forensic is used. Static forensic is an investigation carried out when the computer is turned off, because of the data can change when the computer is turned on [19]. The forensic framework can implement a framework of several standards that can be used in the forensic process according to international standards including the National Institute of Justice (NIJ), Digital Forensics Research Workshop (DFRWS), Integrated Digital Forensics Investigation Framework (IDFIF), Generic Computer Forensic Investigation Model (GCFIM), Systematic Digital Forensic Investigation Model (SRDFIM) or other forensic process frameworks [20].

The evidence is classified into two forms, namely electronic evidence and digital evidence [21]. Electronic evidence is physical evidence that can be recognized visually, so investigators and forensic analysts need to understand the

Corresponding Author

evidence when they are searching for evidence at a crime scene. While digital evidence is very vulnerable to changes in the data, therefore we need extra careful handling to keep digital evidence intact [22].

To make easier by investigators for data collecting related to the cases being investigated, forensic software is needed [23]. Forensic software usually multi-purpose, able to perform multiple tasks in the specific application. Computer forensic software complements the hardware available to law enforcement to obtain and analyze digital evidence gathered from suspect devices.

Research with a similar this topic has been conducted by [24] which is the investigation process and finds digital evidence in steganographic files. The process of steganographic analysis uses software, namely WinHex, InvisibleSecrets, and FTK Imager. The methodology or research stages are systematically carried out, namely literature review, observation & data collection, scenario case, preparation system, investigation & analysis case, and report & documentation.

Study with a similar theme was also carried out with the title Steganographic Engineering Analysis and Steganalysis on Multimedia Files Using the Net Tools and Hex Editor [25]. This research discusses use the WinHex application to perform analysis on messages hidden using the Net Tools into the container image. The method used experimental methods, namely identification problem, literature study, testing, and analysis.

The other reference in [26], steganographic file analysis was carried out by applying the Computer Forensic Investigative Process method which is divided into four stages, namely Acquisition, Identification, Evaluation, and Admission.

Further research was carried out by [27]. This research discusses the importance of computer forensic examiners in knowing the types of steganography tools that can be applied to the victim's computer. The tools used are S-tool and OpenStego.

Based on the background described, the objective of this digital forensics research is to find and analyze evidence in the form of files with text, audio, image, and video formats hidden by criminals by using steganography techniques. The static forensics method and GCFIM framework implemented in order to retrieving data on digital evidence, so that the data obtained can be used as legal evidence in court.

II. Research Method

A. Case Scenario

Digital evidence in this research will obtained from the results of case scenario as shown in Fig. 1.

B. Research Stages

The research was carried out in accordance with the work steps in the GCFIM framework which were added with one initial stage, namely implementation and case scenario. GCFIM describes the stages of research so that research steps can be known systematically and can be used as an investigative model for any digital investigation as shown in Fig. 2.

GCFIM has a back and forth flow, where it is possible for investigators to return to the previous stage due to the possibility of situations that can change such as the crime scene (both physical and digital), the investigation tools used, the crime tools used, and the investigator's level of expertise. The stages in the GCFIM framework are described as follows:

- Pre-Process. This stage is also called the preparation stage. Investigator doing related work before carrying out an investigation, such as preparing letters and official documents from legal authorities, and preparing tools.
- Acquisition & Preservation. At this stage, all relevant data are retrieved, stored, and prepared.
- Analysis. This stage is the main process in a computer forensic investigation, which is an analysis of the data that has been obtained to identify the source of the crime, the motive for the crime, and ultimately to find the person responsible for the crime.
- Presentation. This stage makes a presentation of the results that have been obtained to the competent authorities. This is important considering that the results of the analysis must not only be presented, but also must be supported by adequate/eligible and acceptable evidence. The results of this stage are to prove and/or deny the alleged criminal act.
- Post-Process. Digital and physical evidence must be returned to the rightful owner and stored in a safe place. The investigator reviews the investigation process that has been carried out so that it can be used to improve the further investigation process.

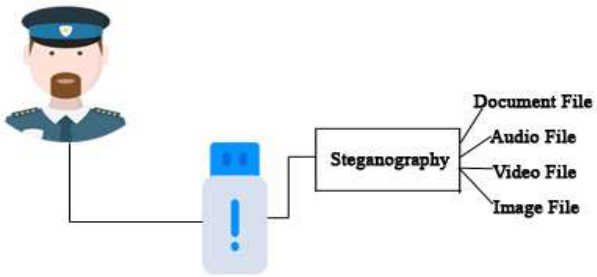


Fig. 1. Case Scenario.

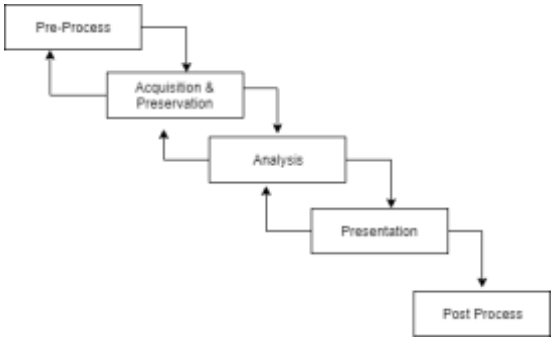


Fig. 2. GCFIM Framework [28].

III. Result and Discussion

A. Implementation Results and Case Scenario

The case scenario is implemented by using the Hiderman application. The hide files process is function to insert steganographic messages into several file formats such as documents, videos, images, and audio which are then stored on flash disk storage media. In this research, the inserted file in the form of stego text. The processing time to hide files is depends on the size of the file inserted. The larger the file size will longer time of insertion process. Fig. 3 is the process of hiding files.

The next step after selecting the container file is to select the files to be hidden or inserted by selecting the Choose the Files You Want to Hide menu as in Fig. 4.

The process in Fig. 4 is to select a secret file that will be inserted into the container file. In this process, the ratio of messages to be hidden can be found. A good ratio when hiding messages is 1 to 10. The hidden files must be 10 times smaller than the container files. After getting the right ratio file, the next step is to select the Hide File (s) menu.

B. Pre-process Results

At this stage, the things that must be prepared by the investigator can be seen in Table I.

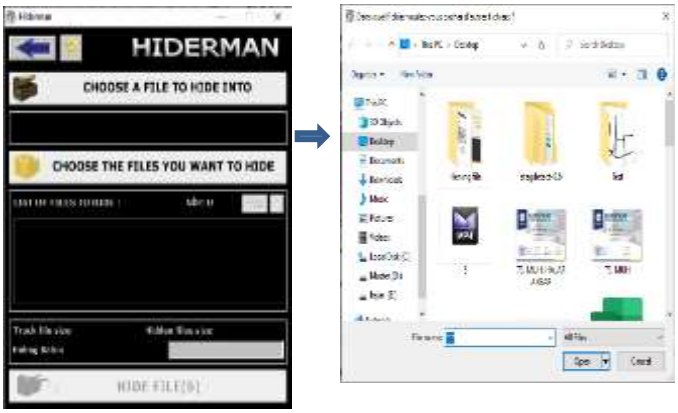


Fig. 3. Choosing the Container File.

Fig. 4. Secret File Selection.

TABLE I. Pre-Process Stage

No	Tool's name	Uses
1	Investigative administration	Search warrant and confiscation warrant
2	Digital camera	To photograph crime scenes and evidence by forensic photography
3	Stationery	To record technical specifications regarding electronic evidence and witness statements
4	Number, measuring scale, Institution label, blank label sticker	To mark each electronic evidence found at the crime scene
5	Chain of Custody Form	Report of the investigation of evidence

C. Acquisition and Preservation Results

This stage is the starting stage for the identification of evidence at the scene of the crime which is continued with the process of acquisition and maintenance of the originality of the evidence. The aim is to secure the evidence from changes in physical form or changes in data by storing it in a safe place. The data acquisition process on physical evidence (flash disk) is carried out using the FTK Imager tool. Choose the create disk image option and the physical drive option is selected for the full acquisition process. The source drive selection option is made with the name "Kingston Data Traveler 2.0". Choose the destination of the storage drive. Then select the image type with the Raw (dd) format. Fig. 5 and 6 is the process of create an image of evidence.

The acquisition results in two hash values, namely Message-Digest Algorithm 5 (MD5) and Secure Hash Algorithm 1 (SHA1) which are used to verify the authenticity of the duplicated image files. The hash value obtained by the recipient then compared with the hash value sent by the sender of the message to check the suitability and authenticity of the message. Fig. 7 is the log result and acquisition hash value on flash disk evidence using the FTK Imager tool.

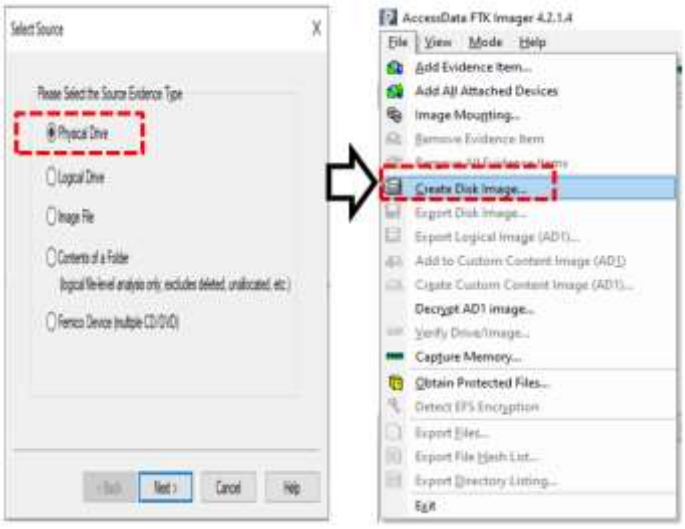


Fig. 5. Create Disk Image Process.

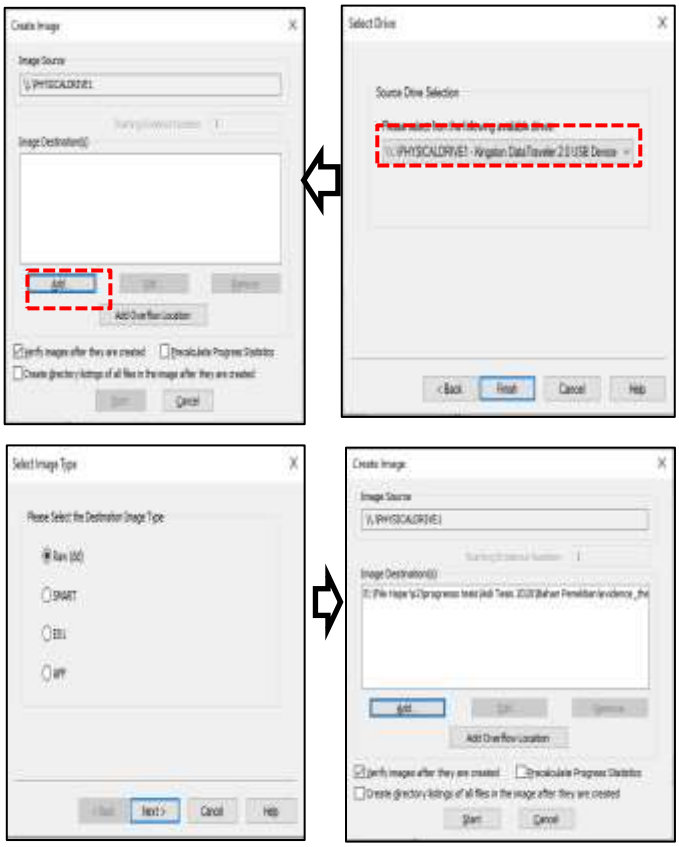


Fig. 6. Create an Image of Evidence.

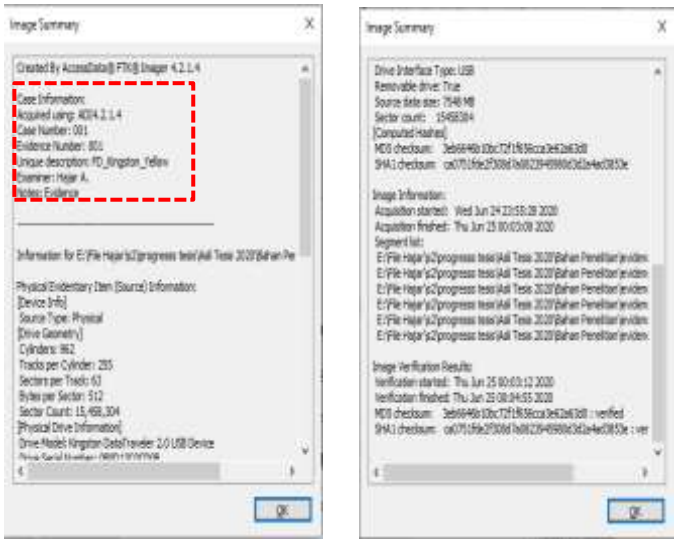


Fig. 7. Case Information and Hash Image.

Based on Fig. 6, the information regarding the MD5 hash value in the image file is "1Hw91XA9c1CuLKp9PhAt1ujZ963ZsagEBf", while the SHA1 hash value is "ca0751fde2f308d7a0823945980d3d2a4ad3853e". Furthermore, the preservation stage is carried out to prove the integrity of the acquired image file is identical to the file on the original evidence.

D. Preservation Results

This stage is retested by matching the MD5 and SHA1 hash values between the hash values of the original evidence and the evidence files of the acquisition or imaging results. Checking the hash value of original evidence is done using the Winhex tool. The MD5 and SHA1 hash values of the original evidence files can be seen as in Fig. 8.

After obtaining the hash value of the original evidence file, the next step is to match the hash value between the imaging evidence and the hash value of the original evidence which can be seen in Table II. The hash value of the acquisition/imaging evidence has the same value as the original evidence. Therefore, it can be concluded that the cloned evidence file is identical to the original evidence.

E. Results of Analysis

The analysis stage is divided into three stages, namely the identification stage, the steganalysis stage, and the extraction stage.

1) Identification stage: Analysis of the "image" file resulted from the acquisition & preservation process is carried out in this stage. The initial analysis process uses the Autopsy tool. Autopsy has several advantages for conducting content analysis and identification, data recovery, and metadata analysis. The process of input cases (case) on the Autopsy tool as shown in Fig. 9 as the initial stage of starting the "image" analysis phase.

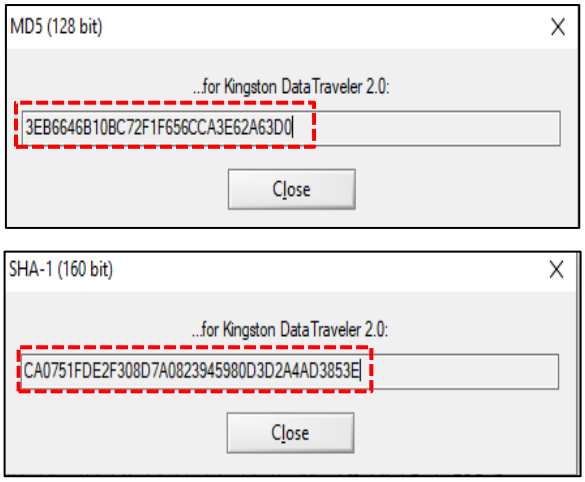


Fig. 8. Testing MD5 and SHA1 Hash Values.

TABLE II. HASH VALUE MATCHING

Original Evidence	
MD5	***SHA1***
3EB6646B10BC72F1F656CCA3E62A63D0	ca0751fde2f308d7a0823945980d3d2a4ad3853e
Image File	
3eb6646b10bc72f1f656cca3e62a63d0	ca0751fde2f308d7a0823945980d3d2a4ad3853e

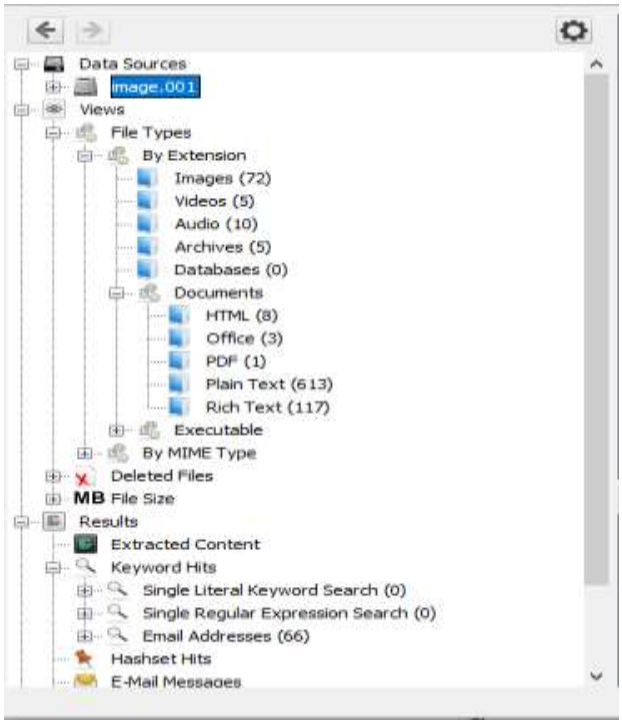


Fig. 9. Image Identification Stage.

Autopsy identify all the details of the data contained in the storage of evidence (flash disk) which has been neatly arranged and has become a data source in Autopsy. It is divided into several components including file types, deleted files, and file size. The file listing that is suspected of having confidential content is a file with the name of the audio, document, image, video folder, and one file in the .txt format which can be seen in Fig. 10.

Furthermore, the file extraction process is carried out in the suspected folder based on Fig. 10. The extraction process aims to obtain files so that re-analysis of the suspected file contents is carried out. The file extraction process can be seen in Fig. 11.

The file extraction process is carried out in order to export the image file based on the suspected folder. Files obtained after the extract process which consists of 4 folders and 1 file with the .txt format as listed in Fig. 11.

2) Steganalysis stage: The steganalysis process is carried out on the extracted files from the initial analysis to identify the files with secret messages that have been inserted. The second stage of the analysis process is shown as in Fig. 12 using the StegSpy tool in each extracted file.

The results of the analysis of the existence of secret files are shown in Table III. Based on the test results on 21 files, it was found that 18 files were identified to contain steganographic messages,

Based on Table III, StegSpy has successfully detected 18 steganographic files that have been inserted in various file formats and provided information about the detected marker values while three files were not detected.

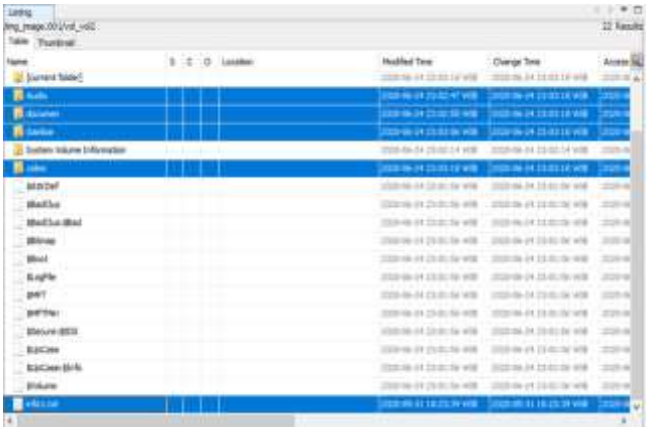


Fig. 10. Confidential Content.

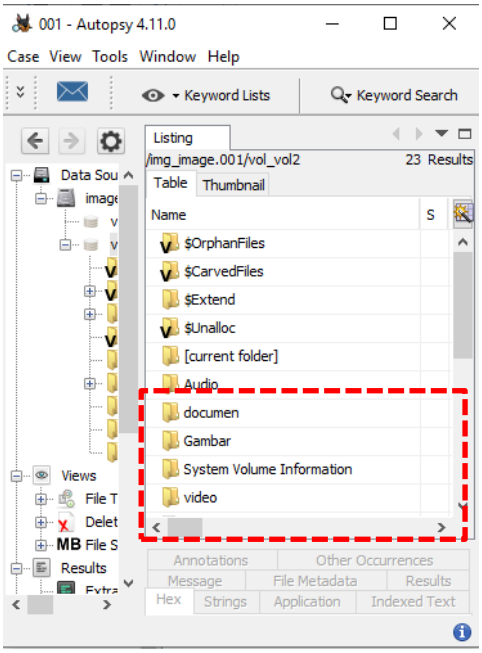


Fig. 11. The Extraction Stage.

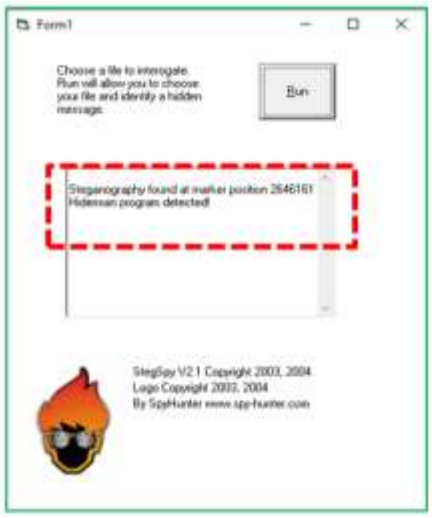


Fig. 12. Steganalysis Stage.

TABLE III. STEGANOGRAPHY FILE ANALYSIS RESULTS

No	File Type	File Name	Format	Information	Marker
1	Audio	evidence6	.wav	Found	2646161
		evidence7	.wav	Found	1073335
		evidence8	.wav	Found	2146284
		evidence9	.wav	Found	5226880
		evidence10	.wav	Found	10406854
2	Image	evidence1	.jpg	Found	1128234
		evidence2	.jpg	Found	5463476
		evidence3	.jpg	Found	785400
		evidence4	.jpg	Found	2546084
		evidence5	.jpg	Not found	-
3	Document	evidence16	.xls	Found	39655
		evidence17	.pdf	Found	941323
		evidence18	.ppt	Found	11093743
		evidence19	.doc	Found	93794
		evidence20	.txt	Found	374135
4	Video	evidence11	.mp4	Found	7832077
		evidence12	.mp4	Found	3517633
		evidence13	.mp4	Found	874744
		evidence14	.mp4	Not found	-
		evidence15	.mp4	Found	2769824
5	document	Info1	.txt	not found.	-

3) Extraction stage: The extraction stage is the analysis process carried out to reveal the presence of steganographic messages that have been detected in the steganalysis process. Based on the extraction results in the previous stage, after observation, there is a file with the file name info1.txt which contains information as in Fig. 13.

Based on Fig. 12, the file with the name info1 in the .txt format is suspected to be the key used to open the secret message contained in the detected file. Furthermore, at this stage an analysis is carried out using the Hiderman forensic tool to decrypt the steganography file using the "trial" key. The process of encrypting steganography files can be seen in Fig. 14.

After selecting a file that is infiltrated with steganographic messages, the next step as shown in Fig. 15, is to select the extract data menu and determine the place where the extracted file is stored.

After the key input process is done, the hidden secret files can be discovered automatically. The secret file obtained is in the form of a .txt text message as shown in Fig. 16.

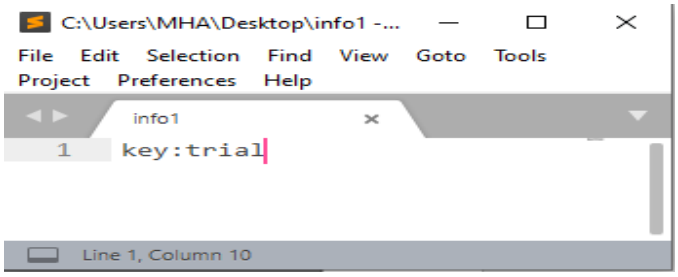


Fig. 13. Info1.txt.

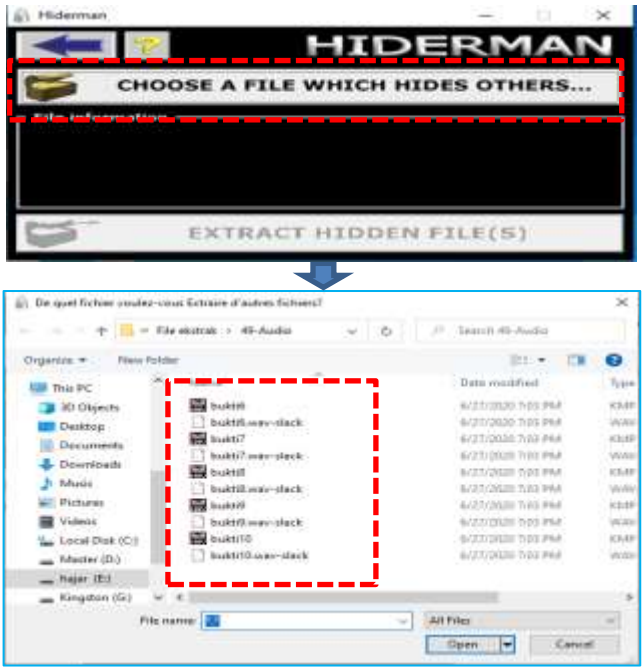


Fig. 14. Selection of the Inserted File.

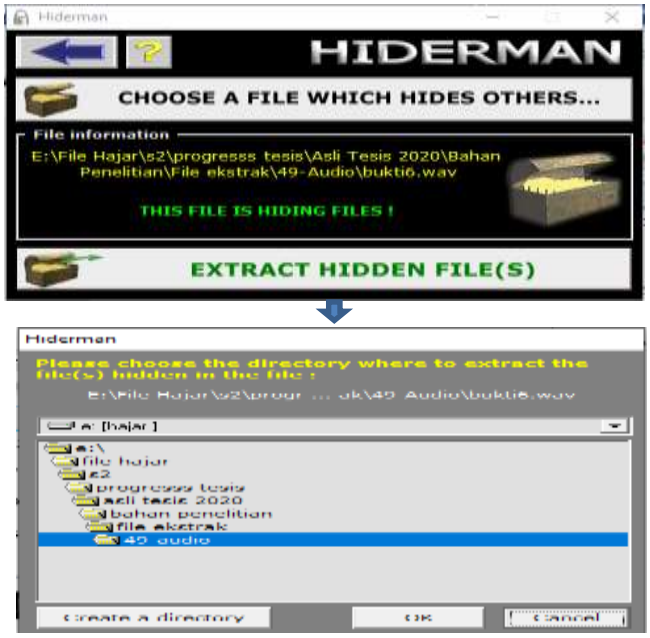


Fig. 15. Extraction Process.

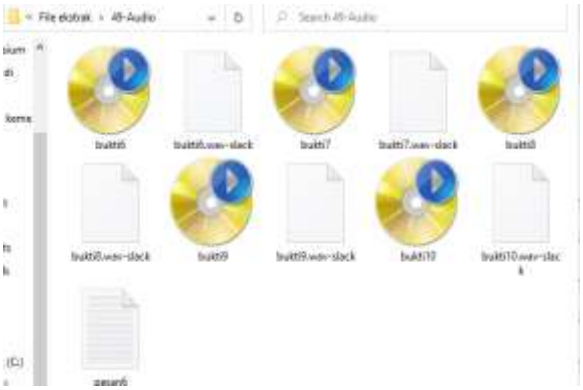


Fig. 16. Extraction Process and Directory Selection.

The final step in the extraction process is to enter the key or password found based on the contents of the info1 file, as shown in Fig. 17.

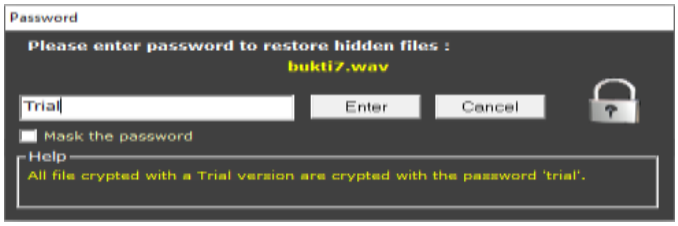


Fig. 17. Confidential Password Files were Found.

Information regarding the confidential files that have been found is shown in Table IV.

F. Presentation Results

After the analysis process on digital evidence was carried out using Stegspy and Hiderman, digital evidence was obtained on the flash disk image file as in Table V. Based on the process of detecting and extracting digital evidence, secret messages regarding delivery schedules are found.

TABLE IV. CONFIDENTIAL FILE INFORMATION FOUND

Message Name	Format	Size	Hash (MD5)
Message1	.txt	1 kb	AADCCD6FD16370F7DDB14DFAEE213BB0
Message2	.txt	1 kb	72C2E79FBA5225F3C5BE3F734795EADF
Message3	.txt	1 kb	A86BB86287E005045AF4C4AD32650732
Message4	.txt	1 kb	94305EEE69C737D258A5B81646F328A9
Message5	.txt	1 kb	A4871E4D6B386E29AF2FCD2025189753
Message6	.txt	1 kb	F500126EE1EB22DD402B8100556EBE95
Message7	.txt	1 kb	2C35E96B69903883A6731F838699309A
Message8	.txt	1 kb	BD41F794F04DEA470F1ACD38FD05877D
Message9	.txt	1 kb	FA1ABD78B1A8D5FD06E7EC36EE18AF6A
Message10	.txt	1 kb	A5C0B0FE889FC53B4CC4AEB7E97831AF
Message11	.txt	1 kb	FF73BBFDBBCB24B59A01FABF2C15ADBD
Message12	.txt	1 kb	D4E4D9DD9B2F1561C3B0D1C02DC34A85
Message13	.txt	1 kb	2D537F93BBB4D61857378E7403D9BA4A
Message14	.txt	1 kb	1Hw91XA9c1CuLKp9PhAt1ujZ963ZsagEBf
Message15	.txt	1 kb	C4ABF8E5D2A505B6A5F6CB2AD98E3795
Message16	.txt	1 kb	1Hw91XA9c1CuLKp9PhAt1ujZ963ZsagEBf
Message17	.ppt	553 kb	5A9A559EE1C31B8A1E0B60BB9164B053
Message18	.doc	16.2 kb	954D8897DE0774600DBD9356229575CA
Message19	.pdf	46.4 kb	25C6796DE638FB818825384BED0D539B
Message20	.pdf	365 kb	D6BF7444584D42C78E5477599188E071

TABLE V. PASTED MESSAGE

Container	Secret file	Size	Message
Evidence1.jpg	Message1.txt	1.07 mb	Monday, January 6, 2020. at 02.30.
Evidence2.jpg	Message2.txt	5.20 mb	Sunday, January 12 2020. Delivery at 04.30.
Evidence3.jpg	Message3.txt	766 KB	Thursday, January 30, 2020. Delivery at 23.30.
Evidence4.jpg	Message4.txt	2.42 MB	Saturday, March 14, 2020. Delivery at 19.30.
evidence6.wav	Message6.txt	2.5 MB	Thursday, April 16, 2020. Delivery at 20.30.
evidence7.wav	Message7.txt	1.02 MB	Saturday, April 18 2020. Delivery at 22.30.
evidence8.wav	Message8.txt	2.04 MB	Tuesday, April 28, 2020. Delivery at 17.30.
evidence9.wav	Message9.txt	4.98 MB	Friday, May 1, 2020. Delivery at 15.30.
evidence10.wav	Message10.txt	9.92 MB	Friday, 15 May 2020. Delivery at 15.30.
Evidence11.mp4	Message11.txt	7.46 MB	Monday, 25 May 2020. Delivery at 12.30.
Evidence12.mp4	Message12.txt	3.35 MB	Monday, 25 May 2020. Delivery at 12.30.
Evidence13.mp4	Message13.txt	854 KB	Sunday, 31 May 2020. Delivery at 13.30.
Evidence15.mp4	Message15.txt	2.64 MB	Wednesday, 3 June 2020. Delivery at 19.30.
Evidence11.mp4	Message11.txt	7.46 MB	Monday, 25 May 2020. Delivery at 12.30.

IV. CONCLUSION

The analysis process uses the static forensics method with the Generic Computer Forensic Investigation Model framework successfully implemented. The secret message that has been inserted using steganography technique was found steganographic messages in the form of stegotext. The success rate of the StegSpy forensic tool based on the detection process of digital evidence containing an average of 85% steganography and 15% unknown files. The accuracy of the Hiderman tool based on digital evidence that has been successfully extracted is 100%.

ACKNOWLEDGMENT

This research is supported by Directorat Riset dan Pengabdian Masyarakat, Direktorat Jenderal Penguatan Riset dan Pengembangan Kementerian Riset, Teknologi dan Pendidikan Tinggi Republik Indonesia. Surat Kontrak Pelaksanaan Penelitian Kementerian Riset dan Teknologi/Badan Riset dan Inovasi Nasional (KEMENRISTEK/BRIN) Tahun Tunggal Tahun Anggaran 2020 Nomor: PTM-027/SKPP.TT/LPPM UAD/VI/2020.

REFERENCES

[1] B. K. Payne and L. Hadzhidimova, "Disciplinary and interdisciplinary trends in cybercrime research: An examination," Int. J. Cyber Criminol., vol. 14, no. 1, pp. 81–105, 2020, doi: 10.5281/zenodo.3741131.

[2] A. Kurniawan, I. Riadi, and A. Luthfi, "Forensic analysis and prevent of cross site scripting in single victim attack using open web application security project (OWASP) framework," J. Theor. Appl. Inf. Technol., vol. 95, no. 6, pp. 1363–1371, 2017.

[3] L. L. Alaydrus and D. Nusraningrum, "Impact of Computer Misuse in the Workplace," KnE Soc. Sci., vol. 2020, pp. 1–7, 2020, doi: 10.18502/kss.v4i7.6838.

[4] A. Yudhana, I. Riadi, and F. Ridho, "DDoS classification using neural network and naïve bayes methods for network forensics," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 11, pp. 177–183, 2018, doi: 10.14569/ijacsa.2018.091125.

[5] R. Sabillon, J. Serra-Ruiz, V. Cavaller, and J. J. Cano, "Digital forensic analysis of cybercrimes: Best practices and methodologies," Int. J. Inf. Secur. Priv., vol. 11, no. 2, pp. 25–37, 2017, doi: 10.4018/IJISP.2017040103.

[6] A. Jain and G. S. Chhabra, "Anti-forensics techniques: An analytical review," 2014 7th Int. Conf. Contemp. Comput. IC3 2014, no. August 2014, pp. 412–418, 2014, doi: 10.1109/IC3.2014.6897209.

[7] L. Widyawati, I. Riadi, and Y. Prayudi, "Comparative Analysis of Image Steganography using SLT, DCT and SLT-DCT Algorithm," MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput., vol. 20, no. 1, pp. 169–182, 2020, doi: 10.30812/matrik.v20i1.701.

[8] M. Khalid, K. Arora, and N. Pal, "A Crypto-Steganography: A Survey," Int. J. Adv. Comput. Sci. Appl., vol. 5, no. 7, pp. 149–155, 2014, doi: 10.14569/ijacsa.2014.050722.

[9] I. Riadi, A. W. Muhammad, and Sunardi, "Neural network-based ddos detection regarding hidden layer variation," J. Theor. Appl. Inf. Technol., vol. 95, no. 15, pp. 3684–3691, 2017.

[10] M. Dalal and M. Juneja, "Video steganalysis to obstruct criminal activities for digital forensics : a survey," vol. 10, no. 4, pp. 338–355, 2018.

[11] T. Sloan and J. Hernandez-Castro, "Forensic analysis of video steganography tools," PeerJ Comput. Sci., 2015, doi: 10.7717/peerj-cs.7.

[12] S. Rathore, "Steganography: Basics and Digital Forensics," Int. J. Sci. Eng. Technol. Res., vol. 4, no. 7, pp. 2589–2593, 2015.

[13] K. K. Sindhu and B. B. Meshram, "Digital Forensics and Cyber Crime Datamining," J. Inf. Secur., 2012, doi: 10.4236/jis.2012.33024.

[14] A. Iswardani and I. Riadi, "Denial of service log analysis using density K-means method," J. Theor. Appl. Inf. Technol., vol. 83, no. 2, pp. 299–302, 2016.

[15] Sunardi, I. Riadi, and A. Sugandi, "Forensic analysis of Docker Swarm cluster using GRR Rapid Response framework," Int. J. Adv. Comput. Sci. Appl., 2019, doi: 10.14569/ijacsa.2019.0100260.

[16] K. K. Sindhu and B. B. Meshram, "Digital Forensic Investigation Tools and Procedures," Int. J. Comput. Netw. Inf. Secur., vol. 4, no. 4, pp. 39–48, 2012, doi: 10.5815/ijcnis.2012.04.05.

[17] I. Riadi, R. Umar, and I. M. Nasrulloh, "Digital Forensic Analysis On Frozen Solid State Drive Using National Institute of Justice (NIJ) METHOD," Elinvo (Electronics, Informatics, Vocat. Educ., 2018, doi: 10.21831/elinvo.v3i1.19308.

[18] M. Rafique and M. N. A. Khan, "Exploring Static and Live Digital Forensics: Methods, Practices and Tools," Int. J. Sci. Eng. Res., 2013.

[19] Sunardi, I. Riadi, and M. H. Akbar, "Application of Static Forensics Method for Extracting Steganographic Files on Digital Evidence Using the DFRWS Framework," Rekayasa Sist. dan Teknol. Inf. (RESTI), vol. 4, no. 3, pp. 576–583, 2020.

[20] A. Yudhana, I. Riadi, and I. Anshori, "Facebook Messenger Digital Evidence Analysis Using Nist Method," IT J. Res. Dev., 2018, doi: 10.25299/itjrd.2018.vol3(1).1658.

[21] D. Mugisha, "DIGITAL FORENSICS : Digital Evidence in judicial System," no. April, 2019.

[22] S. H. Belshaw, "Next Generation of Evidence Collecting: The Need for Digital Forensics in Criminal Justice Education," J. Cybersecurity Educ. Res. Pract., vol. 1, no. 3, 2019.

[23] E. K. J. Melanie, M. V. Naseri, and N. A. B. Sabri, "Image forensics tool with steganography detection," J. Crit. Rev., vol. 7, no. 3, pp. 130–134, 2020, doi: 10.31838/jcr.07.03.24.

[24] A. P. Saputra and N. Widiyasono, "Forensic Digital Analysis of Steganographic Files (Case study: Drug Trafficking)," J. Tek. Inform. dan Sist. Inf., 2017, doi: 10.28932/jutisi.v3i1.594.

[25] Y. B. Utomo and D. Erwanto, "Analisa Teknik Steganografi dan Steganalysis Pada File Multimedia Menggunakan Net Tools dan Hex Editor," Gener. J., vol. 3, no. 1, pp. 16–22, 2019, doi: 10.29407/gj.v3i1.12698.

[26] V. A. Silalahi and I. Sembiring, "Digital Forensics Investigation Analysis on Digital Steganographic Evidence," Artik. Ilm., 2017.

[27] I. A. Yari and S. Zargari, "An Overview and Computer Forensic Challenges in Image Steganography," in Proceedings - 2017 IEEE International Conference on Internet of Things, IEEE Green Computing and Communications, IEEE Cyber, Physical and Social Computing, IEEE Smart Data, iThings-GreenCom-CPSCom-SmartData 2017, 2018, doi: 10.1109/iThings-GreenCom-CPSCom-SmartData.2017.60.

[28] Y. Yusoff, R. Ismail, and Z. Hassan, "Common Phases of Computer Forensics Investigation Models," Int. J. Comput. Sci. Inf. Technol., vol. 3, no. 3, pp. 17–31, 2011, doi: 10.5121/ijcsit.2011.3302.

Multi-Verse Algorithm based Approach for Multi-criteria Path Planning of Unmanned Aerial Vehicles

Raja Jarray[1], Soufiene Bouallègue[1, 2]
Research Laboratory in Automatic Control (LARA)
National Engineering School of Tunis (ENIT)
University of Tunis EL MANAR, BP 37, Le Belvédère, 1002 Tunis, Tunisia[1]
Higher Institute of Industrial Systems of Gabès (ISSIG), 6011 Gabès, Tunisia[2]

***Abstract*—In this paper, a method based on a Multiobjective Multi-Verse Optimizer (MOMVO) is proposed and successfully implemented to solve the unmanned aerial vehicles' path planning problem. The generation of each coordinate of the aircraft is reformulated as a multiobjective optimization problem under operational constraints. The shortest and smoothest path by avoiding all obstacles and threats is the solution of such a hard optimization problem. A set of competitive metaheuristics such as Multiobjective Salp Swarm Algorithm (MSSA), Grey Wolf Optimizer (MOGWO), Particle Swarm Optimization (MOPSO) and Non-dominated Sorting Genetic Algorithm II (NSGA-II) are retained as comparison tools for the problem's resolution. To assess the performance of the reported algorithms and conclude about their effectiveness, an empirical study is firstly performed for solving different multiobjective test functions from the literature. These algorithms are then used to obtain a set of optimal Pareto solutions for the multi-criteria path planning problem. An efficient Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) of Multi Criteria Decision-Making (MCDM) model is investigated to find the optimal solution from the non-dominant ones. Demonstrative results and statistical analysis are presented and compared in order to show the effectiveness of the proposed MOMVO-based path planning technique.**

Keywords*—Unmanned aerial vehicles; path planning problem; multiobjective optimization; multiobjective multi-verse algorithm; decision-making model; nonparametric statistical tests

I. Introduction

The Unmanned Aerial Vehicles (UAVs) have shown their commitment in various military and civil applications [1, 2]. The problem of paths planning, especially within a flight environment with threats and obstacles, is one of the key elements in the framework of design and control of these aerial robots. Such a complex problem can be seen and treated as a hard optimization problem under operational constraints [3]-[15]. The resolution of such a problem allows the unmanned aircraft to autonomously calculate the optimal or almost optimal path from the starting point to the target, based on the requirements and constraints of the activity.

In the literature, various approaches and techniques have been proposed to solve such kind of complex optimization problems. The graph-based techniques are extensively adopted and show some effective advantages. The well-known Voronoi diagram searching method [3], rapidly-exploring random graph algorithm [4], A* algorithm [5], D* Lite approach [6] and artificial potential field algorithm [7] are the main used approaches. Often, it is difficult to consider the movement limitations of the UAVs in these types of planning methods, which means that they cannot normally be used within practical situations [8]. Another category of path planning methods, such as the population-based scalable algorithms, can overcome these shortcomings. As an example, the authors in [9] used the Genetic Algorithm (GA) to solve the shortest path problem in order to scan large agricultural lands and collect data. The authors in [10] developed an algorithm that uses the well-known Particle Swarm Optimization (PSO) method to solve the trajectory planning problem for multiple UAVs in a receding horizon framework. In [11], a two new hybrid metaheuristics that combine the PSO method both with the genetic algorithm and harmony search algorithm have been proposed to solve the UAVs' path planning problem. In [12], the authors have solved the UAVs' path planning problem based on a recent global metaheuristic named Grey Wolf Optimization (GWO). In [13], an improved Ant Colony Optimization (ACO) algorithm has been given by introducing the metropolis criterion into the node filtering mechanism in order to generate the initial trajectory and avoid the risk of falling into the optimal local solution and stagnation. In [14], a modified central force optimization based method has been introduced to address the rotary wing vertical take-off and landing aircraft trajectory planning. The authors in [15] have presented a 3D path planning algorithm based on an adaptive sensitivity decision operator associated with a PSO method.

In addition, most real path planning problems need to be solved by considering different conflicting goals such as price and quality. The conflicting objectives must be addressed simultaneously and the weighted based methods are usually used [14, 16]. Nevertheless, it is difficult to determine the relationship between the weighting factors. Therefore, these objectives should be treated by multiobjective metaheuristics which are applied in many others domains [17, 18]. The idea of using multiobjective optimization concepts for path planning problem formulation and resolution seems a promising solution and it has been exploited in this work.

In [19], a Multiobjective Genetic Algorithm (MOGA) based method has been used to solve the complex path planning problems implying a mission of UAVs and a set of ground control stations. Many objectives have been optimized such as the makespan, the fuel consumption, and covered distance. In [20], the authors developed an improved

Multiobjective Particle Swarm Optimization (MOPSO) algorithm to find collision-free and feasible paths with various minimum factors such as altitude, length and angle variable rate. The authors in [21] have improved a Non-dominated Sorting Genetic Algorithm III (NSGA-III) by adding adaptive genetic operators in the offspring population generation to solve the path planning problems. In [22], an improved multiobjective ACO algorithm has been adopted in which the objective function for optimization is formulated to make UAV drone following a short, safe and smooth path. Such an algorithm assumes that the environment is known in advance. The authors in [23] have used a safety index map (SIM) to catch obstacles in the geography map. Then, a multiobjective path planning approach based on a Crowding Distance NSGA-II (CDNSAG-II) metaheuristic is proposed while considering both path length and safety as the main flight objectives. In [24], the path planning problem has been modeled as a problem with high complexity involving several tasks. Such a modeling approach presented high convergence rate for multiobjective solvers. The authors have used a weighted random generator that can concentrate the search on potentially better regions of the solution space to reduce the convergence rate of the used Multiobjective Evolutionary Algorithm (MOEA) solver. The authors in [25] have solved a multi-UAVs' trajectory planning problem using the concepts of MOPSO metaheuristic.

Based on the aforementioned studies, and regarding the drawbacks of the cited methods especially in terms of complexity and time consuming, the main contribution of this paper is the development of a novel strategy of reformulation and solving of a multi-criteria path planning problem under operational constraints based on a recent and unified MOMVO algorithm. The proposed MOMVO-based method allows the UAVs to autonomously calculate the optimum or near optimal path from the starting point to the target while avoiding all threats and obstacles considered in the flight environment. The choice of a solution among all the optimal Pareto ones requires a higher-level decision-making approach. The Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) is proposed for this path planning methodology. The demonstrative results are presented, statistically analyzed and compared with each other as well as with those obtained by the competitive MOPSO, NSGA-II, MSSA and MOGWO algorithms.

The reminder of this paper is organized as follows. In Section II, the flight environment is topologically modeled and a multiobjective reformulation of the UAVs' path planning problem is derived. This section is ended by the description of the proposed offline path planning algorithm. Section III presents the description of the proposed multiobjective multi-verse algorithm MOMVO as well as its adaptation to solve the reformulated multi-criteria path planning problem. A pseudo-code of such a multiobjective algorithm is given for the soft implementation. In Section IV, numerical simulations are carried out and discussed to show the effectiveness of the proposed MOMVO-based path planning approach. Several comparisons with a set of competitive algorithms are made through this study. Section V concludes the paper.

II. Path Planning Problem Formulation

A. Flight Environment Modeling

In robotics, the path planning procedure is the creation of a plan to guide a UAV, similar to a moving object in the space, from the starting point S to the destination one P with the spatial coordinates denoted as $(x_S, y_S, z_S) = (x_1, y_1, z_1)$ and $(x_P, y_P, z_P) = (x_n, y_n, z_n)$, respectively. The navigation model used in this study is shown in Fig. 1. In a real navigation environment, it is very challenging to define the geometric coordinates of the obstacles and threats of the UAV drone. In this work, a danger zone is characterized by a cylinder model supposed to be static in the 3D flight environment as depicted in Fig. 2.

The x-axis range of the flight space is divided into $n-1$ equal segments which are denoted as $x_1, x_2, x_3, \ldots, x_n$. The perpendicular planes $(L_1, L_2, L_3, \ldots, L_n)$ are passed by these corresponding division points. By taking a waypoint $\boldsymbol{w}_i = (x_i, y_i, z_i)$ at each plane L_i and forming a waypoints' sequence $\boldsymbol{W} = \{S, (x_2, y_2, z_2), \ldots, (x_{n-1}, y_{n-1}, z_{n-1}), P\}$, a flight path is then generated by connecting all these waypoints. In this study, the problem of path planning is solved by optimizing the series of the waypoints in order to generate a shorter and smoother path from the starting point S to the goal point P while avoiding the existing obstacles and threats. Based on the cubic Spline interpolation method, these waypoints are connected to obtain the desired smooth path. In this path's modeling strategy, the x-coordinates of all waypoints are know in advance but those of the y- and z-axis have to be determined. Subsequently, the generation of each coordinate waypoint $\boldsymbol{w}_i = (x_i = x_{known}, y_i, z_i)$,, $i = 2,3,\ldots,n-1$, is formulated as a multiobjective optimization problem with the decision variables $\boldsymbol{\theta} = \{y_i, z_i\}$ and under operational nonlinear and complex constraints. In this mathematical formulation, the variables y_i and z_i denote the y- and z-coordinates of the ith waypoint, respectively.

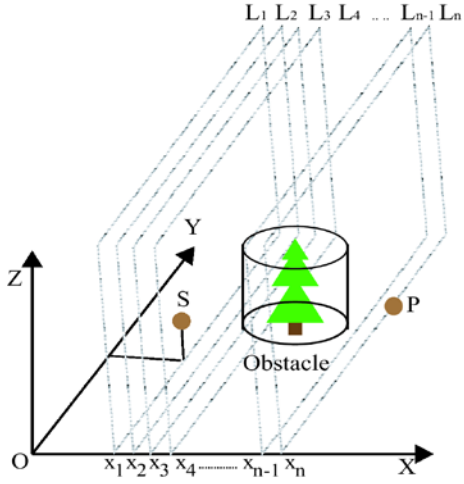


Fig. 1. Geometry of the UAV's Navigation Space.

B. Problem Formulation

In the mathematical multiobjective optimization formalism, a generic constrained problem is defined as follows [26]:

$$\begin{cases} \underset{\boldsymbol{\theta}\in \boldsymbol{D}\subseteq \mathbb{R}^q}{\text{Minimize}}\, F(\boldsymbol{\theta}) = \{ f_1(\boldsymbol{\theta}),\ f_2(\boldsymbol{\theta}), \dots f_M(\boldsymbol{\theta})\} \\ \text{s.t:} \\ g_v(\boldsymbol{\theta}) \le 0 \quad v = 1,2,...,V \\ h_w(\boldsymbol{\theta}) = 0 \quad w = 1,2,...,W \\ \boldsymbol{\theta} \in \boldsymbol{D} \subseteq \mathrm{R}^q \end{cases} \tag{1}$$

where $f_m : \mathbb{R}^q \to \mathbb{R}$, for $m = 1,2,\dots,M$, denote the objective functions to be minimized, $\boldsymbol{D} = \{\boldsymbol{\theta} \in \mathbb{R}^q, \boldsymbol{\theta}_{\min} \le \boldsymbol{\theta} \le \boldsymbol{\theta}_{\max}\}$ is the bounded search domain, $g_v : \mathbb{R}^q \to \mathbb{R}$ and $h_w : \mathbb{R}^q \to \mathbb{R}$ are the inequality and equality constraints, respectively, $q \in \mathbb{N}$ is the dimension of the optimization problem, i.e. the number of decision variables.

For the UAVs' navigation process, the length of the planned path is very important in the case of several missions. A shorter path can extend the life of an UAV and remains desirable in all planning problems. The criteria which can be considered for the path planning process are especially related to the path length and drone's attitude. According to this flight specification and for a given ith waypoint, the related objective function to be minimized in problem (1) can be formulated as follows:

$$f_{1,i}(\boldsymbol{\theta}) = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2} + \sqrt{(x_i - x_n)^2 + (y_i - y_n)^2 + (z_i - z_n)^2} \tag{2}$$

where (x_n, y_n, z_n) denotes the coordinates of the destination point P, n is the number of the waypoints and $\boldsymbol{\theta} = \{y_i, z_i\}$ are the unknown y- and z-axis coordinates of the given ith waypoint, respectively.

In addition, the dynamical characteristics of a given UAV cannot be completely ignored. In order to limit the straightness of the path, the angles between two adjacent segments' $\vec{\varphi}$ and $\vec{\psi}$ are introduced. This path planning specification is modeled by the following cost function:

$$f_{2,i}(\boldsymbol{\theta}) = \arccos\left(\frac{\vec{\varphi}\vec{\psi}}{|\vec{\varphi}||\vec{\psi}|}\right) \tag{3}$$

where $\vec{\varphi}$ means $(x_{i-2} - x_{i-1}, y_{i-2} - y_{i-1}, z_{i-2} - z_{i-1})$ and $\vec{\psi}$ denotes $(x_i - x_{i-1}, y_i - y_{i-1}, z_i - z_{i-1})$.

The collision avoidance is essential for any path planning task. In its own navigation process, a drone cannot in any case cross the threat zones and/or fly over them in order to avoid the risk of being detected by the radars or missiles within a military application. Such an avoidance specification is modeled by the following family of nonlinear and hard constraints:

$$g_{1,i}(\boldsymbol{\theta}) = r_t + \delta_{\min} - \sqrt{(x_i - x_t)^2 + (y_i - y_t)^2} \le 0 \tag{4}$$

where (x_t, y_t, r_t) is the coordinates of the static threat zone, (x_t, y_t) presents the center on the XOY flight plan, r_t is the radius of a given obstacle and $\delta_{\min}$ is the safety distance defined as shown in Fig. 2.

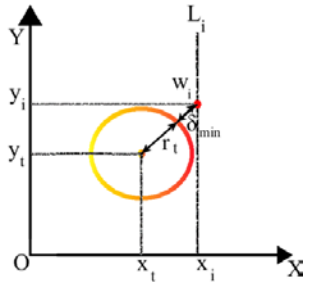


Fig. 2. Illustration of the Threat Zones in Flight Space.

Considering all these defined objectives and constraints, the formulated multiobjective optimization problem for the UAV's path planning according to a given ith waypoint is defined as follows:

$$\begin{cases} \underset{\boldsymbol{\theta}\in \boldsymbol{D}\subseteq \mathbb{R}^2}{\text{Minimize}}\, F(\boldsymbol{\theta}) = \{ f_{1,i}(\boldsymbol{\theta}),\ f_{2,i}(\boldsymbol{\theta})\} \\ \text{s.t}: \\ g_{1,i}(\boldsymbol{\theta}) \le 0 \\ \boldsymbol{\theta} \in \boldsymbol{D} \subseteq \mathrm{R}^2 \end{cases} \tag{5}$$

where $f_{1,i}(.)$, $f_{2,i}(.)$ and $g_{1,i}(.)$ are the cost and constraint functions given in (2), (3) and (4), respectively, $\boldsymbol{\theta} = \{y_i, z_i\}$ is the decision variable of the problem.

To handle with the operational constraints of problem (5), the following static penalty function is used [27]:

$$\phi_{m,i}(\boldsymbol{\theta}) = f_{m,i}(\boldsymbol{\theta}) + \sum_{v=1}^{V} \lambda_{v,i} \max\{0, g_{v,i}(\boldsymbol{\theta})\}^2 \tag{6}$$

where $\lambda_v \in \mathbb{R}^+$ is the weighting coefficient associated to the vth constraint, V is the total number of the inequality types of constraints and $m = 1,2,\dots,M$.

C. Proposed Planning Procedure

In the multiobjective optimization framework, many methods have been developed for the selection of an optimal solution from a given set of Pareto non-dominated ones [28, 29]. In this work, the technique for order of preference by similarity to ideal solution TOPSIS is used to make decision about the optimal solution for problem (5). Algorithm 1 is so proposed for the complete resolution of the UAV's path planning problem (5).

Algorithm 1: offline path planning algorithm

Step 1: Initialization

Initialize the starting point (x_1, y_1, z_1), the destination point (x_n, y_n, z_n), the waypoints matrix $P = [x_1; y_1; z_1]$. Divide the x-axis range into $n-1$ equal portions and set the vector $Px = [x_1\ x_2 \dots x_i \dots x_n]$.

Step 2: Generation of the waypoints

For $i = 2$ to length $|Px| - 1$ do

A multiobjective optimizer is used to obtain a set of optimal Pareto solutions of problem (5).

A multi-criteria decision making method, i.e. TOPSIS, is used to find the optimal solution.

A new waypoint is generated $P_{new} = [Px(i); y_{opt}; z_{opt}]$.

$P = [P\ \ P_{new}]$.

End For

Step 3: Interpolation of the waypoints

Waypoints are linked by a cubic spline interpolation method.

III. Proposed Multiobjective Multi-Verse Optimizer

A. Basic Concepts

The Multi-Verse Optimizer (MVO), originally proposed by Mirjalili et al. [30], is a recent global metaheuristic based on the physics theories about the existence of multi-verse. The interaction among different universes is ensured based on the concepts of white/black holes and worm holes.

The optimization process of the MVO metaheuristic begins with a set of randomly solutions. At each step, the objects from one universe (decision variables) move according to their inflation rates (fitness values) to another via the white/black holes, and displace within a universe or to another via a worm hole. In this process, the white/black holes are used for the improvements of the exploration mechanism, while the worm holes are employed for the exploitation one. The main updating equations in the MVO metaheuristic are given as follows [30]:

$$x_i^j = \begin{cases} \begin{cases} x_j + TDR + \left(ub_j - lb_j \times r_4 + lb_j\right) & r_3 < 0.5 \\ x_j + TDR - \left(ub_j - lb_j \times r_4 + lb_j\right) & r_3 < 0.5 \end{cases} & if \quad r_2 < WEP \\ x_i^j \quad if \quad r_2 \geq WEP \end{cases} \tag{7}$$

where x_i^j denotes the j[th] component in the i[th] solution, x_j indicates the j[th] variable of the best universe, lb_j and ub_j are the lower and upper bounds, respectively, r_2, r_3 and r_4 are random numbers defined in the interval [0, 1], TDR and WEP present the travelling distance rate and the worm hole existence probability, respectively. They are defined as follows [30]:

$$WEP = \min\_pro + iter\left(\frac{\max\_pro - \min\_pro}{Max\_iter}\right) \tag{8}$$

$$TDR = 1 - \frac{(iter)^{\frac{1}{p}}}{Max\_iter} \tag{9}$$

where $\min\_pro$ and $\max\_pro$ are the minimum and maximum probability of the wormhole existence, respectively, $iter$ is the current iteration, Max_iter is the maximum number of iterations and p is the exploitation accuracy.

In order to elaborate a multiobjective version of the MVO metaheuristic for problem (5), a concept of the archive is added to the research mechanism in a similar way with the well-known literature approaches [31]. Like the MVO, the solutions of the MOMVO algorithm are enhanced using black, white and worm holes. For selecting solutions from the archive, the leader selection method is implemented to establish tunnels among solutions. A roulette wheel approach is used to select the fittest solutions. Obviously, a limited number of solutions can be accommodated in the archive. In order to remove the unsatisfactory ones, a probabilistic mechanism given by Equation (10) is proposed as follows [31]:

$$\delta_i = N_i \Big/ c \tag{10}$$

where N_i defines the number of the vicinity solutions and c is a constant which is greater than 1.

B. Pseudo-Code

According to the above evolution equations (7)-(10) and the basic concept of the MOMVO algorithm, a pseudo-code for its software implementation is presented in Algorithm 2.

Algorithm 2: MOMVO

Step 1: Set the parameters of the MOMVO algorithm
Step 2: Randomly initialize the positions of universes.
Step 3: While $(iter < Max\_iter + 1)$ **do**
  Update WEP and TDR by applying Eq. (8) and Eq. (9).
  For each universe **do**
    Boundary checking for the universes inside search space.
    Calculate the inflation rate (fitness) of universes.
  End For
    Sort fitness values.
    Find the non-dominated solutions.
    Normalize the inflation rates of each universe.
    Update the archive.
  If the archive is full **do**
    Delete some solutions from the archive.
  End If
    Update the position of universes according to Eq. (7)
  If any new solution of the archive is outside boundaries **do**
    Update the boundaries to cover the new solution(s).
  End If
    Increment $iter$
Step 4: Stop the algorithm when it reaches Max_iter .

IV. Simulation Results and Discussion

A. Numerical Validation on CEC'2009 Test Suite

In order to evaluate the performance of the reported competitive algorithms MSSA, MOMVO, MOGWO, NSGA-II and MOPSO for problem (5), an empirical study is firstly conducted based on a benchmark of 9 standard multiobjective test problems from the CEC'2009 test suite [32]. The main control parameters of the reported solvers are setting as follows:

- MOMVO [31]: min and max of wormhole existence probability: 0.2 and 1, respectively.
- MSSA [33]: without control parameters.
- MOGWO [34]: grid inflation 0.1, number of grids per each dimension 10, leader selection pressure 4 and extra repository member selection pressure 2.
- NSGA-II [35]: crossover probability 0.7, mutation probability 0.4 and mutation rate 0.02.
- MOPSO [36]: social and cognitive parameters 2, grid inflation 0.1, leader selection pressure parameter 2 and number of grids per each dimension 7.

In order to have a fair comparison, the termination criterion of all competitive algorithms is set as a maximum number of iterations reached 100. The population size is fixed as 50. After numerical experimentations on a PC with i7 Core 2 Duo/2.67 GHz CPU and 6.00 GB RAM, the obtained optimization results show the effectiveness of all reported algorithms with a remarkable superiority of the proposed MOMVO algorithm in terms of convergence fastness and solutions' quality. The obtained Pareto fronts are closer to the well-known ones with satisfied distribution and repartition of solutions.

B. Path Planning Problem Resolution

In this subsection, the effectiveness and validity of the proposed MOMVO-based planning approach are presented and compared to those of the reported competitive algorithms in different flight scenarios as given in Table I. In order to have an equitable comparison, the population size retained for all reported algorithms in the resolution of problem (5) is set as 100 and the maximum number of iterations is equal to 100. Path planning problems involve finding a feasible path from the starting point to the target one by avoiding out all the obstacles and threats. In this work, five experimental scenarios are investigated. Each of them is specified by the number and position of the static threats as shown in Table I.

TABLE I. Information on External Installations of the Flight Environment

Scenarios	Starting point [km]	Destination point [km]	Threats' number
1	[2,2,0]	[8,10,0]	5
2	[1,2,0]	[10,10,0]	7
3	[1,10,0]	[15,2,0]	10
4	[4,4,0]	[19,13,0]	12
5	[1,18,0]	[17,4,0]	15

Since the generation of each waypoint of the flight path is considered as a solution of the formulated multiobjective optimization problem (5), all reported algorithms are executed on such a constrained problem and the obtained Pareto fronts for the generated waypoints at the same plan are given in Fig. 3 to 7. These results show the repartition topology of the non-dominated solutions through the Pareto surfaces. The best compromise solution is selected at each case thanks to the proposed TOPSIS method. These demonstrative results show high optimization performance in terms of convergence dynamics and solutions' distribution. The proposed algorithms have a good coverage of the non-dominated set of solutions that means a high variety among the different solutions of the optimization problem (5) with the considered two objective functions of Equations (2) and (3) and under operational constraints of Equation (4).

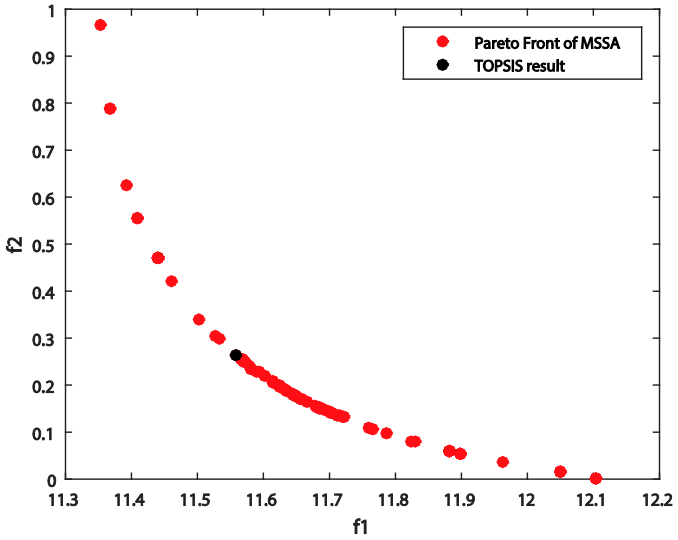


Fig. 3. Pareto Front for the Generation of a Waypoint: MSSA-based Approach.

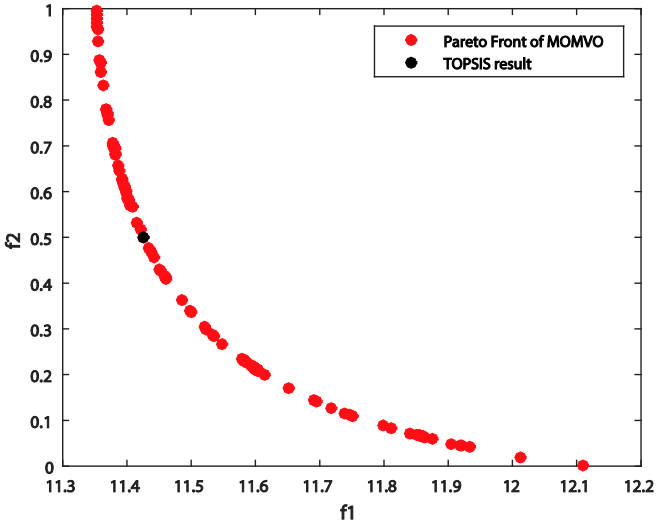


Fig. 4. Pareto Front for the Generation of a Waypoint: MOMVO-based Approach.

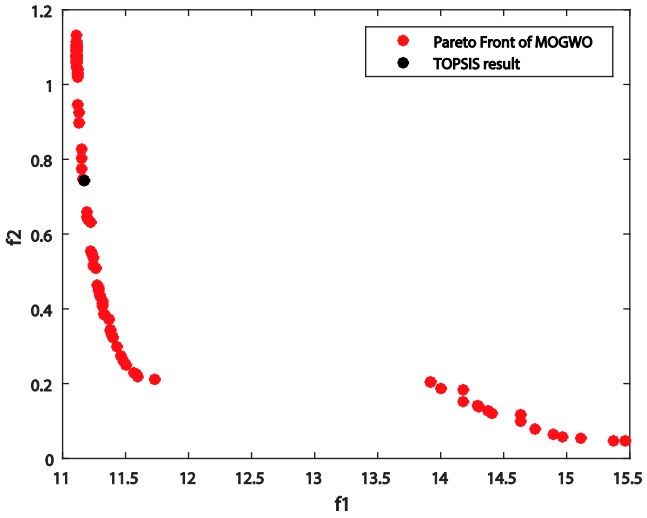


Fig. 5. Pareto Front for the Generation of a Waypoint: MOGWO-based Approach.

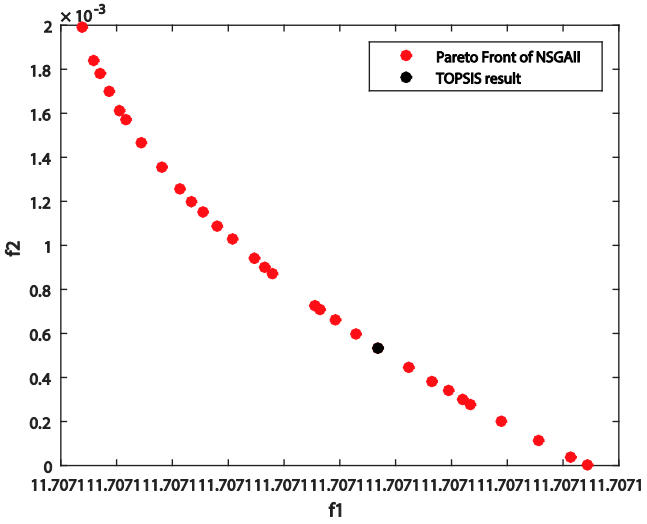


Fig. 6. Pareto Front for the Generation of a Waypoint: NSGA-II-based Approach.

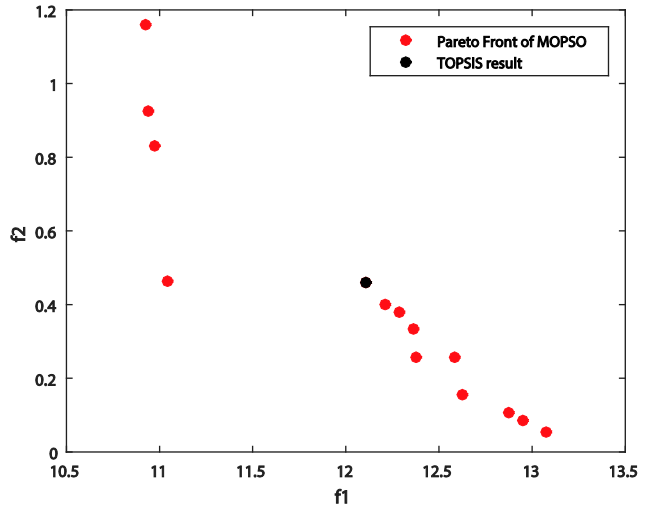


Fig. 7. Pareto Fronts for the Generation of a Waypoint: MOPSO-based Approach.

For the performance comparison purposes, various metrics such as Maximum Spread (MS) [37, 38], Hyper-Volume (HV) [39] and C-metric [40] have been employed in this study. The optimization results related to the MS-metric are presented in Table II. The average performance of the MOMVO has surpassed the other algorithms in terms of having the biggest values for the MS metrics. It may be seen that the coverage of the proposed algorithm tends to be better than other algorithms. Table III shows the comparison of the hyper-volume metrics of different methods. The MOGWO algorithm obtains the largest average of HV values followed by the MOMVO, which means that the MOGWO and MOMVO are the best solvers in terms of diversity and convergence performance.

The comparative results for the proposed algorithm MOMVO with others methods in terms of the C-metric are shown in Table IV. The proposed MOMVO algorithm outperformed all others competitive ones, which dominates more than 21% of the MSSA solutions, 35% of the MOGWO solutions, 1% of the NSGA-II solutions and 76 % of MOPSO solutions on average. The MOGWO algorithm dominates more than 57% of the MOMVO solutions.

TABLE II. COMPARISON OF THE MS-METRIC FOR THE REPORTED ALGORITHMS

	MSSA	MOMVO	MOGWO	NSGAII	MOPSO
Best	114.275	115.395	121.89	64.122	96.2610
Mean	112.477	114.352	114.111	39.6856	63.5578
Worst	111.986	111.745	113.68	11.7029	38.6300
STD	1.0804	1.0712	2.23149	26.5066	28.1470

TABLE III. COMPARISON OF THE HV-METRIC FOR THE REPORTED ALGORITHMS

	MSSA	MOMVO	MOGWO	NSGAII	MOPSO
Best	0.55670	0.61150	4.22660	1.13e-10	2.16790
Mean	0.52158	0.59281	4.02710	2.82e-11	0.54902
Worst	0.47390	0.57180	3.82800	0.0000	0.29900
STD	0.03270	0.01037	0.16850	5.65e-11	0.71760

TABLE IV. COMPARISON OF THE C-METRIC FOR THE REPORTED ALGORITHMS

	Best	Mean	Worst	STD
C (MOMVO, MSSA)	0.34	0.21	0.09	0.0921
C (MSSA, MOMVO)	0.04	0.02	0.00	0.0160
C (MOMVO, MOGWO)	0.39	0.35	0.33	0.0254
C (MOGWO, MOMVO)	0.66	0.57	0.46	0.0776
C (MOMVO, NSGAII)	0.02	0.01	0.00	0.0124
C (NSGAII, MOMVO)	0.25	0.20	0.14	0.0381
C (MOMVO, MOPSO)	1.00	0.67	0.00	0.4306
C (MOPSO, MOMVO)	0.74	0.28	0.00	0.3416

The comparative analysis of MOMVO metaheuristic is performed with MSSA, MOGWO, NSGAII and MOPSO on three performance criteria at the five scenarios, such as the path length, the elapsed time and the 3D planned trajectory. In order to evaluate the capability of the proposed MOMVO algorithm to solve the path planning problem (5), the path length and elapsed time results on each scenario are saved over 10 simulations independently. The statistical results are summarized in Table V.

To analyze the differences between the performances of reported algorithms, nonparametric statistical tests in the sense of Friedman are implemented and discussed [41]. For the five proposed algorithms and five scenarios, the computed value of the χ -distribution is equal to $\chi^2_{F_1} = 10.7059$ for the path length criterion and is $\chi^2_{F_2} = 46.0000$ for the elapsed time one. Based on the distribution table at a confidence level $\alpha = 0.05$, the Friedman statistic, i.e. Iman-Davenport extension of the classical Friedman test, is $\chi^2_{0.95,4,16} = 3.01 < \chi^2_{F_1} < \chi^2_{F_2}$. So the null hypothesis is rejected and there are notable differences between the proposed methods for path planning problem. In order to know which algorithms are different, the post-hoc paired comparison is performed. The algorithms i and j are declared different when the absolute difference of the rank's sum $|R_i - R_j|$ is greater than a critical value. The summation of the ranks of different algorithms is presented in Tables VI and VII. The critical value is equal to 6.1805 for the path length criterion and is 3.3519 for the elapsed time one according to the statistic computation formula given in [42, 43]. The paired comparisons are summarized in Tables VIII and IX. The bold and underlined values in such tables designated that the performances of the reported algorithms are different.

TABLE V. OPTIMIZATION RESULTS OF PROBLEM (5)

Scenarios		MSSA		MOMVO		MOGWO		NSGAII		MOPSO	
		Path length	**Elapsed time**	**Path length**	**Elapsed time**	**Path length**	**Elapsed time**	**Path length**	**Elapsed time**	**Path length**	**Elapsed time**
1	Best	11.716	226.215	11.633	179.835	13.478	609.89	10.367	916.874	12.155	353.330
	Mean	11.805	267.63	11.700	185.214	14.038	694.24	10.684	954.364	12.942	465.251
	Worst	11.889	335.340	11.789	194.796	16.093	769.81	11.014	980.351	13.254	534.214
	STD	0.0487	3.1009	0.0449	2.8908	1.0792	5.5416	0.0510	6.142	0.0562	4.354
2	Best	13.906	463.59	13.814	313.511	15.776	740.541	12.160	1010.02	18.310	770.470
	Mean	13.971	598.90	13.853	352.481	16.259	817.02	13.561	1100.14	18.563	820.540
	Worst	14.060	643.75	13.896	388.173	16.995	868.63	14.547	1502.87	19.547	897.540
	STD	0.0476	1.8751	0.0310	1.0856	0.3817	3.127	0.6751	5.3654	0.5541	3.1452
3	Best	20.052	630.631	19.933	622.273	19.994	920.511	21.040	1246.44	25.035	859.4230
	Mean	20.251	675.421	20.154	658.591	20.169	956.05	21.501	1424.25	26.584	898.3540
	Worst	20.354	780.254	20.280	775.453	20.268	991.8	22.574	1914.24	27.984	950.2450
	STD	0.0541	1.5412	0.0453	1.4516	0.0818	2.0414	0.1554	5.2541	0.1422	3.4251
4	Best	22.764	782.680	22.731	804.85	23.152	1198.13	21.822	1584.63	29.273	1263.090
	Mean	22.815	808.657	22.755	829.669	23.791	1257.24	22.783	2451.35	30.201	1310.250
	Worst	22.867	889.06	22.804	845.214	24.213	1301.32	22.974	2971.35	31.254	1398.210
	STD	0.0353	1.2592	0.0200	1.0241	0.0524	3.2145	0.1642	5.0841	0.1234	4.5412
5	Best	28.953	974.255	28.883	950.989	30.218	2273.75	25.957	2614.11	34.568	1501.90
	Mean	29.076	1201.35	28.967	1024.69	30.451	2468.21	26.874	2781.34	36.354	1654.21
	Worst	29.354	2763.36	29.058	1300.53	30.869	2654.31	27.236	3594.12	37.541	1701.24
	STD	0.1174	550.076	0.0530	112.667	0.3254	2.5471	0.1874	3.1245	0.1542	4.1243

TABLE VI. AVERAGE RANK ANALYSIS OF MEAN PERFORMANCES FOR THE PATH LENGTH CRITERION

	Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5		Rank's sum
	Score	**Rank**	**Score**	**Rank**	**Score**	**Rank**	**Score**	**Rank**	**Score**	**Rank**	
MSSA	11.805	3	13.971	3	20.251	3	22.815	3	29.076	3	15
MOMVO	11.700	2	13.853	2	20.154	1	22.755	1	28.967	2	**8**
MOGWO	14.038	5	16.259	4	20.169	2	23.791	4	30.451	4	19
NSGA-II	10.684	1	13.561	1	21.501	4	22.783	2	26.874	1	9
MOPSO	12.942	4	18.563	5	26.584	5	30.201	5	36.354	5	24

TABLE VII. AVERAGE RANK ANALYSIS OF MEAN PERFORMANCES FOR THE ELAPSED TIME CRITERION

	Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5		Rank's sum
	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	
MSSA	267.63	2	598.90	2	675.42	2	808.65	1	1201.30	2	9
MOMVO	185.21	1	352.48	1	658.59	1	829.66	2	1024.60	1	**6**
MOGWO	694.24	4	817.02	3	956.05	4	1257.20	3	2468.20	4	18
NSGA-II	954.36	5	1100.10	5	1424.20	5	2451.30	5	2781.30	5	25
MOPSO	465.25	3	820.54	4	898.35	3	1310.20	4	1654.20	3	17

TABLE VIII. PAIRED COMPARISON OF THE PROPOSED METAHEURISTICS FOR THE PATH LENGTH CRITERION

$\lvert R_i - R_j \rvert$	MOMVO	MOGWO	NSGA-II	MOPSO
MSSA	**7**	4	6	**9**
MOMVO	-	**11**	1	**16**
MOGWO	-	-	**10**	5
NSGA-II	-	-	-	**15**

TABLE IX. PAIRED COMPARISON OF THE PROPOSED METAHEURISTICS FOR THE ELAPSED TIME CRITERION

$\lvert R_i - R_j \rvert$	MOMVO	MOGWO	NSGA-II	MOPSO
MSSA	3	**9**	**16**	**8**
MOMVO	-	**12**	**19**	**11**
MOGWO	-	-	**7**	1
NSGA-II	-	-	-	**8**

From the results of Tables VI and VIII, i.e. comparison based on the path length criterion, the proposed MOMVO solver outperforms all the MSSA, MOGWO and MOPSO algorithms since the obtained the values of the absolute difference of the rank's sum are greater than the computed critical value 6.1805. However, such an optimizer has the same performance with the NSGA-II algorithm with an absolute difference of rank's sum equal to 1 and according to the final given by Table VI we can confirm that the MOMVO metaheuristic is better than NSGA-II in the case of path length performance metric. For the second criterion, i.e. elapsed time of Tables VII and IX, we found that the MOMVO and MSSA algorithms have the same performance but the MOMVO one is the better since it have the first final rank as given in Table VII. For the rest of competitive algorithms MOGWO, NSGA-II and MOPSO, the proposed MOMVO method has values of absolute difference greater that the computed critical value 3.3519, so it remains better than these mentioned algorithms regarding the final ranking of Table VII.

For the defined performance criteria such as elapsed time, path length and threats avoidance, the planned paths are shown in Fig. 8, 9 and 10 for only the hard flight situations, i.e. 3rd, 4th and 5th flight scenarios with an increasing number of threats. These curves correspond to the optimization results for the mean case. As observed in these figures, the MSSA, MOMVO and MOGWO competitive metaheuristics give the most direct path. They are perfect in all scenarios and can avoid all obstacles and threats, which ensure their high efficiency in flight planning. The path obtained by the MOPSO method avoids all obstacles but takes a long distance in comparison with others algorithms. The NSGA-II algorithm gives a direct path but with a very low level of flight in the case of scenarios 1, 2 and 5 as shown in Fig. 10. The planned path for this solver passes through an obstacle zone in scenarios 3 and 4 as depicted in Fig. 8 and Fig. 9. So, it is too difficult to take into account all the obstacles in certain scenarios.

C. Algorithms' Sensitivity Analysis

In this section, the performance of the two considered main indicators, i.e. path length and execution time, is analyzed with the variations in the population size and iterations values of the competitive algorithms. The performance comparison is given under the 2nd scenario. The results are presented in Tables X and XI. Keeping the iterations constant, the path length decreases linearly with the augment of the population size for all algorithms, on the contrary, the execution time increases. When the population size is constant, the elapsed time varies with the iterations' numbers on the contrary the path length is shorter. The proposed MOMVO algorithm remains robust under these variations and clearly outperforms all others proposed solvers with the shortest path and the minimum elapsed time in most cases. This main capability makes the proposed MOMVO algorithm more adapted for path planning problems.

TABLE XI. PATH LENGTH VARIATION UNDER ITERATIONS AND POPULATION SIZE PARAMETERS OF PROBLEM (6)

Generation	Population size	Path length (km)				
		MSSA	MOMVO	MOGWO	NSGAII	MOPSO
50	20	16.0348	13.8620	16.4859	15.0695	62.7104
	50	13.9310	13.8593	17.0208	13.9520	31.8837
	100	13.8869	13.8411	18.9750	13.8508	20.2795
100	20	15.0420	13.8603	18.4772	14.8723	29.9692
	50	13.9440	13.8549	17.0723	13.8661	23.4170
	100	13.8354	13.8235	15.8578	13.8421	20.6154
200	20	14.1662	13.8602	16.9149	14.5471	19.5114
	50	13.9049	13.8509	16.9512	13.8574	20.8871
	100	13.8244	13.8226	15.8345	13.8314	19.6975

TABLE XII. ELAPSED TIME VARIATION UNDER ITERATIONS AND POPULATION SIZE PARAMETERS OF PROBLEM (6)

Generation	Population size	Elapsed Time (sec)				
		MSSA	MOMVO	MOGWO	NSGAII	MOPSO
50	20	57.3806	56.7261	179.4049	583.8650	279.4850
	50	141.9460	127.4121	371.6397	2851.5570	533.2424
	100	248.4768	279.4000	794.6620	4741.3340	1192.1590
100	20	74.5709	100.4210	574.5162	610.6982	126.6340
	50	176.4819	168.6157	638.9724	3294.8690	1336.0780
	100	365.5790	277.4627	1030.1680	5241.4120	3628.1100
200	20	150.4783	199.9634	432.3340	784.2150	1949.6250
	50	424.0723	420.8711	1121.6300	4145.1420	3980.8930
	100	936.5628	654.8361	2156.1890	5987.1240	4413.2450

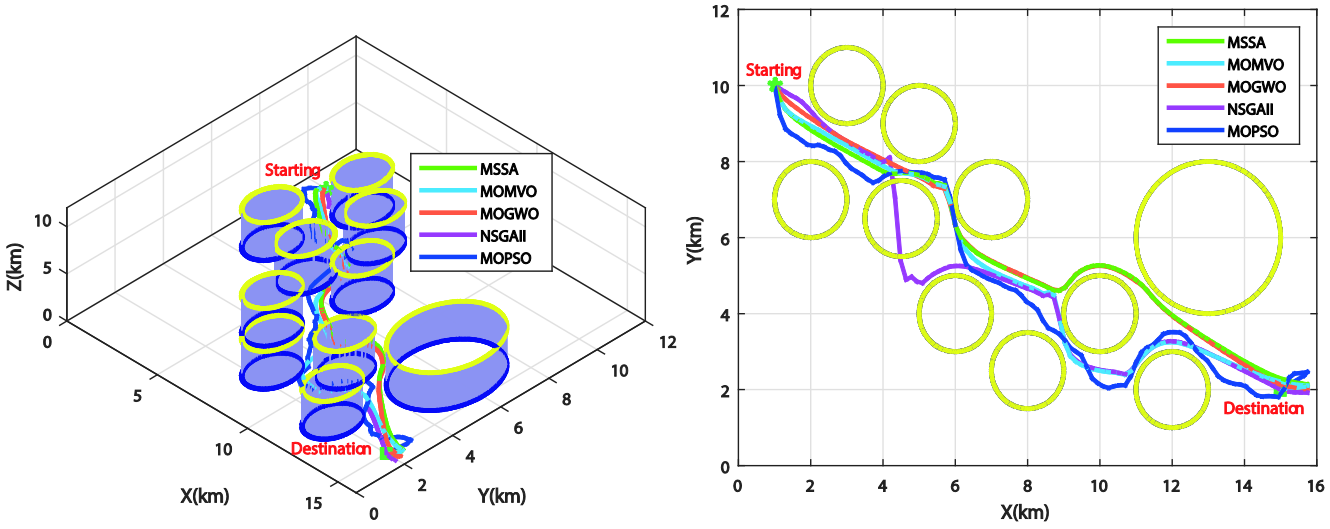


Fig. 8. Performance Comparisons in 3rd Scenario: 10 Threats' Avoidance.

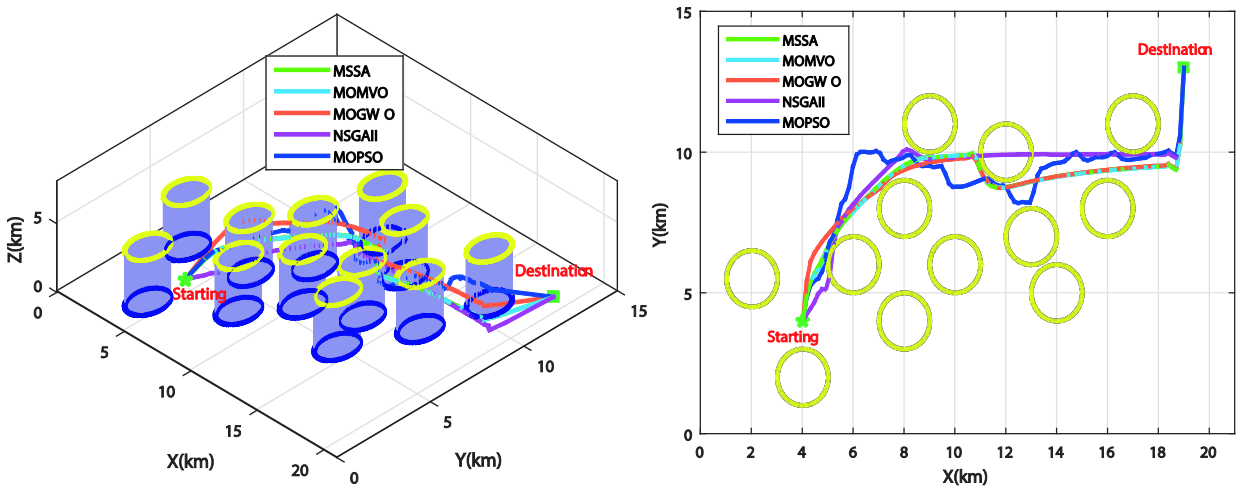


Fig. 9. Performance Comparisons in 4th Scenario: 12 Threats' Avoidance.

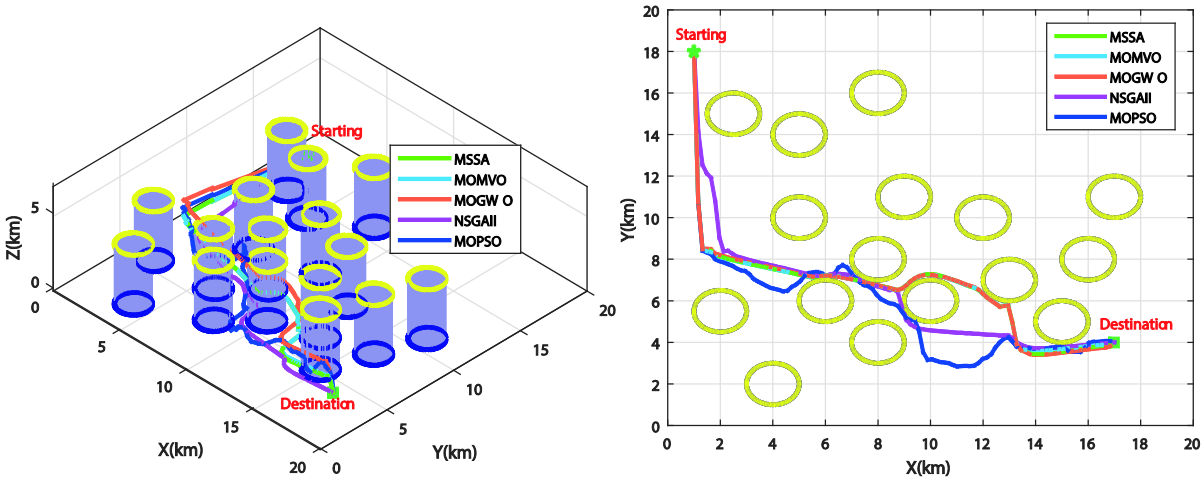


Fig. 10. Performance Comparisons in 5th Scenario: 16 Threats' Avoidance.

V. Conclusion

In this paper, the path planning problem for unmanned aerial vehicles is reformulated by transforming the generation of each flight waypoint into a constrained multiobjective optimization problem. An offline path planning algorithm has been developed and applied within various flight scenarios containing an increasing number of the threats and static obstacles. For an efficient resolution of the formulated multiobjective optimization problem, a recent variant of the MOMVO metaheuristic has been proposed and successfully implemented. A set of competitive algorithms such as MSSA, MOGWO, MOPSO and NSGA-II are retained throughout the study as performance comparison tools. An empirical study of these algorithms has been firstly performed for solving different multi-criteria test functions from the literature. The proposed MOMVO algorithm outperforms all others reported methods in the majority of test functions cases as well as for the real path planning formulated problem. The demonstrative simulations as well as the nonparametric Friedman and paired comparison tests show the effectiveness and superiority of the proposed TOPSIS and MOMVO-based path planning approach in comparison with the reporter competitive algorithms. To further demonstrate the performance of the proposed MOMVO algorithm, several metrics and criteria are employed such as the elapsed time, the path length and threats avoidance capability. The simulation results and comparisons show that the proposed algorithms can successfully solve the 3D UAVs' path planning problem with a remarkable superiority of the MOMVO-based approach. Other improvements should be made in future works as the extension to the paths planning of cooperative multi-UAVs as well as the flight in an environment with dynamic obstacles.

References

[1] A. Idalene, K. Boukhdir and H. Medromi, "UAV Path Planning for Civil Applications", Int. J. of Adv. Comp. Sci. and Appli (IJACSA)., vol. 10, no. 12, pp. 635-641, 2019.

[2] S.J. Chung, A. Paranjape, P. Dames, S. Shen and V. Kumar, "A survey on aerial swarm robotics", IEEE Trans. on Rob., vol. 34, no. 4, pp. 837-855, 2018.

[3] X. Chen, X. Chen and G. Xu, "The path planning algorithm studying about UAV attacks multiple moving targets based on Voronoi diagram," Int. J. of Cont. and Autom., vol. 9, no. 1, pp. 281-292, 2016.

[4] C. A. Cindy and S. O. Rebeca, "Swarm Robotics and Rapidly Exploring Random Graph Algorithms Applied to Environment Exploration and Path Planning", Int. J. of Adv. Comp. Sci. and Appli. (IJACSA), vol. 10, no. 5, pp. 692- 702, 2019.

[5] R. Tianzhu, Z. Rui, X. Jie and D. Zhuoning, “Three-dimensional path planning of UAV based on an improved A* algorithm,” In Proc. of IEEE Chinese Guid., Navig. and Cont. Conf. (CGNCC), pp. 140-145, Nanjing, China, August 12-14, 2016.

[6] H. Kim, J. Jeong, N. Kim and B. Kang, “A Study on 3D Optimal Path Planning for Quadcopter UAV Based on D* Lite,” In Proc. of the Int. Conf. on Unm. Air. Syst. (ICUAS), pp. 787-793, Atlanta, GA, USA, June 11-14, 2019.

[7] I. Iswanto, M. Alfian, W. Oyas and I.C. Adha, “Artificial Potential Field Algorithm Implementation for Quadrotor Path Planning” Int. J. of Adv. Comp. Sci. and Appli. (IJACSA), vol. 10, no. 8, pp. 575 – 585, 2019.

[8] Y.G. Fu, M.Y. Ding, C.P. Zhou and H.P. Hu, “Route planning for unmanned aerial vehicle on the sea using hybrid differential evolution and quantum-behaved particle swarm optimization”, IEEE Trans. on Syst., Man, and Cyb. : Syst., vol. 43, no. 6, pp. 1451-1465, 2013.

[9] A. Gümüşçü, M. Tenekeci and A. Tabanlıoğlu, The shortest path detection for unmanned aerial vehicles via genetic algorithm on aerial imaging of agricultural lands. Int. Adv. Res. and Eng. J., vol. 2, no. 3, pp. 315-319, 2018.

[10] D. M. Vijayakumari, S. Kim, J. Suk and H. Mo, “Receding-horizon trajectory planning for multiple UAVs using particle swarm optimization”, In Proc. of AIAA Scitech 2019 Forum , pp. 1165, San Diego, California, January 7-11, 2019.

[11] B. Abhishek, S. Ranjit, T. Shankar, G. Eappen, P. Sivasankar and A. Rajesh, “Hybrid PSO-HSA and PSO-GA algorithm for 3D path planning in autonomous UAVs,” SN App. Sci., vol. 2, no. 11, pp. 1-16, 2020.

[12] R.K. Dewangan, A. Shukla and W.W. Godfrey, “Three dimensional path planning using Grey wolf optimizer for UAVs”, App. Intelligence, vol. 49, no. 6, pp. 2201-2217, 2019.

[13] B. Li, X. Qi, B. Yu and L. Liu, “Trajectory Planning for UAV Based on Improved ACO Algorithm,” IEEE Access, vol. 8, pp. 2995-3006, 2019.

[14] Y. Chen, J. Yu, Y. Mei, Y. Wang and X. Su, “Modified central force optimization (MCFO) algorithm for 3D UAV path planning”, Neurocomputing, vol. 171, pp. 878-888, 2015.

[15] Y. Liu, X. Zhang, X. Guan and D. Delahaye, “Sensitivity decision based path planning algorithm for unmanned aerial vehicle with improved particle swarm optimization”, Aero. Sci. and Tech., vol. 58, pp. 92-102, 2016.

[16] P.K. Das, H.S. Behera and B.K. Panigrahi, “A hybridization of an improved particle swarm optimization and gravitational search algorithm for multi-robot path planning”, Swa. and Evol. Comp., vol. 28, pp. 14-28, 2016.

[17] R. Madiouni, S Bouallègue, J. Haggège and P. Siarry, “Epsilon-Multiobjective Particle Swarm Optimization Based Tuning of Sensitivity Functions for RST Digital Control Design”, Trans.. of the Inst. of Meas. and Cont. (TIMC), vol. 41, no. 13, pp. 3688-3704, 2019.

[18] R. Madiouni, S. Bouallègue, J. Haggège and P. Siarry, “Robust RST Control Design based on Multi-objective Particle Swarm Optimization Approach”, Int. J. of Cont., Aut., and Syst. (IJCAS), vol. 14, no. 6, pp. 1607-1617, 2016.

[19] C. Ramirez-Atencia, G. Bello-Orgaz, M. D. R-Moreno and D. Camacho, “Solving complex multi-UAV mission planning problems using multi-objective genetic algorithms”, Soft Comp., vol. 21, no. 17, pp. 4883-4900, 2017.

[20] X. Zhen, Z. Enze and C. Qingwei, “Rotary unmanned aerial vehicles path planning in rough terrain based on multi-objective particle swarm optimization”, J. of Syst. Eng. and Elec.,vol.31,no. 1, pp. 130-141, 2020.

[21] J. Liu, W. Wang, X. Li, T. Wang, S. Bai and W. Yanfeng, “Solving a multi-objective mission planning problem for UAV swarms with an improved NSGA-III algorithm,” Int. J. of Comp. Intel. Syst., vol. 11, no. 1, pp. 1067-1081, 2018.

[22] Z. Wang, M. Li, L. Dou, Y. Li, Q. Zhao and J. Li, “A novel multi-objective artificial bee colony algorithm for multi-robot path planning,” In Proc. of Int. Conf. on Info. and Autom. , pp. 481-486, 2015, Lijiang, China , August 8-10, 2015.

[23] Q. Ren, Y. Yao, G. Yang and X. Zhou, “Multi-objective Path Planning for UAV in the Urban Environment Based on CDNSGA-II”, In Pro. of 2019 IEEE Int. Conf. on Ser.-Ori. Syst. Eng., pp. 350-3505, April 2019.

[24] C.R. Atencia, J. Del Ser and D. Camacho, “Weighted strategies to guide a multi-objective evolutionary algorithm for multi-UAV mission planning”, Swa. and Evo. Comp., vol. 44, pp. 480-495, 2019.

[25] S. Thabit and A. Mohades, “Multi-Robot Path Planning Based on Multi-Objective Particle Swarm Optimization”, IEEE Access, vol. 7, pp. 2138-2147, 2019.

[26] Y. Yang, J. Liu, S. Tan, and H. Wang, “A multi-objective differential evolutionary algorithm for constrained multi-objective optimization problems with low feasible ratio”, Applied Soft Computing, vol. 80, pp. 42-56, 2019.

[27] D.A.G.. Vieira, R. Adriano, L.A. de Vasconcelos, “Handling constraints as objectives in a multiobjective genetic based algorithm”, J. of Micro., Opto. and Elec. Appli. (JMOe), vol. 2, no. 6, pp. 50-58, 2002.

[28] C.L. Hwang and K..Yoon, Multiple Attribute Decision Making: Methods and Applications: a State-of-art Survey. Springer-Verlag, New York, 1981.

[29] Z. K. Feng, S. Liu, W. J. Niu, Z. Q. Jiang, B. Luo, and S. M. Miao, “Multi-objective operation of cascade hydropower reservoirs using TOPSIS and gravitational search algorithm with opposition learning and mutation”. Water, vol. 11, no. 10, pp. 2040, 2019.

[30] S. Mirjalili, S.M. Mirjalili and A. Hatamlou, “Multi-verse optimizer: a nature-inspired algorithm for global optimization”, Neur. Comp. and Appli., vol. 27, pp. 495-513, 2016.

[31] S. Mirjalili, P. Jangir, S.Z. Mirjalili, S. Saremi and I.N. Trivedi, “Optimization of problems with multiple objectives using the multi-verse optimization algorithm”, Know. Bas. Syst., vol. 134, pp. 50-71, 2017.

[32] Q. Zhang, A. Zhou, S. Zhao, P. N. Suganthan, W. Liu and S. Tiwari Multiobjective optimization test instances for the CEC 2009 special session and competition. Technical Report CES-487, University of Essex, Nanyang Technological University, (2009).

[33] S.A. Mirjalili, A.H. Gandomib, S.Z. Mirjalilic, S. Saremia, H. Farisd, and S.M. Mirjalili, “Salp Swarm Algorithm: a bio-inspired optimizer for engineering design problems”, Adv. in Eng. Soft., vol. 114, no. 12, pp. 163-191, 2017.

[34] S. Mirjalili, S.M. Mirjalili, S. Saremi and L. Coelho, “Multiobjective grey wolf optimizer: A novel algorithm for multicriterion optimization”, Exp. Syst. with Appli., vol. 47, pp. 106-119, 2016.

[35] K. Deb, A. Pratap, S. Agarwal and T. Meyarivan, “A fast and elitist multi-objective genetic algorithm: NSGAII”, IEEE Trans. on Evol. Comp., vol. 6, no. 2, pp. 182-190, 2002.

[36] C.A.C. Coello, G.T. Pulido and M.S. Lechuga, “Handling multiple objectives with particle swarm optimization”, IEEE Trans. on Evol. Comp., vol. 8, no. 3, pp. 256-260, 2004.

[37] E. Zitzler, K. Deb and L. Thiele, “Comparison of multiobjective evolutionary algorithms: Empirical results”, Evol. Comp., vol. 8, no. 2, pp. 173-195, 2000.

[38] S. Khalilpourazari, B. Naderi and S. Khalilpourazary, “Multi-Objective Stochastic Fractal Search: a powerful algorithm for solving complex multi-objective optimization problems”. Soft Computing, vol. 24, no. 4, pp. 3037-3066, 2020.

[39] E. Zitzler, L. Thiele, M. Laumanns, C.M. Fonseca and V.G.D. Fonseca, “Performance assessment of multiobjective optimizers: an analysis and review”, IEEE Trans. on Evol. Comp., vol. 7, no. 2, pp. 117-132, 2003.

[40] E, Zitzler and L, Thiele, “Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach”, IEEE Trans. on Evol. Comp., vol. 3, no. 4, pp. 257-271, 1999.

[41] W.J. Conover , Practical Nonparametric Statistics, 3rd edition, John Wiley & Sons, New York, 1999.

[42] M. Mazen Alhato, S. Bouallègue, “Thermal exchange optimization based control of a doubly fed induction generator in wind energy conversion systems,” Ind. J. of Elect. Eng. and Comp. Sci. (IJEECS), vol. 20, no. 3, pp. 1252-1260, 2020.

[43] M. Mazen Hato, S. Bouallègue, “Whale Optimization Algorithm for Active Damping of LCL-Filter-based Grid-Connected Converters”, Int. J. of Ren. Ener. Res. (IJRER), vol. 9, no. 2, pp. 986-996, 2019.

Process Level Social Media Business Value Configuration of SMEs in Saudi Arabia

Anwar Shams Eldin[1], Awadia Elnour[2]
Department of Business Administration
University of Bisha, Bisha, KSA

Rugaia Hassan[3]
Department of Information Systems
University of Bisha, Bisha, KSA

***Abstract*—The key enabler of strategic design based on IT is process level value; however, few researchers have tackled the mechanisms through which small and medium-sized enterprises (SMEs) can create value at the process level. This study sheds light on the mechanism of creating social media business value at the process level by identifying the interaction effects of social media and IT resources and the mediating role of management's commitment to innovation as an organizational factor. The research model is based on the IT business value approach, quantitative and descriptive methodology is adopted, and the data are analyzed using structural equation modeling. Among the findings based on 301 SMEs in the Kingdom of Saudi Arabia, that management's commitment to innovation is a necessary condition for social media resources to create dynamic capabilities, and the interaction effects between social media resources and IT resources on social media capability have no impact on the value-generation process at the process level. The result improves the understanding of the theoretical implications of social media business value at the process level, which can be used to guide theorizing about IT business value. SME managers, IT designers, and national decision-makers can use the findings to gain strategic advantage through social media platforms.**

***Keywords*—Interaction effects of social media and IT resources; process level; SMEs; social media business value; social media capabilities; management's commitment to innovation**

I. Introduction

Small and medium-sized enterprises (SMEs) contribute to a country's economic growth despite limited financial resources. In the Kingdom of Saudi Arabia (KSA), SMEs contribute about 33% of the GDP, often using information technology to challenge competitors in various sectors. The use of social media among business and society is increasing, creating strategic advantage [1]. Only a few studies have investigated how SMEs integrate social media technology into their operations and strategies [2]. However, many studies in the Saudi context have sought to investigate how SMEs can create value from social media [1]. Most previous studies in this field draw on the information technology (IT) business value literature [3], exploring the impacts of IT investment at the business and process levels [4]. Recently, scholars have explored the impact at the country (or macro) level [5].

At the process level, the business value is the impact of IT resources on performance at this level, which can be generated from inter-firm diversity, or complementary resources and capabilities [6], [7] creating unique resources that cannot be easily replicated [7] termed as capability, which can be related to two functions—facilitating the functional activities of an organization, which is functional capability (operational capabilities), or facilitating a transformation of the way an organization conducts its functional processes, which is dynamic capability.

However, the relationship between social media operational and dynamic capability in SMEs has rarely been investigated although work such as this can shed light on the mechanisms that build capabilities.

Complementary resources are those that interact to generate capabilities [8]. Although some scholars have explored how IT resources interact with other organizational resources, only a few studies have investigated how IT resources interact with one another [9]: for instance, the interaction between social media resources and other IT resources has, so far, been neglected [10]. This knowledge gap can be filled by formulating a comprehensive model that demonstrates how IT resources interact to generate value.

Many factors purportedly explain the conditions under which business value is created, namely organizational, industrial and macroeconomic factors [7], [9]. Prior studies have failed to conceptualize in a clear way how these factors influence value generation at different organizational levels, and their results have been inconsistent [9]. Thus, the value-generation mechanism at the process level constitutes a gap in the research. In SMEs, managers' strategic decisions are usually influenced by their individual perceptions and characteristics [11]. Thus, management's commitment to using a particular technology or innovation as an organizational factor influences how an organization benefits from using social media. Unfortunately, most previous research has neglected the mechanisms through which this factor affects value generated at the process level.

To exploit the potential of social media to create value in SMEs at the process level, this study explores the mechanism of social media business value creation at the process level, aiming to identify conditions under which social media resources can contribute to creating value at the process level, by investigating the effects of the interaction between social media resources and other IT resources on social media capabilities (operational and dynamic). Moreover, it highlights the linkage between social media capabilities at the process level, and how management's commitment to innovation affects the relation between social media resources and capabilities.

This work is funded by university of Bisha

This can enable IT managers to improve the techniques they use to manage and implement IT and social media through a better understanding of the ways IT and social media generate value at the process level. In turn, this can enable SMEs in KSA to gain value by extending the marketing of their products and services locally and globally. This is a potential source of competitive advantage [8],[12] and is the key enabler of strategic design based on IT at the process level. Moreover, the first-order impacts of IT at the process level can further contribute to impacts on firm-level performance.

The remainder of this paper is organized as follows: Section II provides an overview literature review, besides hypotheses development. Section III describes the research model. Section IV presents the methodology used to conduct the research. Section V provides discussions of the results and conclusions. Section VI proposes future studies.

II. Literature Review and Hypotheses

A. IT Resources and Social Media Resources and Social Media Capbilities Process Level Business Value

IT resources comprise both tangible and intangible components [13] that reflect how the organization uses these resources to generate value [9].

Social media is an IT resource based on Web 2.0 [14], comprising a wide variety of tools and platforms categorized into types according to their potential [15] that allow organizations to interact and exchange information with their customers efficiently and cheaply [14]. From the resource-based view (RBV) perspective, social media can be classified into tangible and intangible assets. The former is the infrastructure of the platform [16] while the latter includes information produced through the use of social media by customers and the firm to generate value at the process level [17].

Prior studies have focused on the impact of social media on business value, describing various forms of value that can be generated using social media at the process level in terms of internal value [15], [18]. The outcome of value generation at this level is capabilities [13], [9].

Based on the dynamic capability approach, capabilities are traditionally described hierarchically as existing at various organizational levels [19], [8] or as multi-dimensional constructs generated by a firm's ability to combine, assemble and integrate resources to create strategic advantage [20] Numerous studies have distinguished dynamic capabilities from other capabilities [21], namely operational capability [19]. This classification can provide a comprehensive view of how an organization can perform the various activities that can give insight into IT business value at the process level.

Social media capabilities refer to how a firm uses social media platforms (resources) to generate business value from its activities [4]. Of the few scholars that have conceptualized social media capabilities, the majority have identified social media in terms of functionality (technological features) [22], [15], marketing [23], or strategic functionality [24] or have concentrated on a single organizational level [16]. Thus, prior studies have not provided comprehensive insights into social media operational capabilities concerning their role in business value creation via functionality. Gaining such insights can help to create a roadmap for capability-building mechanisms, which is necessary for designing effective strategies to achieve superior performance compared with competitors.

Previous studies on IT business value indicate that IT resources positively influence capabilities [9]. In a study on the Sudanese banking industry, Shams Eldin [25] found direct impact between IT resources and dynamic capabilities. In the same context Shams Eldin et al [26] found direct relationship between IT resources and IT functional capabilities. Many scholars have built theoretical models to conceptualize how social media resources configure value at process level. Wittkuhn et al. [16], based on RBV and dynamic capability approach, built a conceptual model to investigate how social customer relation management resource impacts social media customer management capabilities and how both can impact social media customer management performance. An empirical study conducted by Trainor et al. [10] found that social media platform usage positively influences social media capabilities. In SMEs in India, Chatterjee and. Kumar Kar [27] found that social media usage facilitates the use of social media marketing, which brings business benefits to the organization. Moreover, the cost has significant impact in adoption of social media market; this may be due to small budget of SMEs. Therefore, we hypothesize the following:

H1. IT resources have a positive impact on social media capabilities.

H1a. IT resources have a positive impact on social media operational capabilities.

H1b. IT resources have a positive impact on social media dynamic capabilities.

H2. Social media resources have a positive impact on social medial capabilities.

H2a. Social media resources have a positive impact on social media operational capabilities.

H2b. Social media resources have a positive impact on social media dynamic capabilities.

B. Social Media Dynamic Capabilities and Operational Capabilities

Dynamic capabilities result from the transformation of the lower-level capability to confront change in a business environment [21], such as operational capabilities [19], [9]. A few empirical studies have investigated how this is achieved [21]. Pour et al. [28] aimed to investigate the impact of customer relation capabilities on innovation capabilities as dynamic capabilities, the findings of their study indicated to the direct impact of customer relation capabilities on innovations capabilities. In Brazil, Toriani and Angeloni [29] found that use of IT customer relationships management capabilities can support knowledge capabilities as dynamic capabilities. In SMEs based on knowledge management approaches, Cheng and Shiu [30] found a direct relationship between customer involvement and innovation. This is because the strong customer information processing capability enables SMEs to analyze complex information and give new insights from

customer involvement, thereby creating knowledge benefits that help in innovations in their new product/service. We thus hypothesize the following:

H3. Social media operational capabilities have a positive impact on social media dynamic capabilities.

C. The Mediating Role of Management Commitment

Previous literature describes the direct models of IT business value generating. By contrast, the indirect model suggests that the impact of social media resources at the process level is mediated or moderated by organizational factors [31] and that social media resources interact with other organizational resources to generate value at the process level, Trainor et al. [10].

Collectively, these models describe the idea of complementary resources [12], as discussed in prior work [9]. Complementary resources can create unique value to the organization, differentiating it from its competitors.

An example of the factors that can influence the ability of resources to create value and, thus, describe the conditions under which a certain value is created, is organizational factors, which have the potential to influence the relationship between resources and firm performance. Examples include top management's commitment, organizational structure, firm size, location and industry [32].

Management's commitment reflects how managers' individual characteristics influence the way they make decisions [11]. Management's commitment can be categorized by its various forms, such as strategic decision-making, development, digital platform usage, working smarter using available information, conducting action-oriented assessments [33], and commitment to innovation [34]. The present study focuses on how management's commitment to innovation can influence the generation of business value at the process level, which reflects how managers deal with new inventions and technology, such as social media. This also sheds light on management's ability to deal with the impacts of social media on value generation at the process level and provides insights into the role of managers in creating business value [33], in SMEs where the manager plays a central role in the organization.

According to relationship marketing theory, social media usage can be considered an investment relationship between consumers and marketers that influences business value through the meditation of consumer commitment [35]. Thus, in any social network, commitment to the relationship can be considered a mediator between the two parties of a relationship [36]. Following this line of logic, if managers do not commit to developing a relationship between resource usage and capabilities, such a relationship cannot occur. Thus, management's commitment can be considered a mediating factor [37].

Another perspective on how management's commitment influences the relationships between social media usage and performance is based on the innovation system approach and network theory. Nybakk et al. [34] considered innovative commitment a mediator between social networking and economic performance. In the context of SMEs, Cheng and Shiu [30] suggest that the transformation of resources into capabilities requires the commitment of SME managers in developing new capabilities when facing a new market situation, such as the social media environment. This is rarely investigated in SMEs compared with larger firms. To the researchers' knowledge, management's commitment to innovation in SMEs has not received sufficient attention from social media scholars, although managers' decision plays a central role in adopting IT solutions. Accordingly, we hypothesize the following:

H4. Social media usage has a direct impact on management's commitment to innovation.

H5. Management's commitment to innovation has a direct impact on social media capabilities.

H5a. Management's commitment to innovation has a direct impact on social media dynamic capabilities.

H5b. Management's commitment to innovation has a direct impact on social media operational capabilities.

H6. Management's commitment to innovation mediates the relationships between social media usage and social media capabilities.

H6a. Management's commitment to innovation mediates the relationship between social media usage and social media dynamic capabilities.

H6b. Management's commitment to innovation mediates the relationship between social media usage and social media operational capabilities.

D. Interaction Effect of Social Media and IT Resources

Previous studies on IT business value emphasize the role of interaction between resources to create new unique resources that can generate superior value at the process level [9], [12]. In the context of social media, Trainor et al. [10] found a positive relationship between the interaction of customer management systems (as an IT resource) and social media resources and a firm's social media customer relationship capabilities. We thus hypothesize the following:

H7. The interaction of social media resources and IT resources has a positive impact on social media capabilities.

H7a. The interaction of social media resources and IT resources has a positive impact on social media dynamic capabilities.

H7b. The interaction of social media resources and IT resources has a positive impact on social media operational capabilities.

III. Research Model

Fig. 1 illustrate the relationship between the variables in the study, through a two-level model that was designed to explain how social media resources create value. The model comprises the resource level and the capability level at the process level.

The model is based on the prior literature on IT business value, as this approach provides insight into how

complementary resources create value at the process level in terms of capabilities. It also describes a hierarchy of resources and capabilities and how they are related to one another.

In this model, the resources are IT and social media assets (tangible and intangible) that interact with and complement each other to create value at the process level in terms of capabilities, which are social media resources generated at two levels: operational and dynamic capabilities.

For social media resources to create value at the process level, certain conditions must be satisfied regarding organizational factors [38] such as management's commitment to innovation, which is considered as a mediator between social media resources and capabilities.

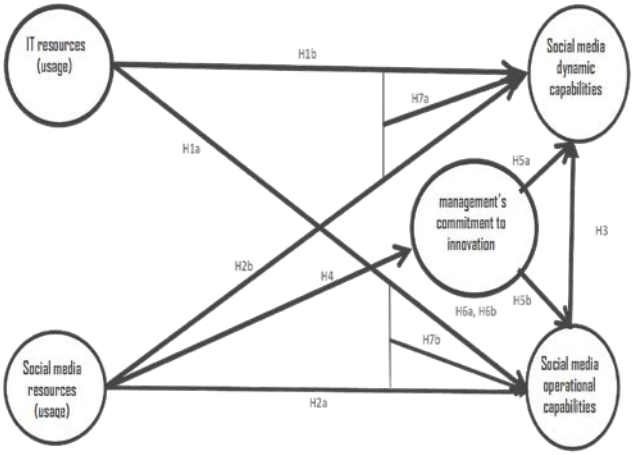


Fig. 1. Research Model.

IV. Research Methodology

A. Sampling and Data Collection

Data were collected through a questionnaire that was distributed to people with comprehensive insights on the organization of Saudi SMEs, such as managers, owners and senior staff members. The questionnaire was initially distributed by email, but very few responses were received, so other electronic methods were used, such as messages through WhatsApp, Twitter, Instagram and Facebook. To increase the response rate, individuals known to organizations' managers were recruited because they could easily allow the collection of data.

Of approximately 2,000 distributed questionnaires, 301 were returned. The questionnaire was distributed to diverse respondents to attempt to cover SMEs in different sectors. The distribution of respondents is as follows: 87.4% were male, 86.4% were aged from 35-45 years, and 77% have a bachelor's degree or college diploma certificate. The distribution of the sector in which respondents work is as follows: 7.7% in the manufacturing industry, 9% in financial services, 8.3% in investment services, 6% in IT services, 15.3% in retail, 6.7% in professional and technical services, 14% in construction services and 33% in other sectors.

B. Survey Instrument

All items in the survey questionnaire were taken from previous studies and adapted to reflect the environment in which Saudi SMEs operate. The key items of the questionnaire were measured through five-point Likert scales, and the questionnaire was divided into six parts. The first part focused on general information about the manager or owner of the enterprise and the enterprise itself. The second part covered social media resources and the third part concentrated on IT. The fourth focused on social media capabilities subdivided into customer relation capabilities, marketing capabilities, production and operational capabilities, technological capabilities, strategic capabilities, knowledge management capabilities, and collaboration and innovation capabilities. The fifth part covered the management commitment to innovation.

Applying Agnihotri et al.'s [39] procedure, which is especially relevant to SMEs, we selected the top social media platforms since these have a large audience.

Eight social media platforms were chosen to reflect the usage of social media resources in KSA SMEs: Facebook, Twitter, LinkedIn, YouTube, WhatsApp, blogs, Snapchat and Instagram. The nature of social media usage, which comprises a variety of types and dimensions, and social media resources are distinct uncorrelated factors. Accordingly, social media resources can be treated as the formative construct, as suggested by Hu and Zhang [40].

The classification of IT resources was adapted from Bardhan et al. [41] to reflect commonly used IT resources in SMEs in KSA. Their approach uses a wide range of IT resources to measure the usage of basic communications technology, group collaboration technology and enterprise computing technology. This broad view of IT usage enables a comprehensive evaluation of how IT resources affect social media at the process level. This construct is considered a reflecting construct, as suggested by Wetering et al. [42].

Social media capabilities were classified into two levels by adapting Protogerou et al.'s [19] classification of capabilities, whom classified them into dynamic and operational capabilities.

In this study, the selected social media dynamic capabilities are those considered necessary for the adaption of an organization's essential processes in a business environment changing due to circumstances such as technological change, including social media innovation and collaborative capabilities [43], social media strategic capabilities [19], and social media knowledge management capabilities [44]. The scale used to rate the social media dynamic capabilities was adapted from Yang and Chen [45], Nguyen et al. [24], Wong et al. [43] and Wang et al. [2].

Social media operational capabilities result from the use of social media to produce goods and services [19]. Operational capabilities were measured through four items: social media production and operational capabilities; social media marketing capabilities; social media customer relationship capabilities; and social media technological capabilities. These items were adapted from Schaupp and Bélanger [17], Trainor et al. [10], Kane [46], Wong et al [43], Marzouk [31] and Wang and Kim [18]. Following prior research, social media capabilities were modelled as reflective constructs [17].

The scale of management's commitment to innovation was adapted from Nybakk et al. [34] and is considered a reflective construct, as suggested by Schaupp and Bélanger [17].

C. Analysis

Partial least squares structural equation modeling analysis was used because of its ability to deal with complex models combining correlations between several constructs and explicit variables, such as interaction and mediation effects, and handle both reflective and formative scales with high predictive accuracy and lower risk than other regression techniques that predict cause-and-effect relationships [47]. Moreover, this modeling technique can handle exploratory models of theories in the early stages of development [47], such as the business value approach considered in this study. Following Chin [48] and Hair et al. [47], the model was analyzed in two stages. The first stage assessed the model's reliability, convergent validity and discriminant validity, and the second stage tested the structural model. Since the research model contains both formative and reflective constructs, each type must be treated differently during evaluation [47].

The reflective constructs were assessed through construct validity, reliability, convergent validity and discriminant validity. Table I presents the results of construct reliability and validity. To confirm convergent validity, the factor loading must be above 0.7, composite reliability (CR) should be higher than 0.7 and average variance extracted (AVE) should be higher than 0.5 [47]. Following Hair et al. [47], a Cronbach's alpha value of 0.7 is considered acceptable. The results in Table I indicate that the model satisfies the requirements for both reliability and convergent validity.

TABLE I. FACTOR LOADING, RELIABILITY AND VALIDITY

Factors and Items	**Factor Loadings**				**Cronbach's Alpha**	**CR**	**AVE**
Organizational factor (OF)					0.876	0.915	0.730
OF1		0.885					
OF2		0.890					
OF3		0.879					
OF4		0.756					
Social media dynamic capabilities (SMDC)					0.964	0.968	0.655
Social media innovation and collaborative capability (SMDc)							
SMDc1			0.848				
SMDc2			0.852				
SMDc3			0.851				
SMDc4			0.840				
Social media knowledge capability (SMDk)							
SMDk1			0.819				
SMDk2			0.835				
SMDk3			0.823				
SMDk4			0.827				
Social media strategic capability							
SMDs1			0.843				
SMDs2			0.833				
SMDs3			0.791				
SMDs4			0.829				
Social media operational capabilities (SMP)					0.968	0.971	0.674
Social media marketing capability \(SMPm)							
SMPm1				0.825			
SMPm2				0.842			
SMPm3				0.835			
SMPm4				0.847			
Social media customer relationship management capability (SMPo)							
SMPo1				0.849			
SMPo2				0.778			
SMPo3				0.846			
SMPo4				0.807			
Social media production and operational capability (SMPp)							
SMPp1				0.831			
SMPp2				0.808			
SMPp3				0.815			
SMPp4				0.818			
Social media technological capability (SMPt)							
SMPt1				0.852			
SMPt2				0.752			
SMPt3				0.811			
MPt4				0.813			
IT resource usage (SMR)					0.872	0.907	0.662
SMR1	0.821						
SMR2	0.788						
SMR3	0.865						
SMR4	0.799						

Table II presents the results for discriminant validity, which was assessed by comparing the AVEs for correlations and other latent constructs, as suggested by Hair et al. [47].

Based on the results in Table II, just one value does not satisfy the requirements of discriminate validity. However, taking these findings in combination with the confirmatory factor analysis results in Table I, the model can be considered valid [49].

The formative constructs were examined through weights and variance inflation factors (VIFs) to test for multicollinearity, as suggested by Hair et al. [47]. Table III presents the results.

As Table III shows, all the social media resources satisfy the condition of VIF being less than 5 [47]. Although some indicators do not show significant output for the weight value, they were not removed from the model following Chin's [48] criteria.

The second stage of the analysis involved testing the structural model using a bootstrapping technique. Interaction effects were tested through the latent variable modeling approach recommended by Hair et al. [47]. The mediating

effects of organizational factors on the relationships between social media resources and IT resources were also tested following the procedure suggested by Hair et al. [47].

Further analysis was conducted to ensure the significance of specific indirect effects through manual calculations of the standard error. This method is recommended by Hair et al. [47]. Table IV presents the results for both direct effects and the specific indirect effects.

Referring to Table IV, the relationship between social media resources and social media operational capabilities was significant ($\beta = -0.143$, $p < 0.020$) while the relationship between social media resources and dynamic capabilities was not significant ($\beta = 0.048$, $p < 0.146$). Thereby, H1 is partially supported. The results in Table IV also give support to H2 as the relationship between IT resources and social media operational ($\beta = 0.414$, $p < 0.000$) and dynamic capabilities ($\beta = 0.074$, $p < 0.083$) was significant. Social media operational capabilities had a significant impact on social media dynamic capabilities ($\beta = 0.642$, $p < 0.000$), supporting H3. Social media resources had a significant impact on management commitment ($\beta = -0.242$, $p < 0.015$), thus supporting H4. H5 is supported due to significant relationship between management's commitment to innovation and both social media dynamic capabilities ($\beta = 0.289$, $p < 0.000$) and operational capabilities ($\beta = 0.445$, $p < 0.000$). The results for indirect effects provide support for the mediating effect predicted by hypothesis H6a and reject hypothesis H6b. The impact of the interaction effects between social media and IT resources on social media operational capabilities ($\beta = 0.083$, $p < 0.568$) and dynamic capabilities ($\beta = -0.099$, $p < 0.498$) was not significant and, thus, H7 is not supported.

TABLE II. DISCRIMINANT VALIDITY TEST

Variables	SMR	OF	SMDC	SMP
SMR	0.814			
OF	0.588	0.854		
SMDC	0.704	0.792	0.809	
SMP	0.713	0.711	0.896	0.821

TABLE III. VALIDITY ANALYSIS OF FORMATIVE CONSTRUCTS

Social Media Resources (SOMR)	Standardized Loading	P Value	VIF
Blog	0.333	0.049	1.188
Facebook	-0.424	0.031	1.096
Instagram	-0.053	0.701	1.250
LinkedIn	0.271	0.065	1.180
Snapchat	-0.006	0.963	1.204
Twitter	-0.131	0.434	1.251
YouTube	0.654	0.002	1.177
WhatsApp	0.016	0.914	1.089

TABLE IV. SUMMARY OF HYPOTHESES TESTING RESULTS

Direct Effect					
	Original Sample (O)	Sample Mean (M)	Standard Deviation (STDEV)	T Statistics (|O/STDEV|)	P Values
SMR -> SMDC	0.074	0.070	0.043	1.734	0.083*
SMR -> SMP	0.414	0.397	0.074	5.558	0.000***
Interaction 2 (SOMR * SMR) -> SMP	0.083	-0.014	0.145	0.571	0.568
Interaction1 (SOMR * SMR) -> SMDC	-0.099	-0.013	0.146	0.677	0.498
OF -> SMDC	0.289	0.293	0.053	5.471	0.000***
OF -> SMP	0.445	0.436	0.069	6.409	0.000***
SMP -> SMDC	0.642	0.634	0.060	10.771	0.000***
SOMR -> SMR	-0.373	-0.374	0.124	3.012	0.003***
SOMR -> OF	-0.242	-0.254	0.099	2.431	0.015**
SOMR -> SMDC	0.048	0.042	0.033	1.453	0.146
SOMR -> SMP	-0.143	-0.122	0.062	2.321	0.020**
Indirect Effect					
SOMR -> OF -> SMDC	-0.070	-0.075	0.033	2.114	0.035
SOMR -> OF -> SMP	-0.107	-0.111	0.049	2.174	0.030

Note: *p<0.10, **p<0.05, ***p<0.01

V. DISCUSSIONS AND CONCLUSIONS

This study investigates how social media resources complement and interact with other IT resources to create value at the process level in Saudi SMEs.

This paper is unlike previous social media process-level studies, which focused on processes management and process culture and did not devote sufficient attention to how value is created [50].

The results of this study indicate that management's commitment to innovation mediates the relationships between social media resources and social media dynamic capabilities. Thus, management's commitment to innovation is a necessary condition for creating dynamic capabilities. This implies that strong commitment from management to innovation can facilitate the usage of social media resources to create dynamic capabilities at the process level. The literature suggests that managers who are committed to innovation are willing to accept risk [51]. Thus, SME managers willing to accept risk and who support innovation can build dynamic capabilities that can create change in an organization. This contrasts with some IT business value scholars, such as Tallon et al. [38], who claim that management committed to innovation must interact with social media resources to create value at the process level.

The interaction between social media resources and IT resources was not found to contribute to value generation at the process level. This contradicts the findings of Trainor et al. [10] and can be attributed to the consideration of the characteristics of SMEs that engage in limited use of information systems [52]. Thus, SMEs in KSA do not benefit from investment in IT and social media to generate synergies at the process level. Accordingly, IT designers should focus on developing IT applications that can be integrated with social media applications and be easily used by SMEs at a low cost.

Although the results of this study demonstrate that the interaction of IT resources and social media resources is not a predictor of social media process-level capabilities, IT resources do contribute to building social media operational and dynamic capabilities, which is aligned with the results of previous studies such as Trainor et al. [10]. This indicates different attributes make different contributions to building new social media capabilities without interacting with one another. Thus, organizations can create unique value if they focus on integrating social media platforms with IT resources.

The results suggest that social media resources directly affect dynamic capabilities because social media can facilitate communication and collaboration and diffuse knowledge-related competencies. This power of connectivity enables organizations to connect to their trading partners, customers and employees, transforming the information that is communicated among these groups. Thus, this enables innovation regarding management's knowledge and creates new opportunities for the company to create business value. This indicates that the way dynamic capabilities are built depends on how individuals choose to use technology [46].

The findings reveal the direct impact of social media resources on social media operational capabilities, such as marketing, customer relationship capabilities, operational capabilities and technological capabilities. This indicates social media can support a firm's operational capabilities and connect it with customers, suppliers and employees. This finding is not surprising as KSA has good internet infrastructure available to customers at suitable prices, which enables the use of mobile and other electronic devices in the work environment.

In summary, the findings indicate that social media resources directly contribute to generating business value at the process level in terms of both operational and dynamic capabilities, which is consistent with Trainor et al. [10]. This can enable SME managers to take steps to introduce social media in the organization to achieve competitive advantage.

Social media operational capabilities contribute to building dynamic capabilities, which supports the hierarchy of resources discussed by Wang and Ahmed [44] and Protogerou et al. [19]. Overall, the results indicate that social media can transform the way firms work at the process level. Thus, the findings support the dynamic capabilities framework proposed by Wilden et al. [53], who indicated building dynamic capabilities requires a routine in terms of operational capabilities.

The study has several managerial implications. First, this research represents a first step toward identifying social media capabilities and determining how they are generated from the perspective of the organization. This enables managers to think about integrating social media operational capabilities into business processes to achieve business goals. This will enable business decision-makers and IT designers to recognize the process capabilities and conditions under which the organization can generate business value from social media. Accordingly, criteria for the effective use of social media can be identified. Second, the results explain how social media usage contributes to generating dynamic capabilities, showing the importance of the role of the innovative manager in building this capability, which is one of the strategic capabilities that can generate competitive advantage. This suggests the SME managers can use innovative techniques to build dynamic capabilities from social media to face the competitors and the changing environments. Third, managers and IT designers can use the results of the study to align social media capabilities with other organizational capabilities to build strategies to maximize the value generated for the organization. This advances the suggestion that SMEs need a new business model that integrates social media with organizational goals. Overall, investigating, the value creation process can lead to recommendations on best practices. Fourth, the model describes complex relations between resources and capabilities to generate some operational and adaptive dynamic capabilities; this proposes that, whenever complex relationships between resources exist, this will lead to generating new capabilities that are not easily replicated, as suggested by RBV, thus, managers can use their innovation to build this complex interaction. Future researches should focus on this complex interaction to provide more insight into the value creation process.

Fifth, the findings can also support national-level strategic plans to increase the contribution of SMEs to KSA's economic growth through social media-based e-commerce [1], through giving the necessary support for use the social media in e-commerce.

The study has many empirical impactions. This result supports social network theory, which suggests that management's commitment to innovation mediates the relationships between social media usage and business value. The findings also support the dynamic capability approach by enabling exploration of managers' role in creating dynamic capabilities, as proposed by Schilke et al. [54].

This study contributes to the literature on social media business value in several ways. To the best of the researchers' knowledge, this is the first study to highlight the attributes of both social media and IT interaction and the mediating role of management's commitment to innovation to create new business value at the process level. Considering different theories can enhance the explanation of how organizational factors can affect value generation at the process level. This gives boundaries to the theory of social media business value, as recommended by Andersson, Cuervo-Cazurra, and Nielsen [55]. The study empirically tests knowledge gaps in the IT business value literature, such as the mechanism by which business value is created through social media at the process level by the disaggregation of social media resources at different levels, as discussed by Schryen [9]. Accordingly, it can be used as a starting point for future research.

In general, this study sheds light on operational capabilities that have not been given sufficient attention in previous social media studies by focusing on capabilities such as customer relationship capabilities and marketing capabilities [18].

VI. Future Research

The results present many challenges for scholars, including the impacts of organizational and environmental factors, and the interactions between these factors, on the relationships between social media resources and capabilities. Future research should investigate interaction effects of social media and IT resources in organizations of different sizes and with different characteristics.

The research model presents a comprehensive, hierarchical list of social media capabilities that can be used in future researches. It should also be noted that the research model neglects the impacts of other organizations' resources and capabilities on value creation through social media at the process level. Thus, future research should focus on how different organizational capabilities contribute to value generation in organizations of different sizes and with different characteristics.

Longitudinal studies could provide insight into how value is generated at the process level and, specifically, how social media dynamic capabilities evolve. Thus, by implementing the research model in different industries and countries, future studies could yield findings that enable the design of effective social media strategies, reducing the knowledge gap between the potential and actual use of social media in business and increasing understanding of how it can create value for organizations [56].

Acknowledgment

The authors are thankful to the Deanship of Scientific Research at University of Bisha Saudi Arabia for funding this work through the Research Project under Grant Number (UB-115-1438).

References

[1] S. Abed, Y. Dwivedi and M. Williams, "Social commerce as a business tool in Saudi Arabia's SMEs," International Journal of Indian Culture and Business Management, vol. 13, no. 1, pp. 1-19, 2016.

[2] Y. Wang, M. Rod, S. Ji and Q. Deng, "Social media capability in B2B marketing: toward a definition and a research model," Journal of Business & Industrial Marketing, vol. 32, no. 8, pp. 1125-1135, 2017.

[3] Trkman and P. Trkman, "A framework for increasing business value from social media", Economic Research-Ekonomska Istraživanja, vol. 31, no. 1, pp. 1091-1110, 2018.

[4] Braojos-Gomez, J. Benitez-Amado and F. Javier Llorens-Montes, "How do small firms learn to develop a social media competence?," International Journal of Information Management, vol. 35, no. 4, pp. 443-458, 2015.

[5] P. Appiahene, N. Ussiph and Y. Missah, "Information Technology Impact on Productivity:A Systematic Review and Meta-Analysis of the Literature", International Journal of Information Communication Technologies and Human Development, vol. 10, no. 3, pp. 39-61, 2018. Available: 10.4018/ijicthd.2018070104.

[6] D. de Lima Oliveira and A. Maçada, "IT capabilities' business value: analysis of multi-level performance in Brazilian organization,s" Gestão & Produção, vol. 24, no. 2, pp. 295-309, 2017.

[7] N. Melville, K. Kraemer and V. Gurbaxani, "Review: Information Technology and Organizational Performance: An Integrative Model of IT Business Value," MIS Quarterly, vol. 28, no. 2, p. 283, 2004.

[8] M. Queiroz, "Business process and IT cospecialization: Conceptualization and suggestions for future research," in 23rd Americas Conference on Information Systems, Boston, Massachusetts, USA, 2017.

[9] G. Schryen, "Revisiting IS business value research: what we already know, what we still need to know, and how we can get there," European Journal of Information Systems, vol. 22, no. 2, pp. 139-169, 2013.

[10] K. Trainor, J. Andzulis, A. Rapp and R. Agnihotri, "Social media technology usage and customer relationship performance: A capabilities-based examination of social CRM," Journal of Business Research, vol. 67, no. 6, pp. 1201-1208, 2014.

[11] L. Hsieh et al., "A multidimensional perspective of SME internationalization speed: The influence of entrepreneurial characteristics," International Business Review, vol. 28, no. 2, pp. 268-283, 2019. Available.

[12] Y. Wang, S. Shi, S. Nevo, S. Li and Y. Chen, "The interaction effect of IT assets and IT management on firm performance: A systems perspective," International Journal of Information Management, vol. 35, no. 5, pp. 580-593, 2015.

[13] H. Mao, S. Liu, J. Zhang and Z. Deng, "Information technology resource, knowledge management capability, and competitive advantage: The moderating role of resource commitment," International Journal of Information Management, vol. 36, no. 6, pp. 1062-1074, 2016.

[14] A. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," Business Horizons, vol. 53, no. 1, pp. 59-68, 2010.

[15] S. Kargaran, M. Jami Pour and H. Moeini, "Successful customer knowledge management implementation through social media capabilities," VINE Journal of Information and Knowledge Management Systems, vol. 47, no. 3, pp. 353-371, 2017.

[16] N. Wittkuhn, T. Lehmkuhl, T. Küpper and R. Jung, "Social CRM performance dimensions: A resource-based view and dynamic capabilities perspective," in 28th BLED e-conference, Bled, Slovenia, 2015.

[17] C. Schaupp, J. Dorminey and R. Dull, "A resource-based view of using social media for material disclosures," in 48th Hawaii International Conference on System Sciences, Hawaii, USA, 2015.

[18] Z. Wang and H. Kim, "Can Social Media Marketing Improve Customer Relationship Capabilities and Firm Performance? Dynamic Capability Perspective," Journal of Interactive Marketing, vol. 39, pp. 15-26, 2017.

[19] A. Protogerou, Y. Caloghirou and S. Lioukas, "Dynamic capabilities and their indirect impact on firm performance," Industrial and Corporate Change, vol. 21, no. 3, pp. 615-647, 2011.

[20] E. Bendoly, A. Bharadwaj and S. Bharadwaj, "Complementary Drivers of New Product Development Performance: Cross-Functional Coordination, Information System Capability, and Intelligence Quality," Production and Operations Management, vol. 21, no. 4, pp. 653-667, 2011.

[21] P. Wójcik, "Exploring Links Between Dynamic Capabilities Perspective and Resource-Based View: A Literature Overview," International Journal of Management and Economics, vol. 45, no. 1, pp. 83-107, 2015.

[22] J. Benitez, A. Castillo, J. Llorens and J. Braojos, "IT-enabled knowledge ambidexterity and innovation performance in small U.S. firms: The moderator role of social media capability," Information & Management, vol. 55, no. 1, pp. 131-143, 2018.

[23] A. Rathore and P. Ilavarasa, "Social media and business practices", Encyclopaedia of Information Science and Technology. IGI Global., USA, pp. 7126 -7139, 2018.

[24] B. Nguyen, X. Yu, T. Melewar and J. Chen, "Brand innovation and social media: Knowledge acquisition from social media, market orientation, and the moderating role of social media strategic capability," Industrial Marketing Management, vol. 51, pp. 11-25, 2015.

[25] A. Shams Eldin, "The relationships between IT resources and dynamic capabilities: Evidence from Sudanese insurance and banking sectors,"

International Journal of Advanced and Applied Sciences, vol. 7, no. 4, pp. 91-102, 2020.

[26] A. Shams Eldin, A. Hafiez and A. Al-Tit, "Impact of IT Resources on IT Capabilities in Sudanese Insurance and Banking Sectors," International Journal of Advanced Computer Science and Applications, vol. 7, no. 6, 2016.

[27] S. Chatterjee and A. Kumar Kar, "Why do small and medium enterprises use social media marketing and what is the impact: Empirical insights from India," International Journal of Information Management, vol. 53, p. 102103, 2020.

[28] M. Pour, E. Mamani and M. Rahimzadeh, "How Customer Relationship Management (CRM) and Innovation Influence Business Performance Mediating Role of Innovation," International Journal of Customer Relationship Marketing and Management, vol. 9, no. 2, pp. 1-15, 2018.

[29] S. Toriani and M. Angeloni, "CRM as a Support for Knowledge Management and Customer Relationship," JISTEM Journal of Information Systems and Technology Management, vol. 8, no. 1, pp. 87-108, 2011.

[30] C. Cheng and E. Shiu, "How to enhance SMEs customer involvement using social media: The role of Social CRM," International Small Business Journal: Researching Entrepreneurship, vol. 37, no. 1, pp. 22-42, 2018.

[31] W. Marzouk, "Usage and Effectiveness of Social Media Marketing in Egypt : An Organization Perspective," Jordan Journal of Business Administration, vol. 12, no. 1, pp. 209-238, 2016.

[32] Wade and J. Hulland, "Review: The Resource-Based View and Information Systems Research: Review, Extension, and Suggestions for Future Research," MIS Quarterly, vol. 28, no. 1, p. 107, 2004.

[33] A. Quaadgras, P. Weill and J. Ross, "Management Commitments that Maximize Business Impact from IT," Journal of Information Technology, vol. 29, no. 2, pp. 114-127, 2014.

[34] E. Nybakk, P. Crespell, E. Hansen and A. Lunnan, "Antecedents to forest owner innovativeness: An investigation of the non-timber forest products and services sector,"Forest Ecology and Management, vol. 257, no. 2, pp. 608-618, 2009.

[35] M. Clark and J. Melancon, "The Influence of Social Media Investment on Relational Outcomes: A Relationship Marketing Perspective," International Journal of Marketing Studies, vol. 5, no. 4, 2013.

[36] A. MorettI, and A. Tuan, "Social media marketing and relationship marketing: revolution or evolution? A first step analysis", Sinergie Italian Journal of Management, no. 93, pp. 115-137, 2018.

[37] R. Ryssel, T. Ritter and H. Georg Gemünden, "The impact of information technology deployment on trust, commitment and value creation in business relationships," Journal of Business & Industrial Marketing, vol. 19, no. 3, pp. 197-207, 2004.

[38] P. Tallon, M. Queiroz, T. Coltman and R. Sharma, "Business Process and Information Technology Alignment: Construct Conceptualization, Empirical Illustration, and Directions for Future Research," Journal of the Association for Information Systems, vol. 17, no. 9, pp. 563-589, 2016.

[39] R. Agnihotri, R. Dingus, M. Hu and M. Krush, "Social media: Influencing customer satisfaction in B2B sales," Industrial Marketing Management, vol. 53, pp. 172-180, 2016.

[40] T. Hu and P. Ping, "Social media usage as a formative construct: Conceptualization, validation and implication," Journal of Information Technology Management, vol. 4, no., pp. 151-164, 2016.

[41] I. Bardhan, V. Krishnan and S. Lin, "Project Performance and the Enabling Role of Information Technology: An Exploratory Study on the Role of Alignment," Manufacturing & Service Operations Management, vol. 9, no. 4, pp. 579-595, 2007.

[42] V. Weterin, R. Versendaal and P. Walraven, "Examining hospital's IT infrastructure capability and digital capabilities: A resource-based perspective," in 24th Americas Conference on Information Systems., New Orleans, Louisiana, USA, 2018.

[43] C. Wong, K. Lai, T. Cheng and Y. Lun, "The role of IT-enabled collaborative decision making in inter-organizational information integration to improve customer service performance," International Journal of Production Economics, vol. 159, pp. 56-65, 2015.

[44] C. Wang and P. Ahmed, "Dynamic capabilities: A review and research agenda,"International Journal of Management Reviews, vol. 9, no. 1, pp. 31-51, 2007.

[45] C. Yang and L. Chen, "Can organizational knowledge capabilities affect knowledge sharing behavior?," Journal of Information Science, vol. 33, no. 1, pp. 95-109, 2007.

[46] C. Kane, "Enterprise social media: Current capabilities and future possibilities," MIS Quarterly Executive, vol. 14, no. 1, pp. 1-16, 2015.

[47] Hair, G. Hult, C. Ringle and M. Sarstedt, A primer on partial least squares structural equation modeling (PLS-SEM). Los Angeles: SAGE, 2017.

[48] W. Chin, "How to write up and report PLS analyses", in Handbook of partial least squares concepts, methods and applications, Berlin: Springer, 2010, pp. 655-690.

[49] A. Farrell and J. Rudd, "Factor analysis and discriminant validity: A brief review of some practical issues," in Australia-New Zealand Marketing Academy Conference., 2009.

[50] J. Prodanova and A. Van Looy, "How Beneficial is Social Media for Business Process Management? A Systematic Literature Review," IEEE Access, vol. 7, pp. 39583-39599, 2019.

[51] A. Sikora and E. Nybakk, "Rural development and forest owner innovativeness in a country in transition: Qualitative and quantitative insights from tourism in Poland," Forest Policy and Economics, vol. 15, pp. 3-11, 2012.

[52] A. Bahaddad, L. Houghton and S. Drew, "Attracting Customer in Saudi Arabia to Buy from Your Business Online," International Journal of Business and Management, vol. 8, no. 7, 2013.

[53] R. Wilden, T. Devinney and G. Dowling, "The Architecture of Dynamic Capability Research: Identifying the Building Blocks of a Configurational Approach," SSRN Electronic Journal, 2016.

[54] O. Schilke, S. Hu and C. Helfat, "Quo Vadis, Dynamic Capabilities? A Content-Analytic Review of the Current State of Knowledge and Recommendations for Future Research," Academy of Management Annals, vol. 12, no. 1, pp. 390-439, 2018.

[55] U. Andersson, A. Cuervo-Cazurra and B. Nielsen, "From the Editors: Explaining interaction effects within and across levels of analysis," Journal of International Business Studies, vol. 45, no. 9, pp. 1063-1071, 2014.

[56] S. Ketonen-Oksi, J. Jussila and H. Kärkkäinen, "Social media based value creation and business models," Industrial Management & Data Systems, vol. 116, no. 8, pp. 1820-1838, 2016.

Recent Progress of Blockchain Initiatives in Government

A Review of Asian Countries

Faizura Haneem[1], Hussin Abu Bakar[2], Nazri Kama[3], Nik Zalbiha Nik Mat[4], Razatulshima Ghazali[5], Yasir Mahmood[6]
ICT Consulting Division, Malaysian Administrative Modernisation and Management Planning Unit (MAMPU), Malaysia[1,2,4,5]
Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia[3,6]
Department of Software Engineering, Faculty of Information Technology[6]
The University of Lahore, Lahore, Pakistan[6]

***Abstract*—Blockchain is a decentralized and distributed ledger technology that aims to ensure transparency, data security, and integrity. There are rising interest and investment by the governments and industries in Blockchain to deliver significant cost savings and increase efficiency. Identifying Blockchain initiatives that are currently implemented in the government world-wide could improve understanding as well as set benchmarks for specific countries. However, although some review studies on Blockchain initiatives have been carried out, there are very few studies that uncover Blockchain initiatives implemented by the government in Asian countries. Hence, this study reviews Blockchain initiatives in the five-top e-government development index (EGDI) countries in Asia; South Korea, Singapore, Japan, United Arab Emirates, and Cyprus. We strategized our review methods by utilizing relevant keyword searches in existing literature, books, academic journals, conferences, and industrial reports. The results of this study will help other researchers and practitioners to recognize the current stage of Blockchain initiatives in the government of Asian countries.**

***Keywords*—Blockchain initiatives; governments; review**

I. Introduction

Blockchain technology is a distributed database that stores data of digital assets and transactions across a peer-to-peer network. Blockchain is used in a modern and innovative way to secure, store, and distribute the data [1]. The data in Blockchain are secured through a chain of blocks that are cryptographically linked together [2]. Unlike a centralized implementation of the Relational Database Management System (RDBMS), Blockchain stores the data in multiple nodes, and this mechanism is called distributed open-ledger [3]. This disruptive technology is expected to change many existing business models or make new ones that could generate significant impacts on entire industries.

This cryptographic link of the records in Blockchain makes the data immutable and traceable whereby once it has been created in the Blockchain, modification is not allowed. This would prevent the data from any forge [1]. Additionally, by storing the data in distributed open ledgers in the whole network, it is almost impossible to tamper and the need for a central entity is no longer necessary [4]. Miner or sealer nodes across the peer-to-peer network are responsible for the validity of the data via a consensus mechanism without any central entity control. The decentralized architecture of Blockchain could also increase efficiency and avoid a single point of failure [1]. In the nutshell, Blockchain could be considered as a data infrastructure that provides security, immutability, and traceability of the records, with high-efficiency and high-reliability of the system.

There are rising interest and investment by the government and industries in Blockchain to deliver significant cost savings and increase efficiency. A report by United Nations [5] stated that there are potential applications of Blockchain in government from identity management, registries, evidence management system, energy and water exchange, to taxes and remittances [5], [6]. Besides, Blockchain could also facilitate Internet-of-Things implementation by enabling the sharing of information and resources that would establish a demand for services between devices [7], [8]. Currently, many pilot projects have been reported [9] in this regard as it could be a tool for governments to improve transparency, prevent fraud, and build trust in the public sector.

In current state-of-the-art, although some review studies on Blockchain initiatives have been carried out, there are very few studies that uncover Blockchain initiatives by the government, particularly in Asian countries context. Since Blockchain is a relatively new and emerging technology, prior studies mostly focus on reviewing the challenges of Blockchain adoption and technical capabilities [4], [10], [11]. In regards to government, a review study by [12] concentrates on current Blockchain architecture applied in the government, and a study by [13] focuses on consensus protocols of Blockchain in the Public Sector. Meanwhile, a study by [10] systematically review challenges of Blockchain adoption by government organizations, and a study by [14] analyses Blockchain initiatives in the public sector but it only focuses on D5 countries Digital 5 (D5) countries which include United Kingdom, United States, Estonia, New Zealand, and Israel. On the other hand. Hence, this paper attempts to address the gap by reviewing recent Blockchain initiatives in the government of Asian countries.

This study reviews Blockchain initiatives in the government of five countries in Asia which have achieved a Very High E-Government Development Index (EGDI) based on the United Nations evaluation in 2020 [15] which are South Korea, Singapore, Japan, Cyprus, and the United Arab

Emirates (UAE), with the EGDI of 0.9560, 0.9150, 0.8989, 0.8731, 0.8555, respectively. This study may guide future researchers and practitioners wishing to embark on Blockchain initiatives especially in the context of government.

The remainder of the paper is organized as follows: Section II presents an overview of the Blockchain concepts; Section III describes the study methodology; Section IV presents the review results and analysis; Section V discusses the results and findings and finally conclusion and recommendations are drawn in Section VI.

II. Literature Review

Recently, more governments join the race to keep pace with the global movement of Blockchain adoption by approaching regulatory legislation and launching pilot projects every day. Governments and public sector organizations are using Blockchain technology to step away from siloed and inefficient centralized structures. Current centralized systems appear expensive and highly-complex, whereas Blockchain is expected to provide more stable, scalable, and cost-effective architectures. Among the main characteristics of Blockchain which could significantly transform the government is as follows:

1) Immutability: Blockchain is a registry of encrypted digital records or transactions called blocks. Every block is then "chained" to the next block, in linear and sequential order, utilizing a cryptographic signature [16]. The Blockchain is a block sequence containing a full list of transaction records whereby each data block is time stamped and linked in chronological order via a cryptographic signature [17]. Since immutability can be described as the ability of the Blockchain ledger to remain unchanged and unaltered. This feature of Blockchain makes no one can interact with the system or change the data stored in the blocks.

2) Traceability: Since each of the transactions on the Blockchain is validated and recorded with a timestamp, users can easily verify and trace the previous records by accessing any node in the distributed network. In the Bitcoin Blockchain, each transaction could be traced to previous transactions iteratively. It improves the traceability and the auditability of the data stored in the Blockchain.

3) Persistency: Since each of the transactions spreading across the network needs to be confirmed and recorded in blocks distributed in the whole network, it is nearly impossible to tamper. Additionally, each broadcasted block would be validated by other nodes and transactions would be checked. Hence, any falsification could be detected easily.

4) Decentralized: In traditional centralized transaction systems, each transaction must be checked by a central trusted entity (e.g. a central bank) which inevitably results in cost and performance bottlenecks on central servers. Differently, a transaction in the Blockchain network can be carried out through a decentralized method between any two peers (P2P) without being authenticated by the central agency. Blockchain can only be changed by consensus between system participants without a central instance and a transaction can never be modified or removed [18]. The elimination of a central instance in a decentralized network of Blockchain means a drastic change to direct transactions between non-intermediaries or intermediary services [19].

III. Study Methodology

This review paper is designed based on a systematic review methodology for IS research [20]. This study simplified the review protocol, enumerated as follows: 1) research questions; 2) search strategy design; 3) study selection; and 4) analysis of findings.

A. Research Questions

Two research questions were formulated: 1) What are the Blockchain initiatives in the government of Asian countries? and 2) How did the initiatives vary by sector?

B. Search Strategy Design

The study thoroughly searched all published works found in the existing literature, books, academic journals, conferences, government, and industrial reports, searching several databases using keywords. First, this study performed a systematic review of Blockchain initiatives in the chosen countries from a database indexed by Scopus, WoS Core Collection, and several prominent digital government reports. Then, using the snowball method, the study extends the search to government and industrial reports such as OECD Working Papers on Public Governance 2018 [21], Blockchain in Government Tracker by The Illinois Blockchain Initiative [9], and E-Government Survey 2020 by the United Nations [15].

Search Keywords: Consistent with prior research by [22], three steps involved in constructing the search keywords were: 1) Identification of alternative spellings and synonyms for major terms; 2) Identification of keywords in relevant papers or books; and 3) Usage of the Boolean OR to incorporate alternative spellings and synonyms. The initial search strings used were “Blockchain”, “initiatives”, “project”, “intention”, “program”, “government”, “public sector”, “Korea”, “Singapore”, “UAE”, “Japan”, “Cyprus”, “Asia countries”. Then, the search strings were combined using “AND” and “OR” Boolean. The search strings were used in each electronic database to retrieve the publications based on the titles, abstracts, and keywords, depending on the advanced search facility provided by each database.

Search Criteria: Three search criteria were used: 1) the language used in the paper was English; and 2) the paper was categorized under journals, proceedings, books, book chapters, and reports 3) the search process involved retrieval of publications from 2015 to 2020.

C. Study Selection

Initially, 814 publications were identified using the search keywords from the selected databases. The search was then extended to the Google Scholar database using the snowball method. This was to ensure that several prominent digital government reports that were not indexed in our selected databases were also considered. From the search, an additional 19 publications were found.

During the search process, metadata of the initial articles were gathered and indexed. The metadata were: 1) Electronic Repository; 2) Title of the literature; 3) Abstract; 4) Year; 5) Publication Type; and 6) DOI/ISBN/ISSN Number. Based on the metadata, the deduplication process was then performed to eliminate any duplicated copies of the articles [23], [24], which then decreased the number of publications to 587.

Next, a practical screening process was conducted which involved screening the titles and abstracts of the publications to select relevant publications. Then, a quality assessment was conducted against the selected publications by evaluating the criteria. Quality assessment criteria involved the retrieval of the full version of these publications and the context of the studies.

D. Analysis of Findings

After conducting the quality assessment, 29 publications were selected for the analysis stage. The following section analyses the findings from these qualified publications.

IV. Findings and Discussion

As described earlier, there are two research questions to be answered by this review study. The following sections describe the findings and discussion based on the research questions.

A. What are the Blockchain Initiatives in the Government of Asian Countries?

Table I represents the Blockchain initiatives in the government of South Korea, Singapore, Japan, Cyprus, and the UAE. The following section describes 30 noticeable initiatives in South Korea, Singapore, Japan, Cyprus, and the UAE.

1) South Korea: South Korea's government is the world leader in online services provision and has the highest EGDI value in Asia has shown a positive movement towards Blockchain [15]. The province of Gyeonggi-do has used a voting system based on Blockchain technology to collect its citizens to vote on community projects. The Seoul Metropolitan Government is developing an ICON project which is expected to incorporate the city-wide initiative's Blockchain applications into the government's core operating IT system used to carry out key administrative activities [9], [25]. The use of ICON for the dissemination of government documents is the first case in which the Government of South Korea used for the Appointment of the Seoul Blockchain Governance [25].

In October 2020. Korea Internet & Security Agency is expected to implement a Blockchain-powered employee ID system using KISA Coin [26]. KISA Coin tokens are issued to reward the best workers, allowing them to buy snacks, drinks, and even office supplies. It is currently being evaluated as a pilot, with the goal that it will be made available to the entire workforce. Among Blockchain and Cryptocurrency regulations, National Tax Services published its preliminary assessment of taxation on cryptocurrency after its 2017 forum [27].

2) Singapore: The Monetary Authority of Singapore (MAS) implements Project Ubin to explore Blockchain to clear and settle payments and securities. The payment network will provide interfaces for other Blockchain networks to be seamlessly connected and provide additional features such as Delivery-versus-Payment settlement with private exchanges, conditional payments, and trade escrow, and payment obligations for trade finance [9], [28]. Meanwhile, the Singapore Economic Development Board, Monetary Authority of Singapore (MAS) in partnership with IBM has established IBM Blockchain Innovation Centre. It is claimed to be the first Blockchain Innovation Center in the world. The Innovation Center is expected to include both global banks and fintech start-ups, and to develop a range of technology pilots for the finance and commercial industries based on the Linux Foundation Hyperledger platform [9], [29].

On the other hand, the Monetary Authority of Singapore (MAS) has created the R3 Blockchain Legal Centre of Excellence (LCoE) as a forum for the global legal community to get the latest updates and discuss best practices on Blockchain technology and the Corda R3 Blockchain network. Law firms are constantly collaborating with clients to provide specialized advice about how to deal with the legal implications of Blockchain technology, such as structuring Corda business networks and drawing up smart legal contracts [9], [30]. Info-communications and Media Development Authority (IMDA) established TradeTrust which consists of a collection of internationally agreed principles and mechanisms that support the exchange of electronic trading documents. This is done through a distributed Blockchain providing interoperability for linking governments and businesses [9], [31]. IMDA also organized Hackathon Competition in which the participants are challenged to create successful Blockchain-based Minimum Viable Products (MVPs) or Proofs-of-Concept (POCs) solutions to industry-related challenges [9], [32].

In the education sector, Ngee Ann Polytechnic (NP) uses Blockchain to check the validity of the polytechnic diplomas - the first public institute of education that uses Blockchain in that manner in Singapore [33]. While Singapore Customs Authority developed the National Trade Platform (NTP) - The new national trading platform will replace the existing TradeNet and TradeXchange platforms for the declaration of permits and other facilities for trade and logistics [33]. In terms of regulations, Singapore has a balanced Blockchain and Cryptocurrency foster by the Monetary Authority of Singapore in taxation and money transmission law [27].

3) Japan: The Ministry of Economy, Trade, and Industry in Japan has developed a Blockchain assessment methodology or form. It is claimed to be the first type of evaluation of Blockchain-based systems taking into account comparability with existing systems & comprehensibility and by analyzing the tradeoff between assessment items [9], [34]. Besides, the ministry also conducted a Survey on Blockchain technologies

to universities and research institutes in studying the applicability of Blockchain technology. It is used to facilitate the mobility of domestic and foreign human resources and safe reliability of research results as well as research for the development of standards [9], [35]. The Ministry of Justice has developed Land's Blockchain-based central repository of land and property registration [9], [36].

Meanwhile, the Ministry of Internal Affairs and Communications has developed a Blockchain-Based Tendering System is expected to assist both the private sector and the government to apply for tenders. Applicants would receive the tax payment certificates and other required documentation from different government agencies using Blockchain [9], [37]. About regulations, the Financial Services Agency of Japan has established Blockchain and Cryptocurrency Regulation in Japan [9], [27].

TABLE I. BLOCKCHAIN INITIATIVES IN THE GOVERNMENT OF ASIAN COUNTRIES

No.	Country	Government Entity	Initiative Name	Sector	Source
1.	South Korea	Gyeonggi-do Province	Local Government Voting	Social	[9], [48]
2.	South Korea	Seoul Metropolitan Government	Blockchain Strategy	Inter-Sector	[9], [15], [49]
3.	South Korea	Seoul Metropolitan Government	Appointment of the Seoul Blockchain Governance	Inter-Sector	[25]
4.	South Korea	Korea Internet & Security Agency	Blockchain-powered employee ID system	Administration	[26]
5.	South Korea	Financial service agencies in South Korea	Blockchain and Cryptocurrency Regulation	Financial	[27]
6.	Singapore	Monetary Authority of Singapore (MAS)	Project Ubin	Financial	[9], [15], [21], [28]
7.	Singapore	Singapore Economic Development Board, Monetary Authority of Singapore (MAS)	IBM Blockchain Innovation Centre	Financial	[9], [29]
8.	Singapore	Monetary Authority of Singapore (MAS)	R3 Blockchain Legal Centre of Excellence (LCoE)	Financial	[9], [21], [30]
9.	Singapore	Info-communications and Media Development Authority (IMDA)	Trade Trust	Financial	[9], [31]
10.	Singapore	Info-Communications and Media Development Authority (IMDA)	Hackathon Competition: Securing IoT Devices with Blockchain	Inter-Sector	[9], [32]
11.	Singapore	Ngee Ann Polytechnic (NP)	Academic credential	Education	[33]
12.	Singapore	Singapore Customs Authority	National Trade Platform (NTP)	Supply-Chain	[33]
13.	Singapore	Financial service agencies in South Singapore	Blockchain and Cryptocurrency Regulation	Financial	[27]
14.	Japan	Ministry of Economy, Trade and Industry	Blockchain Assessment Methodology/Form	Inter-Sector	[9], [34]
15.	Japan	Ministry of Economy, Trade and Industry	Survey on Blockchain Technologies and Related Services	Inter-Sector	[9], [35]
16.	Japan	Ministry of Justice	Blockchain-based Land Registry	Land	[9], [36]
17.	Japan	Ministry of Internal Affairs and Communications	Blockchain-Based Tendering System	Supply Chain	[9], [37]
18.	Japan	Financial service agencies in Japan	Blockchain and Cryptocurrency Regulation	Financial	[9], [27]
19.	Cyprus	Minister of Finance	National Strategy for Cyprus	Inter-Sector	[9], [38]
20.	Cyprus	Financial Services Agency of Cyprus	Cryptocurrency Regulation, Sales Regulation, Money Transmission Law	Financial	[27]
21.	Cyprus	University of Nicosia	Block-Chain Verified Certificates	Education	[35], [50]
22.	UAE	The Emirates Blockchain Strategy 2021	Emirates & Dubai Blockchain Strategy	Inter-Sector	[9], [15], [40], [41]
23.	UAE	Financial service agencies in UAE	Cryptocurrency Regulation, Sales Regulation, Taxation	Financial	[27]
24.	UAE	Dubai Customs, Dubai Trade	Blockchain powered E-commerce Platform	Supply Chain	[9], [42]
25.	UAE	Dubai Immigration and Visas Department	Biometric verification with Blockchain technology	Immigration	[9], [43]
26.	UAE	Dubai Future Foundation	Global Blockchain Council	Inter-Sector	[9], [21], [44]
27.	UAE	Smart Dubai Office	Global Blockchain Challenge	Inter-Sector	[9], [45]
28.	UAE	Dubai Healthcare City Authority and the Dubai Health Authority (DHA)	Health Smart Licensing	Healthcare	[9], [46]
29.	UAE	Dubai Multi Commodities Centre (DMCC)	Securing the Diamond Trade	Supply Chain	[9], [47]
30.	UAE	Roads and Transport Authority (RTA)	Vehicle history Blockchain project	Transportation	[41]

4) Cyprus: The Minister of Finance in Cyprus has established the National Strategy for Cyprus in 2019. The goal of the national strategy is to encourage the advancement of this technology through innovation and pilot applications, through close cooperation between the public and private sectors [9], [38]. The University of Nicosia, Cyprus has established an open standard for verifying academic certificates on the Blockchain [39]. In terms of regulations, Cyprus has established Blockchain and Cryptocurrency Regulation in Japan including [27] Cryptocurrency Regulation, Sales Regulation, Money Transmission Law.

5) United Arab Emirates: The UAE has established The Emirates Blockchain Strategy 2021. The strategy uses three strategic pillars Government Efficiency, Industry Creation, and International Leadership to make "Dubai the first city fully powered by Blockchain by 2020" [9], [40], [41]. The Dubai Customs, Dubai Trade has formed a Blockchain-powered E-commerce Platform - The latest Blockchain-based project, which aims to help the emirate become a global e-commerce center and enable e-commerce companies to set up their businesses in Dubai [9] [42]. Meanwhile, the Dubai Immigration and Visas Department has combining biometric authentication and Blockchain technology to develop digital passports for seamless entry at Dubai Airport [9], [43]. The Global Blockchain Council was formed by Dubai Future Foundation to examine, discuss current and future applications, and coordinate transactions through the Blockchain platform [9], [44].

Smart Dubai Office has organized the Global Blockchain Challenge which has received 700 applications from 79 countries around the world in 2019. The participated projects covered 20 different fields, including, but not limited to, real estate, asset management, payments, energy, education, healthcare, and supply chain management [9], [45]. In healthcare, the Dubai Healthcare City Authority and the Dubai Health Authority (DHA) decided to connect the licensing data of healthcare professionals with a Blockchain system in 2018. It enables potential healthcare providers to apply from anywhere in the world to receive a license [9], [46]. In securing the Diamond Trade, Dubai Multi Commodities Centre (DMCC) has announced the launch of its ecosystem in the Crypto Valley [9], [47].

While in transportation management, the Roads and Transport Authority (RTA) is implementing a Vehicle history Blockchain project. The project aims to provide a transparent record of the vehicle's past from the manufacturer to the scrap yard including car makers, dealers, regulators, insurance firms, buyers, sellers, and garages [41]. With regards to regulations, the Financial Services Agency of Japan has established Blockchain and Cryptocurrency Regulation in Japan including Cryptocurrency Regulation, Sales Regulation, Taxation, Money Transmission Law [9], [27].

B. How did the Initiatives Vary by Sector?

Blockchain initiatives and implementation in the Asian countries could be tracked through several sectors – Inter-sector, Financial, Supply Chain, Education, Administration, Immigration, Healthcare, Land, Transportation, and Social. The Tree-map analysis in Fig. 1 shows the fraction of Blockchain initiatives by sectors. Based on this categorization, it could guide and create a lesson learned for other researchers and practitioners.

1) Inter-sector Initiatives: Inter-sector initiatives include programs that have been implemented by the government that beneficial across all sectors such as financial, education, healthcare, and supply chain. These programs play a fundamental role to foster Blockchain initiatives in all sectors. Among vital inter-sector initiatives involves Blockchain strategy or master plan, council or governance, hackathon competition, assessment model, and survey on Blockchain.

As indicated in Fig. 2, this study revealed three countries – South Korea [49], Cyprus [38], and UAE [40] have published on their Blockchain Strategy in 2019, 2019, and 2018 respectively. The establishment of the Blockchain strategy sets out a roadmap for the launch of Blockchain technology and will provide economic opportunities for all sectors. Blockchain council or governance also being formed in South Korea [25] and Dubai [9], [44] as part of the efforts to embrace the new developments and practices of Blockchain at a global level. The Blockchain council or governance is intended to discover, confer current, and future applications of Blockchain in the country which consists of potential main players in the Blockchain industry. It incorporates government agencies, private companies, leading banks, free zones, and international Blockchain technology firms.

The study also discovered Hackathon competitions were organized in Singapore [9], [32], and Dubai [9], [45] on an annual basis. The Hackathon competitions were held to encourage technology understanding and adoption and inspire businesses to pursue creativity of the business model resulting from Blockchain and other emerging technology. Meanwhile, in Japan, assessment methodology and survey on Blockchain have been conducted in producing reusable evaluation form intended to be used not only to equate a traditional system with a Blockchain-based system but also to compare various Blockchain-based systems [34].

2) Financial: Regarding the financial sector, all five countries have already embarked on cryptocurrency regulation comprised of standard terms, sales regulation, taxation, and anti-money laundering laws. According to [51], as cryptocurrency seems to be a disruptor to conventional currency, regulations should be set up to prevent criminal misuse of this disruptive technology. Law enforcement must be proactively approached, and regulatory agencies are expected to inform the related agencies about how cryptocurrencies function, provide technical assistance, and promote discussion about topics of common concern.

Singapore has shown interest in the financial sector which explores Blockchain to clear and settle payments and securities through the project called Project Ubin and established the Innovation Center which includes both global banks and fintech start-ups. As the Blockchain is the backbone of the cryptocurrency such as Bitcoin and Ethereum, financial

firms are trying to cope with this technology due to high increase in demand for financial services and the enormous increase in competition worldwide in the financial sector [52].

3) Supply Chain and Immigration: Fig. 1 shows that three countries have developed programs in this sector in tracking the declaration of permits and other facilities for trade and logistics in Singapore [33], tender application in Japan [9], [37], and managing identity via biometric authentication in Dubai [9], [44]. This indicates that supply chain and immigration are among the potential sectors for Blockchain implementation.

4) Education and Healthcare: The analysis result also pointed out that the two countries have organized programs under Healthcare and education. In 2018, UAE has a collaboration with the Estonian company to develop a Blockchain-based licensing system that stores healthcare professional licenses [9], [46]. Estonia has used Blockchain as an extra layer of security to ensure the integrity of health records since 2016 [53]. In the meantime, through Blockchain, Singapore has demonstrated an educational initiative with academic credentials [33]. This means that among possible sectors for the adoption of Blockchain are healthcare and education.

5) Administration, Social, Land and Transportation: Other sectors that the Asian countries have embarked on Blockchain initiatives are Administration, Social, Land & Transportation. Realizing the advantages of Blockchain in improving efficiency, the Korean government has used this opportunity for the implementation of the employee system which is expected to be rolling out to the entire workforce [26]. With regards to social, a Blockchain-based voting system has been implemented in Gyeonggi-do Province which has increased the transparency of the voting result [48]. Meanwhile, Japan already working on a Blockchain-based Land Registry since 2018 [9], [36], and a Blockchain-based transportation tracking system has been developed by the UAE Roads and Transport Authority which aims to provide transparent records of the vehicle's history [41].

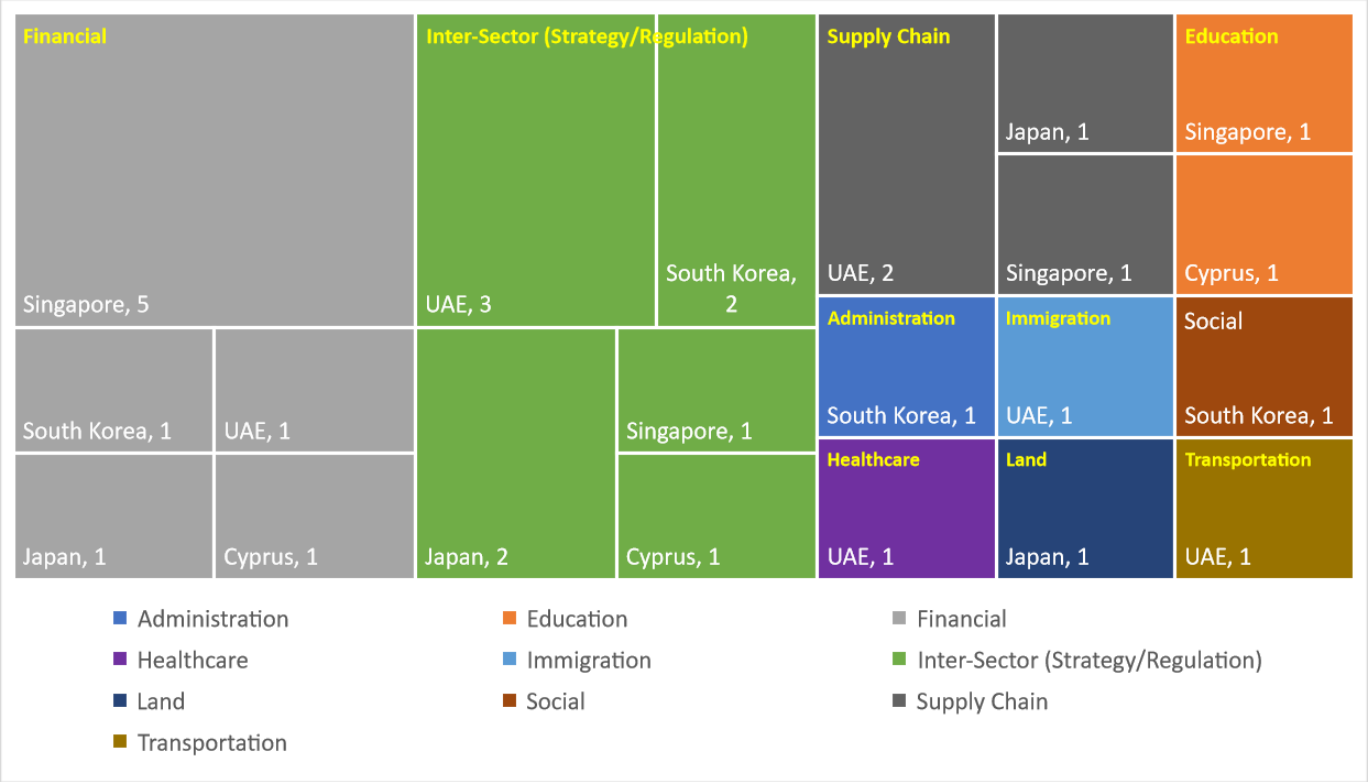


Fig. 1. Blockchain Initiatives by the Government in Asian Countries.

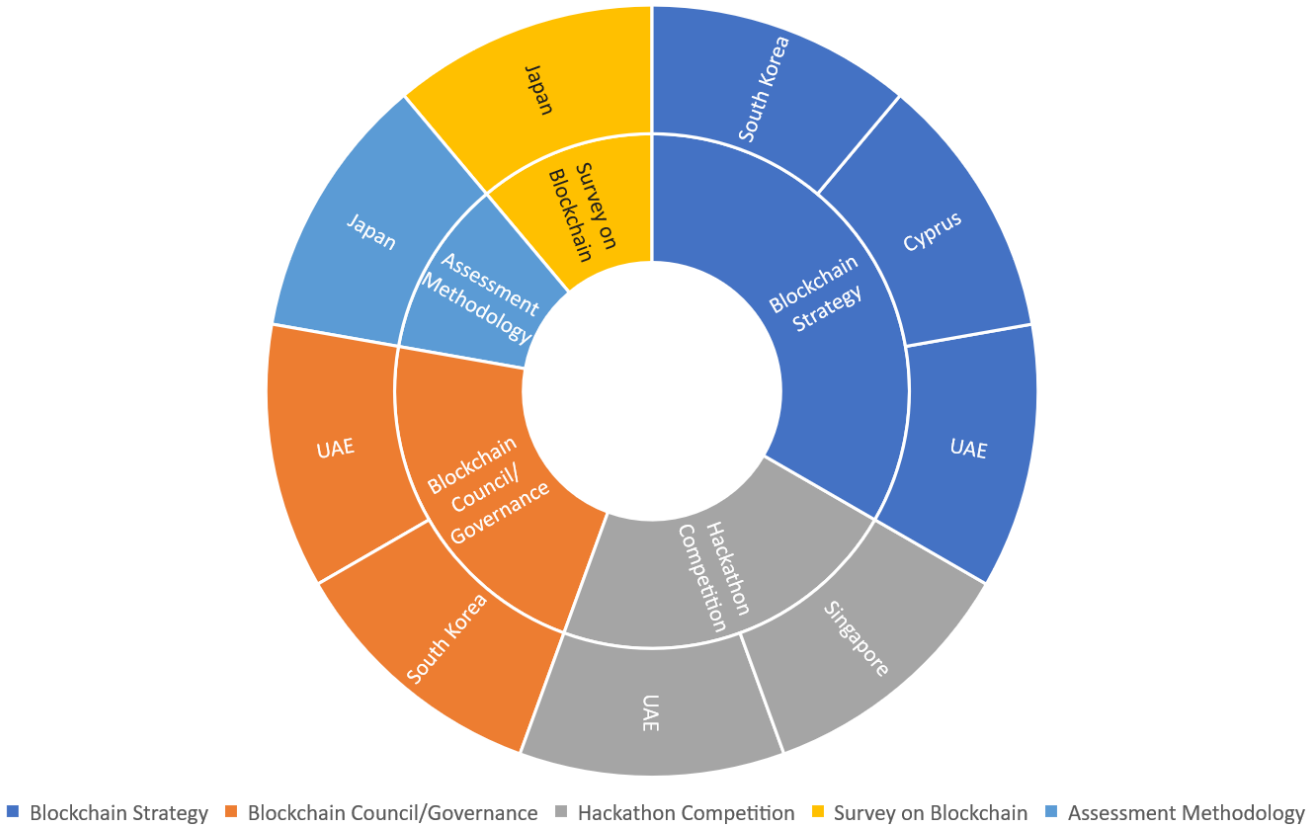


Fig. 2. Inter-Sector Blockchain Initiatives by the Government in Asian Countries.

V. Conclusion

To conclude, this study has achieved the aims to review Blockchain initiatives in the government of five countries in Asia that have achieved a Very High E-Government Development Index (EGDI) based on the United Nations evaluation in 2020. This study has been one of the first attempts to thoroughly review recent literature on Blockchain applications and initiatives by the government in Asian countries. This study has highlighted prominent Blockchain initiatives implemented by the government entities in these countries. Among noticeable sectors are financial, supply chain, immigration, education, healthcare, administration, social, land & transportation. This study has identified that the national strategy, Blockchain council, and innovation center establishment are important to set out a roadmap and to sustain the Blockchain initiatives in a long run. Besides, introducing standard methodology or guidelines, and appropriate laws and regulations are important to strengthening the process of supporting Blockchain implementation by the public sector. Additionally, the skill development programs and the organization of competition i.e. Hackathon are also vital to encourage Blockchain understanding and adoption. For future work, further studies should thoroughly review recent progress in other Asian countries to compare the movement of this disruptive technology. This study could guide and provide researchers and practitioners with the success stories and lessons learnt of the Blockchain initiatives by the government in Asian countries.

Acknowledgment

The study is financially sponsored by Fundamental Research Grant, vote no. 21H12, Universiti Teknologi Malaysia.

References

[1] M. Pilkington, “Blockchain technology: principles and applications,” in Research handbook on digital transformations, Edward Elgar Publishing, 2016.

[2] D. Drescher, Blockchain basics, vol. 276. Springer, 2017.

[3] M. Nofer, P. Gomber, O. Hinz, and D. Schiereck, “Blockchain,” Bus. Inf. Syst. Eng., vol. 59, no. 3, pp. 183–187, 2017.

[4] M. Crosby, P. Pattanayak, S. Verma, V. Kalyanaraman, and others, “Blockchain technology: Beyond bitcoin,” Appl. Innov., vol. 2, no. 6–10, p. 71, 2016.

[5] U. Nations, “E-Government Survey 2018,” 2018.

[6] A. Government, “The National Blockchain Roadmap,” 2019.

[7] T. Ahram, A. Sargolzaei, S. Sargolzaei, J. Daniels, and B. Amaba, “Blockchain technology innovations,” in 2017 IEEE Technology & Engineering Management Conference (TEMSCON), 2017, pp. 137–141.

[8] K. Christidis and M. Devetsikiotis, “Blockchains and smart contracts for the internet of things,” IEEE Access, vol. 4, pp. 2292–2303, 2016.

[9] Illinois, “Blockchain in Government Tracker.” [Online]. Available: https://airtable.com/universe/expsQEGKoZO2lExKK/blockchain-in-government-tracker. [Accessed: 08-Apr-2020].

[10] F. R. Batubara, J. Ubacht, and M. Janssen, “Challenges of blockchain technology adoption for e-government: a systematic literature review,” in Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, 2018, pp. 1–9.

[11] J. Yli-Huumo, D. Ko, S. Choi, S. Park, and K. Smolander, “Where is current research on blockchain technology?—a systematic review,” PLoS One, vol. 11, no. 10, p. e0163477, 2016.

[12] E. A. Franciscon, M. P. Nascimento, J. Granatyr, M. R. Weffort, O. R. Lessing, and E. E. Scalabrin, "A Systematic Literature Review of Blockchain Architectures Applied to Public Services," in 2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2019, pp. 33–38.

[13] A. Shahaab, B. Lidgey, C. Hewage, and I. Khan, "Applicability and Appropriateness of Distributed Ledgers Consensus Protocols in Public and Private Sectors: A Systematic Review," IEEE Access, vol. 7, pp. 43622–43636, 2019.

[14] A. Ojo and S. Adebayo, "Blockchain as a Next Generation Government Information Infrastructure: A Review of Initiatives in D5 Countries," Springer Int. Publ., vol. 32, pp. 283–298, 2017.

[15] UN, "UN E-Government Survey 2020," 2020.

[16] S. Bogart and K. Rice, "The blockchain report: welcome to the internet of value," Needham Insights, 2015.

[17] D. L. K. Chuen, Handbook of digital currency: Bitcoin, innovation, financial instruments, and big data. Academic Press, 2015.

[18] K. Fanning and D. P. Centers, "Blockchain and its coming impact on financial services," J. Corp. Account. Financ., vol. 27, no. 5, pp. 53–57, 2016.

[19] D. Tapscott and A. Tapscott, Blockchain revolution: how the technology behind bitcoin is changing money, business, and the world. Penguin, 2016.

[20] C. Okoli and K. Schabram, "Working Papers on Information Systems A Guide to Conducting a Systematic Literature Review of Information Systems Research," Sprout Work. Pap. Inf. Syst., vol. 10, no. 26, pp. 1–51, 2010.

[21] A. H. Jamie Berryhill, Théo Bourgery, "Blockchains Unchained : Blockchain Technology and its Use in the Public Sector," 2019.

[22] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," Tech. Rep. EBSE-2007-01, 2007.

[23] Q. He, Z. Li, and X. Zhang, "Data deduplication techniques," in Future Information Technology and Management Engineering (FITME), 2010 International Conference, 2010, vol. 1, pp. 430–433.

[24] F. Haneem, N. Kama, R. Ali, and A. Selamat, "Applying Data Analytics Approach in Systematic Literature Review: Master Data Management Case Study," Front. Artif. Intell. Appl., vol. 297, pp. 705–715, 2017.

[25] J. Young, "Here's How the Gov't of Seoul is Using a Public Blockchain in the Real-World." [Online]. Available: https://www.ccn.com/heres-how-the-govt-of-seoul-is-using-a-public-blockchain-in-the-real-world/. [Accessed: 05-Oct-2020].

[26] F. Erazo, "South Korean government agency to use blockchain-based employee ID system." [Online]. Available: https://cointelegraph.com/news/south-korean-government-agency-to-use-blockchain-based-employee-id-system.

[27] GLI, "Blockchain & Cryptocurrency Regulation," 2020. [Online]. Available: https://www.globallegalinsights.com/practice-areas/blockchain-laws-and-regulations/japan#chaptercontent2.

[28] MAS, "Project Ubin: Central Bank Digital Money using Distributed Ledger Technology." [Online]. Available: https://www.mas.gov.sg/schemes-and-initiatives/project-ubin.

[29] A. Mizrahi, "Singapore Government Joins IBM to Create Blockchain Innovation Center." [Online]. Available: https://www.financemagnates.com/cryptocurrency/innovation/singapore-government-joins-ibm-to-create-blockchain-innovation-center/. [Accessed: 05-Oct-2020].

[30] R3, Can Blockchain Future-Proof Supply Chains? - A Brexit Case Study. 2019.

[31] IMDA, "TradeTrust." [Online]. Available: https://www.imda.gov.sg/programme-listing/international-trade-and-logistics/tradetrust.

[32] IMDA, "Blockchain Challenge." [Online]. Available: https://www.imda.gov.sg/programme-listing/blockchain-challenge. [Accessed: 05-Oct-2020].

[33] C. Lago, "How Singapore is using blockchain outside of cryptocurrencies." [Online]. Available: https://www.cio.com/article/3291758/how-singapore-is-using-blockchain-outside-of-crypto-currencies.html.

[34] METI, "Document Titled 'Evaluation Forms for Blockchain-based Systems ver. 1.0' Released," 2017. [Online]. Available: https://www.meti.go.jp/english/press/2017/0329_004.html. [Accessed: 05-Oct-2020].

[35] METI, "FY 2018 Industrial Technology Survey," 2019.

[36] M. Gonzalez, "Blockchain in Japan," 2018.

[37] Nikkei-Asia, "Japan looks to blockchains for more secure e-government systems," 2017. [Online]. Available: https://asia.nikkei.com/Business/Biotechnology/Japan-looks-to-blockchains-for-more-secure-e-government-systems. [Accessed: 05-Oct-2020].

[38] R. of Cyprus, "A National Strategy for Cyprus," 2019.

[39] UNIC, "Blockchain Certificates." [Online]. Available: https://www.unic.ac.cy/iff/blockchain-certificates/. [Accessed: 11-Oct-2020].

[40] "Smart Dubai Blockchain." [Online]. Available: https://www.smartdubai.ae/initiatives/blockchain. [Accessed: 06-Oct-2020].

[41] UAE, "Blockchain in the UAE government," 2020. [Online]. Available: https://u.ae/en/about-the-uae/digital-uae/blockchain-in-the-uae-government. [Accessed: 11-Oct-2020].

[42] N. Lung, "Dubai launches world's first B2B Smart Commerce Platform for Free Zones," 2018. [Online]. Available: https://opengovasia.com/dubai-launches-worlds-first-b2b-smart-commerce-platform-for-free-zones/.

[43] P. Bhunia, "Government of Dubai to develop world's first gate-less border using biometrics and Blockchain," 2017. [Online]. Available: https://opengovasia.com/government-of-dubai-to-develop-worlds-first-gate-less-border-using-biometrics-and-blockchain/.

[44] "Global Blockchain Council." [Online]. Available: https://www.dmcc.ae/about-us/global-blockchain-council.

[45] "Smart Dubai Global Blockchain Challenge." [Online]. Available: https://www.smartdubai.ae/newsroom/news/smart-dubai-s-3rd-global-blockchain-challenge.

[46] O. Brytskyi, "Lessons to learn from the UAE in blockchain healthcare." [Online]. Available: https://espeoblockchain.com/blog/blockchain-healthcare. [Accessed: 11-Oct-2020].

[47] J. Bourne, "Dubai Multi Commodities Centre to launch 'crypto valley' ecosystem," 2020. [Online]. Available: https://blockchaintechnology-news.com/2020/01/dubai-multi-commodities-centre-to-launch-crypto-valley-ecosystem/. [Accessed: 11-Oct-2020].

[48] S. Das, "A South Korean Province Used Blockchain Tech for Resident Voting." [Online]. Available: https://www.ccn.com/south-korean-province-used-blockchain-tech-resident-voting/. [Accessed: 05-Oct-2020].

[49] R. Meyer, "The City of Seoul Will Create a Cryptocurrency for Citizen Rewards," 2019. [Online]. Available: https://www.coindesk.com/the-city-of-seoul-will-offer-blockchain-services-this-fall. [Accessed: 15-Oct-2020].

[50] Data61, "Distributed Ledgers - Scenarios for the Australian economy over the coming decades," 2017.

[51] J. Dewey, Blockchain & Cryptocurrency Regulation. 2019.

[52] M. Al-Essa, "The Impact of Blockchain Technology on Financial Technology (FinTech)," 2019.

[53] A. Alketbi, Q. Nasir, and M. A. Talib, "Blockchain for government services-Use cases, security benefits and challenges," in 2018 15th Learning and Technology Conference, L and T 2018, 2018, pp. 112–119.

Voice-Disorder Identification of Laryngeal Cancer Patients

Mr.G.B.Gour[1]
Department of Electronics and Communication Engineering
BLDEAs V.P.Dr.P.G.Halakatti College of Engineering and Technology, Vijayapur, Karnataka, India

Dr.V.Udayashankara[2]
Department of Electronics and Instrumentation Engineering
Sri Jayachamarajendra College of Engineering, Mysuru, Karnataka, India

Dr.Dinesh K. Badakh[3]
Department of Radiation Oncology
Sri Sri Siddhivinayak Ganapati Cancer Hospital, Miraj, Maharastra, India

Dr.Yogesh A Kulkarni[4]
Department of Medicine
Nargis Dutt Memorial Cancer Hospital, Barshi Makarastra, India

***Abstract*—This Previous studies have shown that much of laryngeal cancer-based work was carried out with a minimal set of linear features. Much of the work was focused on the study of larynx preservation, quality of life around radiotherapy, or surgery. The voice disorder database was not solely limited to laryngeal cancer. In the context of this, the paper proposes a non-invasive voice disorder detection of laryngeal cancer patients. The sustained vowel /a/ was recorded with 55 laryngeal cases and 55 healthy cases. Owing to the non-linearity property of the vocal cords, seven non-linear parameters along with biologically inspired 39 Mel-Frequency Cepstral Coefficients (MFCC) are extracted. This forms a laryngeal dataset of size 110X46. The wrapper method is used for better feature selection and to enhance the discriminating ability of the present work. The classification is carried out using a tuned support vector machine (SVM) with grid search and random forest (RF). The present work has shown an improved accuracy of 76.56% with SVM and 80% in the case of random forest. The forward selection of features along with the involvement of non-linear features has played a significant role in the better performance of the present system.**

***Keywords*—Support Vector Machine (SVM); random forest; Mel Frequency Cepstral Coefficients (MFCC); voice disorder detection; laryngeal cancer; non-linear features**

I. Introduction

One of the main categories of head and neck cancer is laryngeal cancer (LC). The oncologists can treat the LC by subjective evaluations and invasive diagnostic methods. LC patients are found to be reluctant in many situations. The stroboscope does not record the status of vocal function with cycle-to-cycle information. Moreover, the parameters and rating information provided by stroboscope are visual perceptual ratings and subject to reliability and validity errors. The video laryngoscope (VLS) has limitations of irritating patients and a need for topical anesthesia [1]. The voice analysis provides clinicians an alternative, non-invasive, and objective analysis tool in this regard. So far the LC based works were carried out using statistical and machine learning approaches with linear features only. The experimental study was conducted by [2,3] with 80-1925 LC cases. The statistical analysis was carried out with only few linear features. Jitter, shimmer, noise-to-harmonic ratio (NHR) and maximum phonation time (MPT) have shown a significant difference between cancer and dysfunctional groups in the results ($p<0.05$). The perturbation measures, NHR and MPT were verified with laryngoscopic evaluations for which LC cases are reluctant. To carry out the LC classification, a prototype distribution type map (PDM) was proposed by [4]. But, this PDM modeling was done using neural maps which was quite complex and time consuming with increased number of iterations. Moreover, efforts have been made to preserve the larynx after surgery, radiotherapy, and rehabilitation [10]. It is clear from the previous studies that,

- Much of the LC database used was found to contain LC cases along with dysphonia cases.
- Only linear based voice acoustic features were used.
- Much of the work was focused on voice analysis of LC cases to study the quality of life and preservation of larynx w.r.t surgery or radiotherapy.
- Much of the study was carried out on LC were lacking voice disorder analysis using machine learning approaches.

The main objectives of the present study are:

- To create and use a specific LC database concerned with the pathologies related to adjacent regions of the larynx.
- To assess the performance of linear and non-linear descriptors to discriminate between the patients with LC pathologies and healthy cases.
- To develop a cost-effective (using open-source platform Ubuntu 15.0 and GNU Octave 4.0) and non-invasive voice acoustic classification tool.

Hence, the paper presents a non-invasive voice disorder detection of LC cases with optimized feature selection using forward selection method and, classification by tuned SVM

and random forest. Rest of this paper is organized as follows. Section II describes the background study with reviews in the existing research, Section III illustrates materials and proposed methodology for LC detection, Section IV depicts extensive experimental evaluation for the proposed method and finally Section V presents concluding remarks with future scope.

II. Background Study

Researchers have adopted many clinical and acoustic methods in classifying the LC cases but with little attention to detection of LC. The clinical methods have been suffering from patients comfortless to laryngoscopy and variations in measurements. A more diversified LC database was adopted in the following studies. Data dependent random forest was proposed by [5] in fusing knowledge with increased classification accuracy. Here, 110 LC cases with different pathologies (tumors, polyps, cysts, papillomata, keratosis, carcinoma and paralysis) were included in the study. The auto-associative neural network was used to differentiate LC cases (139-laryngitis cases, 211-hyper-functional dysphonia and 212-recurrent laryngeal nerve paralysis) from healthy group. Frequency-based features like MFCC, Cepstral-based features, HNR, and LP-based parameters were used to form a 14-dimensional vector for each subgroup. Over 37 linear features were used to train neural network with 87.5% accuracy [6]. The authors believes that with more protuberant features describing dynamics of vocals cords along with better feature selection methods can enhance the accuracy of the LC detection system.

However, much of the experimental work was carried out to understand the impact of radiotherapy on LC cases at particular periodic follow up. The retrospective study was conducted on 115 early-stage (in-situ, T1-T2) LC cases to assess the improvements in visual, acoustic, and patient-reported findings [7]. A similar study was conducted on laryngopharyngeal cancer cases [8]. To determine the effect of supra-clavicular RT on the physiology and functioning of the vocal fold, an experimental study was performed on 29 female patients diagnosed with breast cancer who underwent supra-clavicular RT, reported at intervals of 1 and 6 months before and after treatment. [9]. In all these radiation-based studies, a limited number of linear voice acoustic parameters were used with an average significant variation. These variations were used in the assessment of the early stage of LC or analysis of voice quality. The work was carried out to know the impact of voice rehabilitation on laryngeal cancer patients after radiotherapy using jitter, shimmer, quality of life scores (QOL), and voice handicap-index (VHI). These studies have provided a better road map towards multi-classification among LC cases. But, no specific voice rehabilitation was found to be necessary for laryngeal cancer patients after radiotherapy as there was no statistical significance found with these parameters [10].

III. Materials and Methods

A. Database

The present study involves cases with laryngeal pathology. The voice samples of laryngeal cancer patients came for radiotherapy, were collected after taking the written consent from each case. The cases have given sufficient information about the non-invasive procedure being followed while recording the voice samples. The ethical approval was taken in advance from the Sri Siddhivinayak Ganapati Cancer Hospital, Miraj, and Nargis Dutt Memorial Cancer Hospital, Barshi, (Maharastra), India. The sustained vowel /a/ was recorded using an Omni-directional microphone for 1-3 seconds. The recording has performed at a sampling frequency of 44.1 kHz with a 16-bit resolution using the Praat software. The operated laryngeal cases were not included in the recording. A total of 55 laryngeal cases in the age group of 62.8±10.8, having 49 males and 6 female cases are included as shown in the Table I. The voice samples for the control group of 55 cases are collected in the age group of 63.1±10.9. This forms the voice dis-order database for LC having a total of 110 cases.

B. Proposed Algorithm

The methodology adopted in the present work is as shown in Fig. 1. Pre-processing is optionally adopted in the present work based on the type of classifier used. As shown in Fig. 1, the speech frames of 25 ms with 50% overlap are used throughout work. A pre-emphasizing filter with $\alpha = 0.97$ is used to boost the high-frequency components.

$$y_p(n)=\mathrm{y}(n) - \alpha \mathrm{y}(n-1) \quad (1)$$

Then, the speech enhancement is carried out using a two-step noise reduction method (TSNR) using a Wiener filter [11]. This stage is followed by the extraction of the MFCC and non-linear features as discussed in the section C. To validate the irrelevance to the present function, the features were investigated. By utilizing forward function selection, the identification of voice dysfunction is optimized and comparative analysis is established between optimized and non-optimized features with a strongly validated mix. In classification process, SVM with grid search and random forest methods are adopted owing to their ability in dealing with low dimensional features space.

TABLE I. Non-Operated LC Cases Included in the Study

Voice Pathologies	**Number of Cases**
Ca-larynx	24
Ca-laryngo-pharynx	1
Ca-Supraglottis	7
Ca-epiglottis	12
Ca-Pharynx	4
Ca-Hypopharynx	6
Ca-Cricopharynx	1
Total	**55**

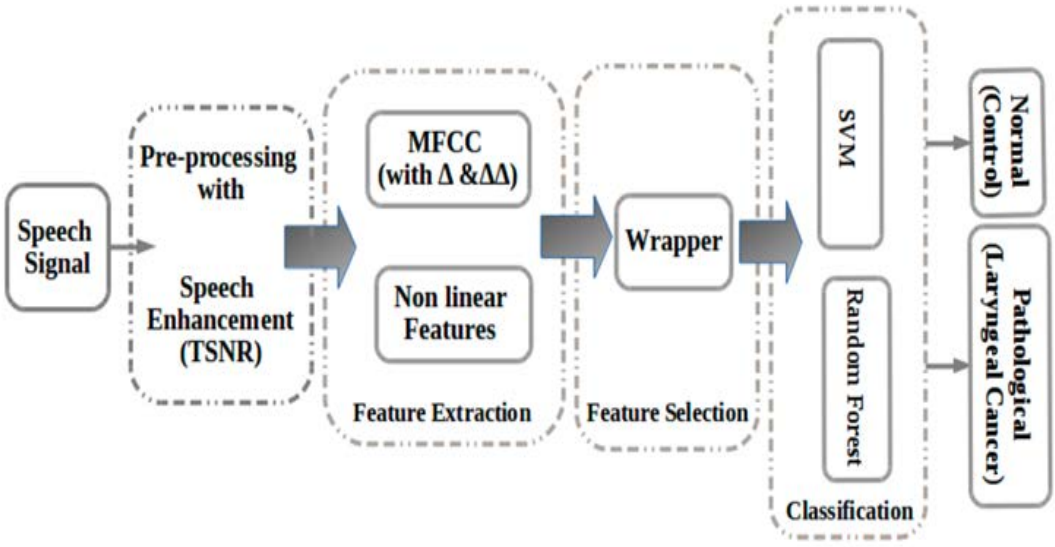


Fig. 1. Methodology of the Present Work.

C. Feature Extraction

1) MFCC: In the feature extraction, linear and biologically inspired 39-dimensional MFCC features are extracted per frame as shown in Fig. 2. Here, the 13 MFCC is derived from each of the speech frames by taking the discrete Fourier transform (DFT). To obtain its power spectrum, it is then passed through a triangular filter bank with 24 filters, uniformly positioned on the Mel scale. Log energy is computed for each of the banks, which is found to be sensitive to small variations in the articulatory movements. Then the MFCC of 13 coefficients is obtained by using DCT. A total of the 39-dimensional vector of MFCC are obtained with its derivatives. The 39-dimensional vectors are fed to the supervised SVM for the classification of LC pathology or healthy [12].

2) Non-linear features: It is clear from the previous studies that, the linear time-invariant property is no longer holding good for the non-linear structure of vocal cords and their dynamic behavior. The non-linear features are capable of producing the trajectories in phase space from the vibrations produced by a dynamical system like the vocal folds. The 7-dimensional non-linear data subset was obtained by the extraction following features [13,14,15];

A. Mutual information (MI): It provides a flexible approach in evaluating dynamical variables by using the method of delays. It refers to the reconstruction of phase space of voice signal using with minimum delay, called the first minimum of mutual information. The MI is computed by using the equation (2).

$$S=\sum_{ij} p_{ij}(\tau)\ln\left(\frac{p_{ij}(\tau)}{p_i p_j}\right) \quad (2)$$

3) False nearest neighborhood (FNN): If there are two nearest neighbours $\vec{s_i}$ and $\vec{s_j}$ in dimension m with distance between them is $\left\|\vec{s_i}-\vec{s_j}\right\|$ then, $\overrightarrow{s_{i+1}}$and $\overrightarrow{s_{j+1}}$are the maps of the respective $\vec{s_i}$and $\vec{s_j}$in m+1 dimension. The divergence rate of these points while travelling from dimension m to m+1 is shown by (3),

$$R_i=\frac{|s_{i+1}-s_{j+1}|}{\|\vec{s_i}-\vec{s_j}\|} \quad (3)$$

These two points become false neighbors if the distance between them exceeds a certain threshold. Then the condition for maximum embedding dimension is the fraction of points for Ri> Rt threshold gives the estimation for embedding dimension m. Here, pi is the probability of finding a time series value in the i-the interval, and pi (τ) is the joint probability of pi and pj corresponding two points. MI measures the mutual dependence of the points pi and pj.

4) Correlation Sum and Dimension (CD): The correlation dimension quantitatively describes the complexity or irregularity of the trajectory in phase space. This irregularity is confined to the correlation of two points on the trajectory known as the correlation dimension. The convergence of finite correlation can be obtained by straight-line fitting of the log-log plot of the correlation sum which is given by equation (4).

$$C_i^m(r)=\sum_{i=1}^{N}\left(\frac{2}{N(N-1)}\sum_{j=i+1}^{N}\Theta\left(r-\left\|\vec{s_i}-\vec{s_j}\right\|\right)\right) \quad (4)$$

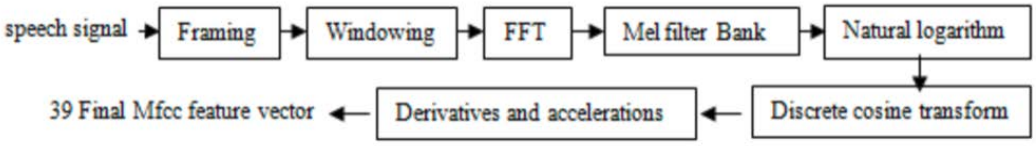


Fig. 2. Estimation of MFCC.

The correlation dimension quantitatively describes the complexity or irregularity of the trajectory in phase space. This irregularity is confined to the correlation of two points on the trajectory known as the correlation dimension. The convergence of finite correlation can be obtained by straight-line fitting of the log-log plot of the correlation sum which is given by equation (5),

$$D_{\mathrm{TT}}(r) = \frac{C(r)}{\int_0^r \left(\frac{C(r')}{r'}\right) \mathrm{dr'}} \quad (5)$$

Information dimension or correlation dimension of order 2 (CE_2): Information dimension CE_2 can be estimated from the modified correlation sum.

5) Largest Laypunov Exponent (LLE): This feature depicts the average divergence rate of the neighbor trajectories as calculated by Rosenstein's method. The separation or the average divergence of points in a trajectory is d (t) = C $e^{\lambda 1 t}$. Here λ_1 is the LLE and C is a constant.

6) Renyi Entropy (RE_1, RE_2): In a dynamic system like vocal cords, Renyi entropies describe the loss of information in time. There will be an evolution of a nearby point in phase space to far points. In random systems, Renyi entropies tend to infinity.

At the end of this section, 39-dimensional MFCC and 7-dimensional non-linear parameters provide a 46-dimensional feature vector.

D. Feature Selection

The feature selection methods help in revising the model with reduced complexity and optimized accuracy. The commonly categorized feature selection methods are filtering, wrappers, and embedding methods [16]. The wrapper method is based on a predictive model. The forward selection (FS) method is used as it is less prone to overfitting and having low computational cost, than backward selection. The FS method provides a reduced, un-correlated significant feature sub-set with average computational cost [17,18]. The steps involved in the forward selection are:

- Begins with null model M_0 with zero predictors.
- Pre assuming the significant levels S_{add} and S_{drop} to add and drop features respectively.
- For k = 0, ..p-1 (0 to 9 features), means considering all (p-k) models that augument with the predictors in M_k.
- Augumenting of significant model for S_{add} < cross-validation error (CV) and droping the previously added predictor for S_{drop} > CV error.
- The last step is replicated before a final optimum set of characteristics is obtained among M_1 ,M_2 , . . .M_p.

E. Classification Algorithms

Two supervised classification algorithms, SVM and random forest are used in this work owing to the size of the present dataset and number of features. Random forest is sensitive to the small changes in training data (bagging) that enable us to include in the present work. Both algorithms are fast in training and testing data-sets. In order to infer the most productive classifier for the detection of voice disorders, the comparative analysis is compiled between their success rates.

1) Support Vector Machine with a grid search: SVM is a supervised binary classifier using kernel tricks to trace the best hyperplane with the maximum margin between two classes. In a higher-dimensional space, the non-linear data is mapped, where the kernel becomes linear. Commonly used kernels with SVM are linear, polynomial, radial, and sigmoid. The SVM is used to classify the LC cases due to low computational costs and less prone to over-fitting. The data comprising of n-dimension feature vectors are first labeled by using the Audacity tool, scaled and normalized before feeding to the SVM. In all the cases, SVM-kernel tricking is employed with the help of a grid search. SVM is tuned with hyper-plane parameters with different kernels, gamma, and C values [19, 20]. The grid search will build and evaluate the model for each combination of hyper-plane parameters. Then, best hyper-plane parameters with good cross-validation (CV) accuracy. In the present work 10-fold, cross-validation is used. The C (1 to 500), gamma (0.00003 to 0.002), and the kernels are traced to find the best fit of these hyper-parameters. The grid search is separately performed in using different data-sets during experimental investigations as discussed in section III.

2) Random forest: Random forest is an ensemble of a set of more number of individual and uncorrelated decision trees. Because of RF's sensitivity to changes in training data, it allows each tree to randomly sample from the main data-set with replacement known as bagging. Hence, the bagging process provides us with a sample size of N, which is less in size than the actual whole data-set. But with replacement, each of the attributes may be repeated in each chunk. In the model, this forces much more variance among the trees and eventually results in less correlation between trees. Therefore, trees that are not only learned on different sets of data (bagged) but also use different features will make decisions eventually [20,21]. The typical steps to be followed for the core working of RF are;

- For a given training data of {(a_1,b_1), (a_2,b_2)…,(a_n,b_n)} with X_i predictors which corresponds to the root node.
- Each non-terminal node splits up into two descedent nodes based on the value of one of the predictor variables.
- Categorical predictor variable creates partitions on either side of the split point with different sub-set of categories.
- 4. This process proceeds recursively before the conditions for stopping are satisfied.
- 5. For all events in the terminal nodes, an estimated value occurred by averaging the computation.
- Response of the most frequent class for classification problems.

An in-depth tree is kept at 5 to address the over-fitting of data.

F. Evaluation Process

By using precision, sensitivity, specificity, precision, and the Area Under Curve (AUC), the efficacy of the proposed algorithm is evaluated. The sensitivity evaluates the likelihood of pathological samples to be detected by the algorithm. In turn. The specificity assesses the algorithm's ability to classify typical samples. Precision reflects the percentage of pathological samples from the pathological class that is well identified. Besides, accuracy measures the correct classification rate of the algorithm. AUC assesses the capacity to differentiate between the normal and abnormal samples. The AUC offers an alternative means of calculating the performance of the method suggested. These measures are based on the following notions:

$$\text{Accuracy}=((\text{TP}+\text{TN})/\text{total}) * 100 \quad (6)$$

$$\text{Sensitivity(Recall)} = (\text{TP}./(\text{TP}+\text{FN})) * 100 \quad (7)$$

$$\text{Specificity}=\text{TN}/(\text{TN}+\text{FP}) * 100 \quad (8)$$

$$\text{Precision}=(\text{TP}/(\text{TP}+\text{FP})) * 100 \quad (9)$$

$$\text{AUC}=0.5x(\text{Specificity}+\text{Sensitivity}) \quad (10)$$

IV. Experimental Results with Discussion

According to Fig. 1, the speech signal related to sustained vowel /a/ is framed using a 25 ms window and 50% overlapping. Then, for each 25 ms speech frame, a 49-dimensional feature vector is extracted as discussed in section. Now with 110 cases, gives 110X46 sized main dataset. In this research, the dataset was divided a 70% of the training data and 30% for validation. All simulations were conducted in GNU-Octave using python.

The forward selection has been applied to a 46-dimensional data-set (MFCC & NL) and also on 39-dimensional MFCC separately. FS will select a sub-set of maximally performing 10 features, which improves accuracy with minimal computational time. In the present work, FS iteration is supported with 10 features due to more computational time, as shown in Table II. The best score of 0.80 is achieved. MFCC along with two non-linear features, CD and LLE are found to be significant. Where in the above Table II, Δn, and ΔΔn denote n^{th} coefficient of first and second derivatives of MFCC respectively. Similarly, 10-MFCC significant parameters are selected using the wrapper method.

After performing forward feature selection, the experiments are conducted in two ways, with MFCC and Non-linear parameters and with only linear parameters MFCC to assess the impact of having non-linear features in the prediction of laryngeal cancer.

A. Evaluation based on SVM Performance

The 110 by 46 sized data-set is applied to SVM (with grid search). Table III shows the SVM performance with tuned hyper-parameter (C, gamma values) before and after applying the wrapper. It is observed that, hyper-parameters selected with gamma = 0.0001 with radial basis (RB) kernel for the whole dataset (C=30) and linear kernel (C=100) in case of MFCC data.

Table III shows performance of SVM in terms of accuracy (%), sensitivity (%), specificity (%), area under curve (AUC) and precision (%). The forward selection has shown a significant impact of involving non-linear features in the experimental study. The SVM with (RB) kernel has shown an improved accuracy from 72.85% to 75.66% for the whole dataset and accuracy enhancement from 62.82% to 66.66% in the case of the MFCC dataset. Thus wrapper method along with (RB) kernel of SVM has played a significant role in optimizing the dataset. This resulted in good discrimination of laryngeal cancer. Fig. 3 shows an increase in accuracy with feature selection (optimized) related to MFCC dataset from 62.82% to 66.66%. The maximum accuracy of 75.66% is achieved with a dataset having 2.81% optimization. Moreover, an average sensitivity of 74% has shown better discrimination ability of SVM with a complete dataset along with a better AUC rate of 76.56% as shown in Fig. 4.

B. Process Evaluation based on Random Forest Performance

Table IV shows the performance of random forest presented in terms of accuracy (%), sensitivity (%), specificity (%), the area under the curve (AUC), and precision (%). It is clear from Fig. 5 and 6, that with the whole dataset, 3.44% of optimization is achieved. This is responsible for the maximum enhancement in the accuracy from 76.56% to 80% along with a maximum AUC rate of 79.80%. Hence, from the experimental observations, it is evident that random forest is showing better accuracy (80%) and discriminating ability (79.80%) with a complete dataset.

TABLE II. Average Scores of Features using Forward Selection Applies on the Complete Dataset

Average Score	Feature Names
0.72	ΔΔ8
0.70	MFCC10,ΔΔ8
0.76	MFCC10, MFCC11, ΔΔ8
0.74	MFCC2, MFCC10, MFCC11, ΔΔ8
0.74	MFCC2, MFCC4, MFCC10, MFCC11, ΔΔ8
0.76	MFCC2, MFCC4, MFCC10, MFCC11, ΔΔ8, ΔΔ-log-energy
0.77	MFCC4, MFCC10, MFCC11, ΔΔ8,ΔΔ12, ΔΔ-log-energy,CD
0.78	MFCC10, MFCC11, Δ8, ΔΔ8,ΔΔ12, ΔΔ-log-energy,CD,LLE
0.77	MFCC8,MFCC10, MFCC11, Δ8, ΔΔ8,ΔΔ12, ΔΔ-log-energy,CD,LLE
0.80	MFCC8,MFCC10, MFCC11, Δ6, Δ8, ΔΔ8,ΔΔ12, ΔΔ-log-energy,CD,LLE

TABLE III. SVM PERFORMANCE BEFORE AND AFTER APPLYING FORWARD SELECTION METHOD

Wrapper	Features	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	Precision (%)	C	Kernel
Without Feature selection	MFCC + Non-linear parameters	72.85	73.33	72.00	72.66	82.5	30	Radial Basis
With Feature selection	MFCC parameters	62.82	61.82	65.21	63.52	80.95	100	Linear
Without Feature selection	MFCC + Non-linear parameters	75.66	74.00	79.17	76.56	83.33	30	Radial Basis
With Feature selection	MFCC parameters	66.66	64.81	70.83	67.82	83.33	50	Linear

TABLE IV. PERFORMANCE OF RANDOM FOREST BEFORE AD AFTER APPLYING FORWARD SELECTION METHOD

Wrapper	Features	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	Precision (%)
Without Feature selection	MFCC + Non-linear parameters	76.56	68.75	72.72	70.74	84.61
With Feature selection	MFCC parameters	68.57	60.71	63.63	62.17	80.95
Without Feature selection	MFCC + Non-linear parameters	80.00	80.43	79.17	79.80	88.09
With Feature selection	MFCC parameters	69.33	68.62	70.83	69.72	83.33

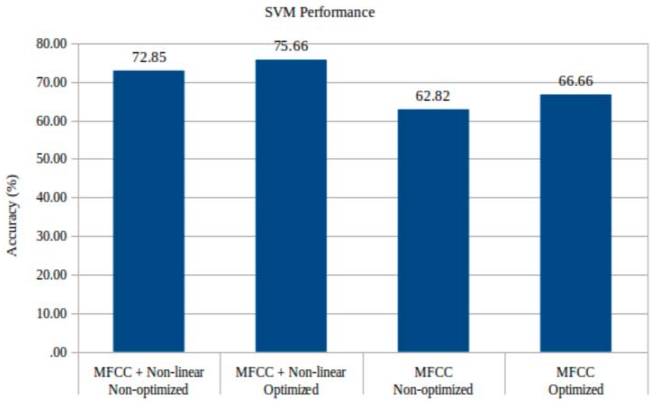


Fig. 3. SVM Performance in Terms of Accuracy(%).

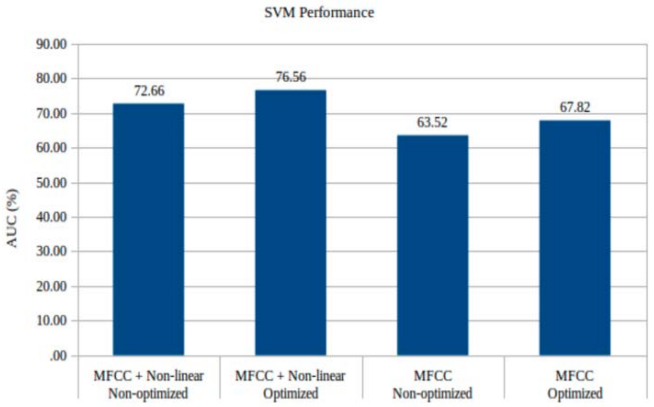


Fig. 4. SVM Performance in Terms of AUC(%).

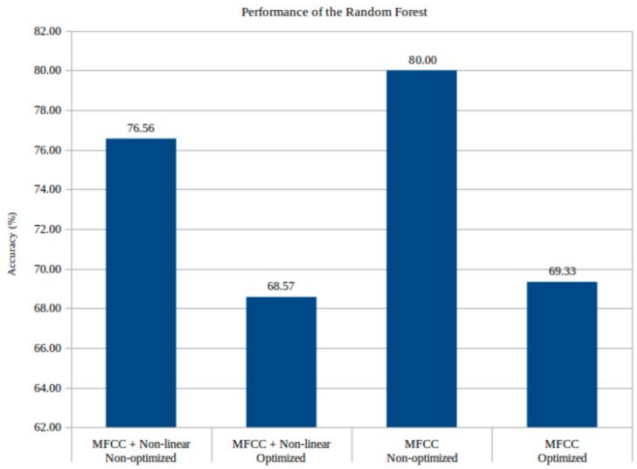


Fig. 5. Performance of Random Forest in Terms of Accuuracy(%).

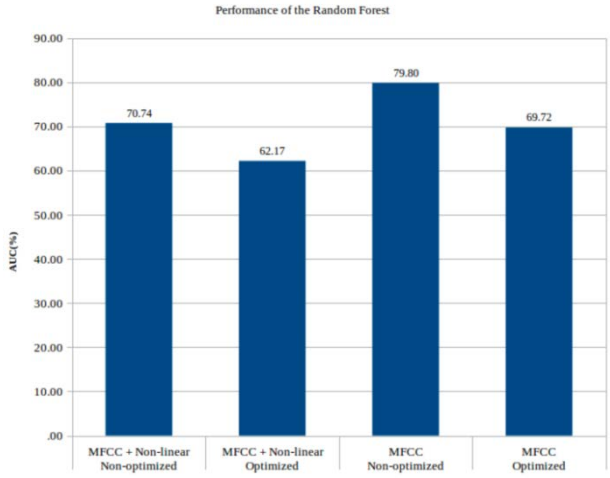


Fig. 6. Performance of Random Forest in Terms of AUC(%).

V. Conclusion and Future Scope

The paper presents a non-invasive laryngeal cancer detection platform. Both linear and non-linear features are tested over 110 LC cases. Features are optimized with the forward selection method. The system's performance is evaluated with SVM and random forest. The main findings are:

- Hyper-parameter tuning using grid search, helped in the identification of significant features.
- Better optimization of features resulted in improved accuracy of 76.56% with SVM and 80% in the case of random forest.
- Better discriminating abilities are observed with 75.66% with SVM and 79.80% with random forest.

The observations suggest that the SVM with RB kernel and forward selection of features involving non-linear parameters can be used for the development of a more enhanced non-invasive diagnostic tool for laryngeal cancer. With such findings along with more significant features and deep learning techniques, a better diagnostic tool can be developed for the detection of laryngeal cancer.

Acknowledgment

The ethical approval was taken in advance from the Sri Siddhivinayak Ganapati Cancer Hospital, Miraj, and Nargis Dutt Memorial Cancer Hospital, Barshi, (Maharastra), India. The written informed consent was taken from each case with proper understanding of procedures in their local languages. All actions carried out in studies involving human participants were consistent with the ethical standards of the ethics review committee for Institutional Research. This research received no specific grant from public, commercial, or nonprofit organizations funding bodies.

References

[1] G. Cummings: Otolaryngology:" Head & Neck Surgery", 4th ed. Mosby.Inc,2005.

[2] Om Karlsen, Lorentz Sandvik, John-Helge Heimdal, Hans Jorgen Aarstad, " Acoustic Voice Analysis and Maximum Phonation Time in Relation to Voice Handicap Index Score and Larynx Disease", Journal of Voice, Vol. 34, No. 1, 2020.

[3] Young Ae Kang,Jae Won Chang,Ho-Ryun Won,Bon Seok Koo, " Comparison Between Early Glottic Carcinoma and Epithelial Dysplastic Lesions of the Vocal Fold Via Voice Analysis", 24 April 2020, https://doi.org/10.1016/j.jvoice.2020.03.005.

[4] Stefan Hadjitodorov, Boyan Boyanov, and Bernard Teston,"Laryngeal Pathology Detection by Means of Class-Specific Neural Maps", IEEE Transactions On Information Technology in Biomedicine, Vol. 4, No. 1, March 2000.

[5] Adas Gelzinis, Antanas Verikas, Evaldas Vaiciukynas, Marija Bacauskiene, "Exploring sustained phonation recorded with acoustic and contact microphones to screen for laryngeal disorders", 978-1-4799-4527-6/14/2014 IEEE.

[6] Daria Hemmerling, "Voice Pathology Distinction Using Autoassociative Neural Networks", 2017 25th European Signal Processing Conference (EUSIPCO).

[7] Matthew R. Naunheim, Jonathan Garneau, Chris Park, Linda Carroll, Leanne Goldberg, Peak Woo, "Voice Outcomes After Radiation for Early-Stage Laryngeal Cancer", Journal of Voice, Volume 34, Issue 3, May 2020, Pages 460-464.

[8] Priya Dubey, Jacqueline B. Fernandes, Mahesh Bhat,"Acoustic Analysis of Voice in Laryngopharyngeal Cancers Pre and Post Radiotherapy", Indian Journal of Otolaryngology and Head & Neck Surgery, July 2020.

[9] Saime Sagiroglu, Neslihan Kurtul, "The Effect of Supraclavicular Radiotherapy on Acoustic Voice Quality Index (AVQI), Spectral Amplitude and Perturbation Values", Journal of Voice, Volume 34, Issue 4, July 2020, Pages 649.e7-649.e13.

[10] Taito, M., Taito, S., Banno, M. et al. "Voice rehabilitation for laryngeal cancer after radiotherapy: a systematic review and meta-analysis", Eur Arch Otorhinolaryngol 276, 1573–1583 (2019). https://doi.org/10.1007/s00405-019-05452-2.

[11] Cyril Plapous, Member, IEEE, Claude Marro, and Pascal Scalart," Improved Signal-to-Noise Ratio Estimation for Speech Enhancement", IEEE Transactions on Audio, Speech and Language Processing, Institute of Electrical and Electronics Engineers, 2006.

[12] HyunBum Kim, Juhyeong Jeon, Yeon Jae Han , YoungHoon Joo , Jonghwan Lee ,Seungchul Lee and Sun Im, „"Convolutional Neural Network Classifies Pathological Voice Change in Laryngeal Cancer with High Accuracy", J. Clin. Med. 2020, 9, 3415; doi:10.3390/jcm9113415.

[13] Jack J. Jiang, Yu Zhang, and Clancy McGilligan," Chaos in Voice, From Modeling to Measurement", Journal of Voice, Vol. 20, No. 1, pp. 2–17, 0892-1997/$32.00, 2006 The Voice Foundation, doi:10.1016/j.jvoice. 2005.01.001.

[14] Patricia Henríquez, Jesús B. Alonso, Miguel A. Ferrer, Carlos M. Travieso, Juan I. Godino L lorente, and Fernando Díaz-de-María, " Characterization of Healthy and Pathological Voice Through Measures Based on Nonlinear Dynamics", IEEE Transactions On Audio, Speech, And Language Processing, Vol. 17, No. 6, August 2009.

[15] Iasonas Kokkinos, Student Member, IEEE, and Petros Maragos, Fellow, IEEE," Nonlinear Speech Analysis Using Models for Chaotic Systems", IEEE Transactions On Speech And Audio Processing, Vol. 13, No. 6, November 2005.

[16] Girish Chandrashekar,Ferat Sahin,"A survey on feature selection methods", Computers & Electrical Engineering,Volume 40, Issue 1, January 2014, Pages 16-28Brant, Rollin. Forward Selection. MDSC 643.02 Lecture Materials. https://www.stat.ubc.ca/~rollin/teach/643w04/lec/node41.html on July 7, 2018.

[17] Gabriel Solana-Lavalle, Juan-Carlos Galan-Hernandez, Roberto Rosas-Romero " Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features",Biocybernetics and Biomedical Engineering, 40 (2020) 505-516.

[18] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin," A Practical Guide to Support Vector Classification", Department of Computer Science, National Taiwan University, Taipei 106, Taiwan, May 19, 2016, http://www.csie.ntu.edu.tw/~cjlin.

[19] Iwan Syarif , Adam Prugel-Bennett, Gary Wills, " SVM Parameter Optimization Using Grid Search and Genetic Algorithm to Improve Classification Performance", TELKOMNIKA, Vol.14, No.4, December 2016, pp. 1502~1509, ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v14i4.3956.

[20] Adele Cutler, D. Richard Cutler and John R. Stevens,"Random Forests", Chapter in Machine Learning, January 2011, DOI: 10.1007/978-1-4419-9326-7_5.

[21] Zuzana Dankovi cova, David Sovak, Peter Drotár and Liberios Vokorokos, "Machine Learning Approach to Dysphonia Detection", Appl. Sci. 2018, 8, 1927; doi:10.3390/app8101927.

An Improved Time-Based One Time Password Authentication Framework for Electronic Payments

Md Arif Hassan[1]*, Zarina Shukur[2], Mohammad Kamrul Hasan[3]
Center for Cyber Security, Faculty of Information Technology
National University Malaysia (UKM), 43600 UKM, Bangi, Selangor, Malaysia

***Abstract*—One-time Password is important in present day scenario in the purposes of improving the security of electronic payments. Security sensitive environment or perhaps organization avoid the resources from unauthorized access by allowing different access control mechanism as user authentication. There are several safety issues in one Password based authentication. However, studies show that OTP sent over SMS are causing different causes and issues, which lead to precious time, delay in transaction. User authentication can be raised with more levels within the procedure of multi-factor authentication scheme. Time-based One-time Password and biometrics are one of the widely accepted mechanisms that incorporate multi-factor authentication. In this paper, we approach the Time-based OTP authentication algorithm with biometric fingerprints to secure an electronic payment. This algorithm uses a secret key exchanged between the client and the server and uses a certain password through the algorithm. The shuffle of the TOTP approach better wear by screening the key as being a QR code, as revealed in the majority movable applications are able to read. It offers confidentiality at the application level within the system to protect user credential within equal entities (the user and the server) for preventing brute force and dictionary attacks. Thus, the proposed system design is possible for users because of the lack of the concern of holding its own hardware token or additional charges from the short message service. Our suggested approach has been found to improve safety performance substantially compared to existing methods with regard to authentication and authorization. This research hopes to boost research effort on further advancement of cryptosystems surrounding multi-factor authentication.**

Keywords*—Electronic payments; One Time Password (OTP); Quick Response (QR) code; Time based One Time Password (TOTP)

I. Introduction

Mostly all online services and the websites are today implementing multi step authentication to offer protection to the customers of theirs. Multi-factor authentication is a technique of the digital device access influence that a person is able to pass effectively showing different authentication stages. In this, rather than asking only the individual piece of info as passwords, users are requested to provide a number of extra info and that helps make it harder for any intruder to bogus the identity of the real user. This info could be an OTP that will be delivered by the server on the registered mobile of consumer or perhaps there could be certain security concerns. This particular procedure makes it hard for the opponent to access the internet account even if the assailant understands the username as well as password of the user. This more info is able to consist of different aspects as fingerprints, security tokens [1], biometric authentication, and so on. It has emerged as an alternative means to enhance protection by needing the user to give over one authentication factor instead of just one password. Authentication issues are of those kinds: Knowledge - something which the person knows [2], like a password and a username; Possession - a thing the person has specifically a hardware token [3]; Inherence - a thing verifies the person is, like fingerprints, iris, facial recognition, palm print [4-7]. Biometrics technology enjoys a wider acceptance because of including fingerprint biometrics and more user-friendly applications on digital devices [5]. Two forms of biometric authentication are available, respectively physiological and behavioral approaches [8-9]. Fingerprint is the most popular biometric process. As authentication is highly user friendly, it is increasingly used to login functionality in fingerprints [10]. Among the many other applications of theirs, QR codes are popular for the multi-factor authentication to transmit info through the authenticating device on the mobile device which is accredited as being an AIM Standard, an ISO standard and a JIS Standard [11]. In the beginning, the QR code is created to be utilized in the auto industries. However, these days, it has been popular in the ad so that a customer is able to utilize the smartphone and scan to find out much more info about the marketed products.

The barcode scanner programs are developed that are suitable for smartphones as IOS and android. QR Code is a kind of 2D bar codes that was created by Denso Wave, within 1994 [12]. The symbol things in 2D bar codes include light and dark squares. The 2D specifications set the encoding of the information, the dimensions of quiet zones before and also after the barcode, the finder or maybe place detection patterns, as well as blunder detection and correction of information [12]. Barcodes present an inexpensive and simple way to encode textual info about objects or items in a type which machines are able to read, retrieve, validate, and procedure [12]. The QR code has the increased capacity that will keep 7,089 numeric, 4,296 alphanumeric, and also 2,953 binary characters [13]. Now, QR Codes have forty designs, which range from one to forty, so the scale of each edition is different. The size of QR code is dependent on the vertical and horizontal sizes of the QR version employed [14]. It can be checked out with smartphones equipped by using a digital camera. A software program client placed on the smartphone controls the camera to browse and understand the coded info, letting mobile users to connect to the net with a point and click of the phones of theirs, therefore making mobile surfing easier. It's clear by reasonably equipped mobile cell phones with cameras and also QR

*Corresponding Author

scanners, info like Url, SMS, contact info and plain text could be embedded into the two dimensional matrix [13]. Data can be encrypted inside a QR code to offer the confidentiality of info lodged in the code [15]. The barcode and QR code are presented in Fig. 1, respectively.

One of the more trendy implementations is Google authenticator that is working with QR codes. The shuffle of the TOTP approach better wear by screening the key as being a QR code, as revealed in the majority movable applications are able to read. This is easier and acceptable to utilize in looking at the mechanical input of the same secret. After the TOTP authenticator is enabled, owners are going to be ready to allow MFA individually within their user profile that adds a layer of protection and postulate an added authentication code from a dependable device. Fig. 2 displays the TOTP based QR generation procedure flow diagram [16].

Several authentications methods have been developed to ensure the security of electronic transactions. Until now, there are many methods used for authentication in electronic payment. Onetime passwords (OTP) are produced on demand by Internet centralized party and delivered to the customer via a correspondence channel in which a registered getting device is assumed to have the client's possession. Probably the most prominent illustrations will be the SMS OTPs given by banking apps [17]. The majority of the OTP authentication methods are network reliant. The issue is that network-dependent devices provide a secure network connectivity between the device and the authentication server. For instance, SMS based program is going to need to transmit onetime password via an SMS within the user device. As in deep SMS primarily based two-step verification methods, the server will send out an SMS on the user's device, the person might have to purchase the price of SMS. The issue with SMS based OTP is it is just and the SMS network the cell phone is subscribed to? Recent studies show that OTPs over SMS are causing different causes and issues, which lead precious time, delay in transaction [18]. SMS OTP might also have financial problems in case the carrier charges the subscriber for having SMS communications. In this paper, we approach the Time-Based OTP Authentication Algorithm for electrical payment. There will be no spoofing or perhaps tempering of the transmitted information in between. In this manner, only legitimate user will gain a chance to access the account. The entire program will operate with absolutely no system expecting the registration stage. The proposed system maintains zero SMS policy with no additional charges for SMS. It will encrypt and secure the information inside the system from any misuse.

We structure this paper into seven sections. Section 1 discussed above together with introducing the multi-factor authentication techniques. Section 2 offers a brief knowledge of the literature review with OTP techniques. Section 3 points out the part of the proposed method architecture. Section 4 presents the system architecture of the proposed system and result and implementation are presented in Section 5. Section 6 discussion on performance key factor and Section 7 concludes the paper.

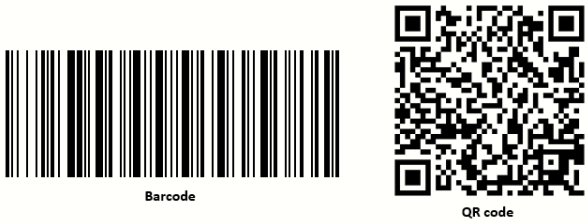


Fig. 1. Barcode and QR Code.

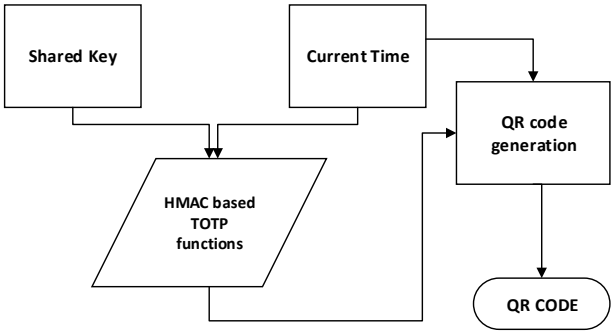


Fig. 2. TOTP based QR Generation Procedure Flow Diagram.

II. Literature Review

To preserve details on the net as protected as servers and possible, many clients implement different kinds of cryptographic methods to encrypt sensitive details and authenticate individuals at the opposite end of the connection [19-20]. Today that much more confidential information is stored virtual [21], it is supreme that community security oak updated with modern threats [22]. The bulk of sites in use today use the conventional verification pattern of supplying a password and a username more than a secure interconnection. The user name is used-to find what online account a client would like to access, even though the password is accustomed to confirming the identity of the customer. However, it seems secure in concept several passwords even now wind up being affected [23-24]. This is because of two things: vulnerable passwords and also quicker password cracking hardware [25-26]. In order to stop these attacks, two-factor authentication contemplated solving securing online transactions and also identifying the authentic individual and logging them right into a method or even the program.

The most used method of boosting the protection of an account is asking for additional info out of the computer user. Rather than asking only one piece of info out of the person, the server can ask for additional info, making it a lot harder for an assailant to bogus the identity of the person. With the hand of the fantasy, they have approached the OTP primarily based authentication [27-28], do the related work approach by [29]. A onetime password (OTP) method depends on the capability for just a unit to make a onetime code that will be delivered towards the server for verification. If the code is discovered to be accurate, subsequently that consumed is provided a chance to access the account. A onetime Password (OTP) is the one of the important part of the mobile networking [30]. OTP is a password or maybe code that is effective just for one login session or maybe transaction holding a computer or maybe

some electronic device. OTPs had been released to stay away from the flaws, which are connected with fixed passwords. Even though they are legitimate for a little time and they also instantly expire after the specified time span. A technological mechanism to reduce the risk of an unauthorized person getting to access the account. The most important advantage of OTP is in contrast to a static password. OTP, security technique shield for the various password-based attacks, specifically password sniffing and reply attack [31]. TOTP is one of the principal requirements for the onetime password. In generally, TOTP, the token creates a numeric code, typically six or maybe eight digits [25]. TOTP makes use of time in increments known as time action, and that is typically thirty or maybe sixty seconds. What this means is that every OTP is legitimate for the duration of the precious time action. The TOTP is regarded as a much more safe Onetime Password remedy. A high-level diagram of TOTP enrolment process are shown in Fig. 3 [23].

A Time-based authentication of multi-factor tokens improved cryptocurrency security approach by [32]. Tahar et al. (2019), in their research, they developed the protection and enhancement algorithm for MFA Crypto-monetary (CR) to set up an additional safeguard layer when looking for the target through the onetime password (TOTP) technology in time. The user first requires a username and password for logging into every 2FA-enabled entity; as a second factor, the user will then create a TOTP virtually through the token. A similar concept based TOTP based challenge response protocol for e-commerce approach in [33]. Aina et al. (2018) on their paper, they approached Scan2Pass payment for banking system. The system is depended both server side and client side. After the registration in server side, the user needs to input their username and password and generate a QR code. In client side, the user has to open; his mobile application to screening the QR code after input the user authentication details. Do the similar work proposed by [34]. Abhishek et al. (2020), in their article they proposed TOTP Based Authentication Using QR Code for payments. The QR code is read, and the system tracks the TOTP on the server side. The consumer is permitted to join whether this TOTP matches in the QR code. Moreover, Chowdhury et al. [35] suggested the usage of OTP and QR code for payment transfers in the online banking system.

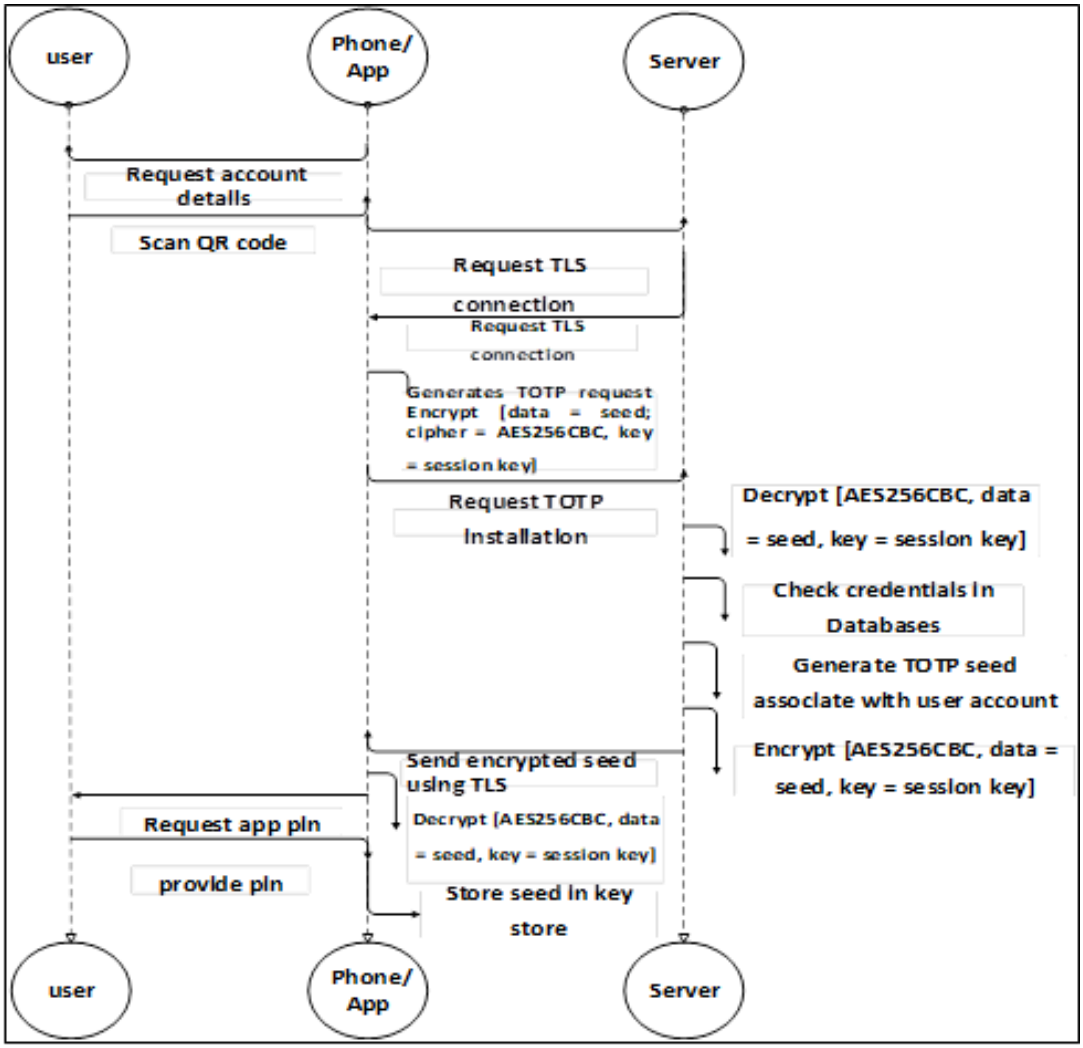


Fig. 3. A High-Level Diagram of TOTP Enrolment Process.

A QR code along with a secret key shared between the client and the server is generated in this system and is used to generate an OTP, which is integrated into the QR code. For the transaction to be completed, this QR code is then checked to verify the OTP. A new OTP must then be created for every session to provide additional protection. In addition, a TOTP based two-factor authentication using smart phones as software tokens proposed in [36]. The proposed system would use mobile phones to create software tokens that will be used to authenticate an Internet Banking application, using specific onetime passwords (OTPs). The user ID, IMEI phone number, timestamp and PIN of the server compute the mutual secret of the TOTP algorithm on their device. After the analysis of the previous study, it can be found that most of the study proposed username and password based authentication system along QR code. Furthermore, the can be improved using biometric features for client side authentication. Authentication techniques, which hinge on much more when compared to one component, typically are tougher to compromise as opposed to just one element system. There is a strong need to produce the strategy powerful and efficient a multi-factor authentication component is necessary to boost protection for electronic transactions. Table I lists the existing systems and their proposed properties.

TABLE I. RELATED METHODS WITH THEIR PROPERTIES

Author	Method	Finding	Drawback
Tahar et al. [32]	Password /TOTP	This study proposed a framework for the security enhancement of the Cryptocurrency using time based token. The user first needs username and password, then user the TOTP token for authentication.	In password-based authentication, many security issues exist. Intruders may try different methods to steal passwords using password-based attacks.
Aina et al. [33]	Pin/TOTP	This study has proposed a challenge–response protocol for enhanced e-commerce security using time based token. The proposed Scan2Pass method used pin for authentication.	In pin -based authentication, various security issues exist. Intruders may try different methods to steal passwords using password-based attacks.
Abhishek et al.[34]	Password /TOTP	This study has proposed TOTP Based Authentication Using QR Code for Gateway Entry System. The proposed technique used username and password for TOTP authentication.	In password-based authentication, many security issues exist. Intruders may try different methods to steal passwords using password-based attacks.
Choudhary et al. [35]	Password /OTP	This proposed technique used Mobile OTP with the combination of QR-code, which is a variant of the 2D barcode. The proposed method used username and password for OTP verification.	Security challenges with password-based authentication are many. Intruders may use password-based attacks to attempt various ways of stealing passwords.

III. PROPOSED SYSTEM

In the proposed method, we have utilized TOTP as a starting algorithm to produce needed onetime passwords. TOTP is dependent on HTOP; However, HTOP is used counter whereas TOTP is a time-based algorithm. TOTP is going to generate an innovative worth after a determined period. This particular occasion is known as the time step. TOTP supports HMAC-SHA2 and HMAC-SHA1 hash functions [37]. The proposed system has two phases, namely: registration stage, an authentication phase. A comprehensive explanation of each phase is provided below. Before making use of this service, the user should register the information of theirs during a procedure known as the registration phase. Verification of that information may just be achieved by a procedure known as an authentication phase. Each of the suggested materials and strategies are completed in the system during both registration process as well as the login procedure, their process flow is reviewed in this area. In Table II, we provided the symbol used in the proposed technique.

$$ENTOTP = EN\ (PKIDi\ (TOTP)) \quad (1)$$

$$ENTOTP = QRDEC\ (QR\ (ENTOTP)) \quad (2)$$

$$TOTP = DEC\ (ENTOTP) \quad (3)$$

A. Registration Phase

After the registration is done, the client app creates an eight digit onetime password (OTP) that may be utilized for the authentication aim. The registration process of the proposed system can be seen in Fig. 4. However the registration process of the proposed system as working as follows.

Step 1: The user input his credential information *IDi* on the server.

Step 2: The server determines the client's info and recovers the client's public key *PKIDi*

Step 3: the server then choices an arbitrary string TOTP, have a period slot, and encrypts it together with the public element to get (1)

Step 4: The server generates the *QR* code in the payment side.

Step 5: The client decodes the QR code with (2)

Step 6: The arbitrary string is encrypted together with the client's public key *PKIDi*, the client is able to read the *T*OTP string just over the device of user by (3) and type in the *T*OTP within the terminal with an actual keyboard.

Step 7: Registration Successful.

TABLE II. LIST OF THE SYMBOLS USED

Notation	Description
IDi	Client details identification
PKIDi	Client's public key
EN	Encryption string
*T*OTP	Times based one-time password
QR	Quick Response
DEC	Decryption

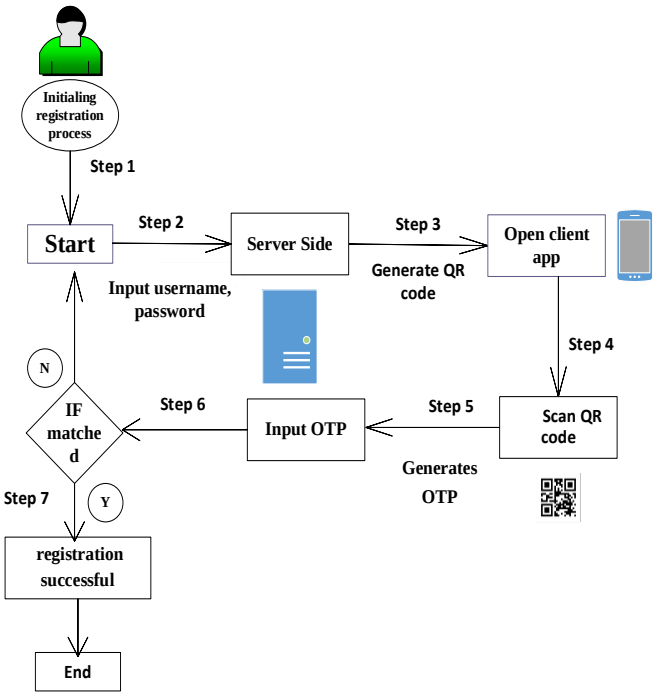


Fig. 4. Registration Phase.

B. Authentication Phase

The authentication service has to authenticate the client whenever the client wants to access the system. The authorization service checks server data and database identification Identities. The value submitted by the client would be compared to the current value of the server. When the values are both identical, the authentication is successful; the new value will be used to change the old value for the server. Otherwise, the authentication of the client will fail. Fig. 5 illustrates the method of authentication of the proposed system. The details authentication steps of the proposed system are mentioned below:

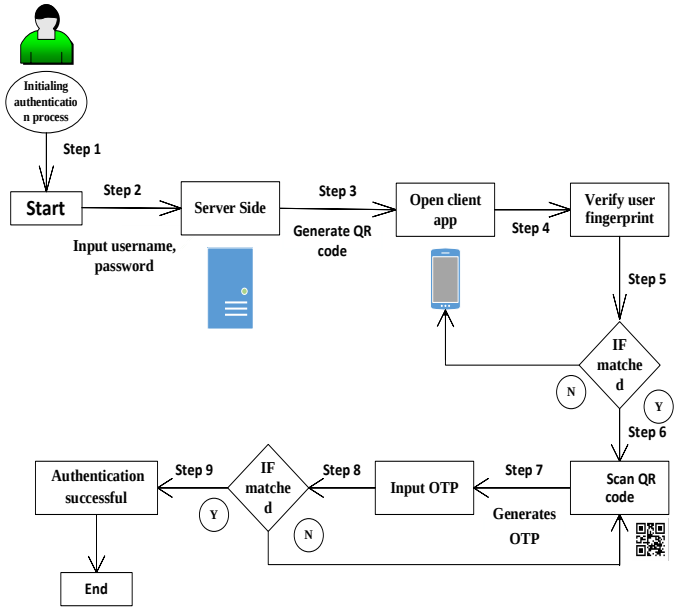


Fig. 5. Authentication Phase.

Step 1: The user input his credential information on the server.

Step 2: The server determines the client's info and regains the client's public key

Step 3: The server generates the *QR* code in the payment side.

Step 4: In user side, the user will open the application.

Step 5: The client input his fingerprint to verify

Step 6: Once the app verify the registered user, then the app ready for decode the QR code.

Step 7: the user will get TOTP number after decode the QR code.

Step 8: the user will input the TOTP number in the server side, if matched,

Step 9: Authentication successful.

IV. System Architecture

In this paper, we proposed TOTP based on authentication for enhanced electronic payments authentication security. The system design includes various entities, like a prospect, a smartphone, a user's PC and a server. The user is an individual with little to no knowledge of cryptographic codes, such as passwords and complicated mathematical equations. The terminal of a user is a computer of a user that is used to connect to a server for money transfers [38]. The user has a smartphone that stores the public key certification of the digital certificate or the server furnished with a camera. The server is the method entity belonging to the monetary institution that interrelates with the user by carrying out all the back end operations. In deep agreement with the present moment, TOTP uses a secret shared between client and server to produce a onetime use code [39]. Through executing the disgust secret through the algorithm, the client experiences the code with the server being able, during the whole algorithm, to confirm the published chip with a similar secret. The cipher is equally relevant for an imbued amount of time, usually thirty seconds [32]. The flow looks like firstly operator logs directly into an application program with username as well as the password, now view a text field asking to type in the newest launch and code TOTP client on their cell phone. Fig. 6 displays the proposed framework architecture of the proposed system.

The user gets a TOTP token by scanning the QR code. In the first phase, users open the Internet browser for login their account details getting a username password together with TOTP. Within the next stage, it provides an authentication need on the identity authentication server. In the last stage, verification on the request is used by confirming the allowed individual through identity authentication server. The request may be accepted in the last stage and maybe denied. The onetime password is made on the subject of the server using seed exchange, after which provided via a Transport Layer Security (TLS) tunnel about the client mobile program. The client will solely be authenticated whether it suits the password on the server on the server part. It is moreover secure than the SMS solution, since the transmission of the cipher is not intermediate. The function is the algorithm. To stay behind safe, mutual confidentiality should be reserved for this process.

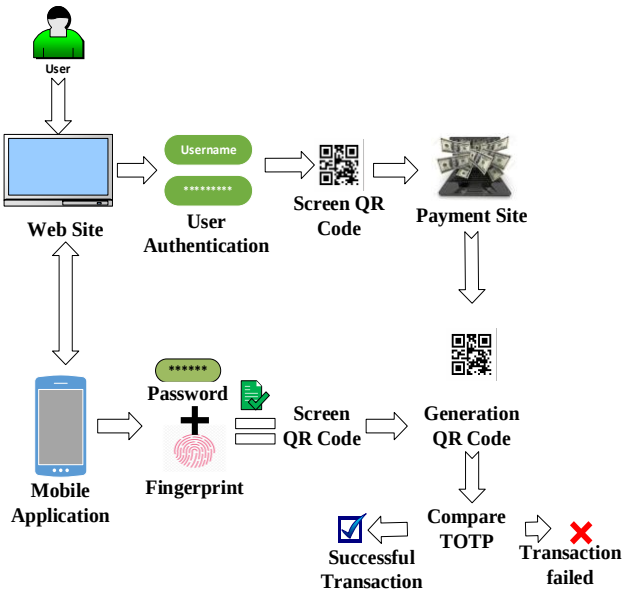


Fig. 6. Proposed System Architecture.

V. Design Consideration

The suggested solution works with a smartphone on the person's side. The smartphone plays a significant part within connecting the breach between the server and the user. In order to offer secure user authentication device, which works mutual authentication in between entities, thus, the proposed method uses the TOTP algorithm of RFC 6234 to compute the OTP required authenticating the user and finishing the login process [27]. The android application syndicates three components: the shared secret, timestamp and server challenge [33], in the mobile to make a token of 8 or more 8 digits so long as it applies the TOTP algorithm. Random details are essential, and they are utilized by the 8 bytes utilizes tokens and the shared secret. The system is depended both server side and client side. Some parameters are needed for the establishment of a TOTP authentication. The following steps are descrying how this framework works:

- For the TOTP generation, users and servers will know or be able to measure the current UNIX period.
- A secret key must be shared between user and server. The hidden key may be used as a pre-existing key between the parties. On the other side, the secret key may be produced by means of a main agreement protocol during agreement between the parties. This is a secure communication.
- The HMAC-Based One Time Password (OTP) will be the main component for the algorithm.
- The same time value is required for both the user and the server.
- For each user there must be a single specific private key.
- The key must be generated randomly or by key derivative algorithms and the keys should be protected from unauthorized access.
- In order to login, first-time users must register. At the registration stage, the user from the server will provide the QR code for authentication with the username and password.
- The user application runs on their device and needs user registered fingerprints on their phone to authenticate it.
- The registered device should only be used for a transaction, so each time a valid fingerprint needs to be checked.
- In order to access the system, the user is compulsory to input the approved fingerprints.
- Once users successfully enter the username and password, then the user side application needs to open a QR scan request. Remember that the user must encrypt their application via fingerprint before logging on the services.
- Complete after registration. The QR code scan page will be sent to the user with the same hidden key once the login is done and stored in the database. This key will produce the TOTP encoded in the QR code. Therefore, the QR code is verified using a QR code reader. The TOTP is then compared with the server TOTP.
- The user is permitted to enter if both TOTP match, otherwise access is not permitted.

VI. Discussion and Analysis

In this paper, we use the Time Based onetime password authentication algorithm to secure an electronic payment. The TOTP method is generally utilized in applications, which have to limit time like mobile banking and applications transactions. This section summarizes the key functionality and discussing regarding the OTP authentication system their methods. In the earlier methods found there are already various stages in the authentication task, as there they have worn SMS OTP Authentication within the authentication phase. Right here we have used TOTP its combat with specific QR Code of user that could be a fruitful method for supplying great protection on the authentication procedure. Here we have compared the usability considerations of SMS OTP and TOTP. The comparison of existing methods with the proposed system outcome is shown in Table III, where the usability considerations are discussed in Table IV for both SMS OTP AND TOTP.

The important paradigm of SMS OTP that is the Mobile Transaction Authorization Number, that's put on to authorize transactions of the person. In this particular mechanism, the OTP is delivered as a text message on the user's mobile device. Nevertheless, the protection of SMS OTP depends on the confidentiality of SMS, which is trusted by the security of movable networks [40]. While authenticator Apps count during a shared secret, which both the server and the App have to store. This "seed" is mixed with the period to produce the multi-factor authentication code. In our method, the TOTP based onetime password authentication for secure electronic payment process aims to be raised by utilizing TLS connection

between server and client Apps. Because the seed is discussed making use of the secure link, therefore it is never, exposed.

User verification has become more and more important than ever for electronic payments. Various authentication stages were described in previous approaches, as they did with the knowledge-based methods in the authentication stage. The security mechanism for usernames and passwords that can easily be accessed through guessing and password based attacks [41-42]. There is also a possibility to develop user authentication methods for multi-factor implementations. This study suggested a user authentication framework focused on TOTP for electronic payments that are concrete with biometric features. In addition, the proposed study recommends the possibility of biometric fingerprints verification during user authentication. The fingerprint method appears to be one of the most secure means for authentication in the electronic payments world in order to reduce future security vulnerabilities [43-44].

However, the proposed system is free of cost. The program-offering site likewise should make use of this product to improve the protection of the program, charging no extra cost. Because user have no SMS, services associated with the device so there will be absolutely no cost of SMS to user and server. This method could be lodged in a broad range of applications to provide multi factor authentication.

TABLE III. COMPARING THE RESULT OF EXISITNG AND PROPOSED METHODS

Attribute	**Tahar et al.** [32]	**Aina et al.** [33]	**Abhishek et al.**[34]	**Choudhary et al.** [35]	**Our method**
Authentication technique	TOTP	TOTP	TOTP	TOTP	TOTP
Method	Password	Pin	Password	Password	Fingerprint
Methods type	knowledge	knowledge	knowledge	knowledge	Biometric
Authentication type	Two-Factor	Two-Factor	Two-Factor	Two-Factor	Multi-Factor
Password based Attack	Yes	Yes	Yes	Yes	No

TABLE IV. COMPARING THE USABILITY CONSIDERATIONS OF SMS OTP AND TOTP METHOD

System	**SMS OTP**	**Our proposed method**
Token	Cellphone	Smartphone
Client App	No	Yes
Enrollment	SMS	QR Code
Derivation	SMS	Offline
Resettable	N/A	Yes
Cost	Not Free	Free
Service Access	Restricted	Worldwide
Service Provider	Cellular Network	Not Required
Secure Seed	Fixed	Dynamic
Availability	No	Yes

VII. CONCLUSION

Strengthened multi-factor authentication guarantees the protection of personal data for internet companies and protects them from collapsing or losing money. With Time-based multi-factor authentication algorithm, we improved protection of electronic payments. Our proposed methods uses mechanisms of TOTP, where it facilitates the user device authentication creating the onetime codes. Enabled MFA and worked with the TOTP method to include an additional level of protection for an electronic payment program. We presented our proposed method is building an additional biometric authentication layer that is going to provide additional is safe against famous attacks such as spoofed, MITMF and tempering. The real information of the user is saved anomalously in database. In addition, the algorithm is used to operate an identical secret via the algorithm using a shared secret key between the client and the server. Our system has the benefit to authenticate the only legitimate user will acquire a chance to use the account where the system is free of cost. Our suggested solution has shown that security efficiency for authentication and authorization has been improved significantly compared to the existing method. Finally, the effort could be put on using modern environments such as cloud computing, banking systems, e-commerce, and mobile devices. In the future, we will apply in actual time as a potential task. In addition, we have focused on incorporating other protection elements into the approaches suggested.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful feedback. We are also grateful to Prof. Dr. Zarina Shukur for helping to perform this research. This research was funded by a research grant code from Ya-Tas Ismail - University Kebangsan Malaysia EP-2018-012.

REFERENCES

[1] V. Khattri and D. K. Singh, “Implementation of an Additional Factor for Secure Authentication in Online Transactions,” J. Organ. Comput. Electron. Commer., vol. 29, no. 4, pp. 258–273, 2019.

[2] H. Venugopal and N. Viswanath, “A robust and secure authentication mechanism in online banking,” Proc. 2016 Online Int. Conf. Green Eng. Technol. IC-GET 2016, pp. 0–2, 2016.

[3] K. Skračić, P. Pale, and Z. Kostanjčar, “Authentication approach using one-time challenge generation based on user behavior patterns captured in transactional data sets,” Comput. Secur., vol. 67, pp. 107–121, 2017.

[4] M. A. J. Kartini Mohamed, Fatemah Sidi, “Strengthening User Authentication for Better protection of mobile application system,” J. Theor. Appl. Inf. Technol., vol. 85, no. 3, 2016.

[5] O. Ogbanufe and D. J. Kim, “Comparing fingerprint-based biometrics authentication versus traditional authentication methods for e-payment,” Decis. Support Syst., vol. 106, pp. 1–14, 2018.

[6] N. A. Karim and Z. Shukur, “Review of user authentication methods in online examination,” Asian J. Inf. Technol., vol. 14, no. 5, pp. 166–175, 2015.

[7] A. Hassan, Z. Shukur, and M. K. and A. S. A.-K. Hasan, “A Review on Electronic Payments Security,” Symmetry (Basel)., vol. 12, no. 8, p. 24, 2020.

[8] W. Meng, D. S. Wong, S. Furnell, and J. Zhou, “Surveying the development of biometric user authentication on mobile phones,” IEEE Commun. Surv. Tutorials, vol. 17, no. 3, pp. 1268–1293, 2015.

[9] F. Sinigaglia, R. Carbone, G. Costa, and N. Zannone, “A Survey on Multi-Factor Authentication for Online Banking in the Wild,” Comput. Secur., no. February, p. 101745, 2020.

[10] O. S. Okpara and G. Bekaroo, "Cam-Wallet: Fingerprint-based authentication in M-wallets using embedded cameras," IEEE Int. Conf. Environ. Electr. Eng., 2017.

[11] Q. Code, "QR Code Standardization | QRcode.com | DENSO WAVE." [Online]. Available: https://www.qrcode.com/en/about/standards.html. [Accessed: 31-Mar-2020].

[12] M. H. Sherif, Protocols for Electronic Commerce, vol. 53, no. 9. 2016.

[13] T. S. & R. A. Sharvil Shetty, "QR-Code based Digital Wallet," Int. J. Adv. Res. Comput. Sci., vol. 5, no. 7, pp. 105–110, 2014.

[14] A. Alhothaily, A. Alrawais, T. Song, B. Lin, and X. Cheng, "Quickcash: Secure transfer payment systems," Sensors (Switzerland), vol. 17, no. 6, pp. 1–20, 2017.

[15] A. A. Lezhebokov, Y. A. Kravchenko, and V. V. Bova, "Support system for QR-code-based educational processes," 8th IEEE Int. Conf. Appl. Inf. Commun. Technol. AICT 2014 - Conf. Proc., pp. 4–7, 2014.

[16] J. Physical, "Physical presence verification using TOTP and QR codes," Int. Conf. ICT Syst. Secur. Priv. Prot. - IFIP SEC 2019, Lisbon (Portugal), 2019, 2019.

[17] E. Esiner, S. H. Hanley, and A. Datta, "DMZtore: A dispersed Data Storage System with Decentralized Multi-factor Access Control (Demo)," Proc. - Int. Conf. Distrib. Comput. Syst., vol. 2016–Augus, pp. 757–758, 2016.

[18] S. P. Dhanashri Ghosalkar, "OTP over SMS: Time Delay Issues and Causes," pp. 1–7, 2019.

[19] L. Xuanzhi and K. Ahmad, "Factors Affecting Customers Satisfaction on System Quality for E-Commerce," Proc. Int. Conf. Electr. Eng. Informatics, vol. 2019–July, no. July, pp. 360–364, 2019.

[20] S. S. Alam, M. H. Ali, N. A. Omar, and W. M. H. W. Hussain, "Customer satisfaction in online shopping in growing markets: An empirical study," Int. J. Asian Bus. Inf. Manag., vol. 11, no. 1, pp. 78–91, 2020.

[21] M. A. Hassan and Z. Shukur, "Review of Digital Wallet Requirements," 2019 Int. Conf. Cybersecurity, ICoCSec 2019, pp. 43–48, 2019.

[22] P. Aigbe and J. Akpojaro, "Analysis of Security Issues in Electronic Payment Systems," Int. J. Comput. Appl., vol. 108, no. 10, pp. 10–14, 2014.

[23] A. O. Alsayed and A. L. Bilgrami, "E-Banking Security: Internet Hacking, , Analysis and Prevention of Fraudulent Activities," Int. J. Emerg. Technol. Adv. Eng., vol. 7, no. 1, pp. 109–115, 2017.

[24] J. Gualdoni, A. Kurtz, I. Myzyri, M. Wheeler, and S. Rizvi, "Secure Online Transaction Algorithm: Securing Online Transaction Using Two-Factor Authentication," Procedia Comput. Sci., vol. 114, pp. 93–99, 2017.

[25] M. L. T. Uymatiao and W. E. S. Yu, "Time-based OTP authentication via secure tunnel (TOAST): A mobile TOTP scheme using TLS seed exchange and encrypted offline keystore," ICIST 2014 - Proc. 2014 4th IEEE Int. Conf. Inf. Sci. Technol., pp. 225–229, 2014.

[26] A. Hassan, Z. Shukur, and M. K. Hasan, "An Efficient Secure Electronic Payment System for E-Commerce," computers, vol. 9, no. 3, p. 13, 2020.

[27] M. Harish, R. Karthick, R. M. Rajan, and V. Vetriselvi, A New Approach to Securing Online Transactions—The Smart Wallet, vol. 500, no. January. Springer Singapore, 2019.

[28] K. Vengatesan, A. Kumar, and M. Parthibhan, Advanced Access Control Mechanism for Cloud Based E-wallet, vol. 31, no. August 2016. Springer International Publishing, 2020.

[29] R. Mohan and N. Partheeban, "Secure Multimodal Mobile Authentication Using One Time Password," Int. J. Recent Technol. Eng., vol. 1, no. 1, pp. 131–136, 2014.

[30] S. Islam, A. H. A. Hashim, M. H. Habaebi, and M. K. Hasan, "Design and Implementation of a Multihoming-Based Scheme to Support Mobility Management in NEMO," Wirel. Pers. Commun., vol. 95, no. 2, pp. 457–473, 2017.

[31] R. Idayathulla, "Enhanced adaptive security system for SMS – based One Time Password," vol. 5, no. 4, pp. 538–541, 2019.

[32] K. A. Taher, T. Nahar, and S. A. Hossain, "Enhanced cryptocurrency security by time-based token multi-factor authentication algorithm," 1st Int. Conf. Robot. Electr. Signal Process. Tech. ICREST 2019, pp. 308–312, 2019.

[33] F. Aina, S. Yousef, and O. Osanaiye, Design and Implementation of Challenge Response Protocol for Enhanced e-Commerce Security, vol. 3. Springer International Publishing, 2018.

[34] Abhishek Arvind, Pradyumna Mahajan, and Rishikesh Chalke, "TOTP Based Authentication Using QR Code For Gateway Entry System," Int. J. Eng. Comput. Sci., vol. 9, no. 05, pp. 25023–25028, 2020.

[35] A. Choudhary, S. Rajak, A. Shinde, S. Warkhade, and P. G. F.S., "Online Banking System using Mobile-OTP with QR-code," Ijarcce, vol. 6, no. 4, pp. 657–661, 2017.

[36] C. A. Soare, "Internet Banking Two-Factor Authentication using Smartphones," J. Mobile, Embed. Distrib. Syst., vol. 4, no. 1, pp. 12–18, 2012.

[37] C. Sudar, S. K. Arjun, and L. R. Deepthi, "Time-based one-time password for Wi-Fi authentication and security," 2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017, vol. 2017–Janua, pp. 1212–1215, 2017.

[38] R. Divya and S. Muthukumarasamy, "An impervious QR-based visual authentication protocols to prevent black-bag cryptanalysis," Proc. 2015 IEEE 9th Int. Conf. Intell. Syst. Control. ISCO 2015, 2015.

[39] V. Shukla, A. Chaturvedi, and N. Srivastava, "A new one time password mechanism for client-server applications," J. Discret. Math. Sci. Cryptogr., vol. 22, no. 8, pp. 1393–1406, 2019.

[40] R. M. Ibrahim, "A Review on Online-Banking Security Models , Successes , and Failures," Int. Conf. Electr. Electron. Comput. Commun. Mech. Comput. (EECCMC)" IEEE EECCMC, no. February, 2018.

[41] Mohammed and Yassin, "Efficient and Flexible Multi-Factor Authentication Protocol Based on Fuzzy Extractor of Administrator's Fingerprint and Smart Mobile Device," Cryptography, vol. 3, no. 3, p. 24, 2019.

[42] S. F. Tan and A. Samsudin, "Enhanced Security of Internet Banking Authentication with EXtended Honey Encryption (XHE) Scheme," pp. 201–216, 2018.

[43] A. Gupta, D. Kaushik, and S. Gupta, "Integration of Biometric Security System to Improve the Protection of Digital Wallet," SSRN Electron. J., no. Icicc, pp. 1–6, 2020.

[44] A. S. Robert Hunt, Jeremy Kalas, Patrick Lowe, "Biometric security," Biometrics Concepts, Methodol. Tools, Appl., pp. 1399–1418, 2016

An Efficient Digital Space Vector PWM Module for 3-Φ Voltage Source Inverter (VSI) on FPGA

Shalini Vashishtha[1]
Research Scholar
Dept. of ECE, SJBIT Research Center
Bangalore, India

Rekha K.R[2]
Professor
Dept. of ECE, SJBIT
Bangalore, India

***Abstract*—The realization of digital control circuitry based PWM strategies provides many advantages. It includes better prototyping, higher switching frequency, simple hardware, and flexibility by overcoming the limitations of analog control strategies. In this article, The Digital space vector-based Pulse width Modulation ((DSV-PWM) is designed. The DSV-PWM Module includes, mainly, Xdq reference frame, Sector generation, Square root, switching time generation, Carry-save adder (CSA), and PWM Generation module. These modules are designed using simple logical operations, and combinational circuits to improve the DSV-PWM performance. The DSV-PWM Module is synthesized and implemented on a cost-effective Artix-7 FPGA device. The present work utilizes a < 1% chip area, operates at 597.83 MHz of maximum frequency, and utilizes 110mW of total power on FPGA Device. The DSV-PWM module is also compared with the existing SV-PWM approach with better improvement in hardware constraints like chip area, operating frequency, and dynamic power (mW).**

***Keywords*—Digital space vector PWM; 3-phase voltage source inverter; sector generation module; switching time generation; FPGA; Verilog-HDL; Xilinx**

I. INTRODUCTION

The development of AC motor drives with present technology becomes major highlight for delivering a suitable amount of energy to the motor. This is achieved with the drastic improvements and advancement in the microprocessor, power electronics technologies and switching power converters. The amount of energy is computed by Pulse width Modulated (PWM) signals. Most of the existing analog control circuitry provides excellent dynamic response, but failed to maintain circuit complexity, circuit modification, and has few functions. The present AC motor drives adopt the processor-based or advance technology like DSP, ASIC, or FPGA based digital control strategy. The digital control strategy overcomes the existing analog control strategy limitations with more significant improvement in circuit development. The commonly used PWM approaches include the Space Vector (SV)-PWM method, Sinusoidal-PWM method, and Hysteresis-PWM method [1-2].

The 3Φ--Voltage Source Inverter (VSI) provides the Variable supply voltage to AC motor drives and is used majorly in modern industrial applications. The PWM methods are used to modulate the VSI output voltage using different controlling strategies. The SV-PWM based VSI uses many topologies, which includes, High Power 2-level VSI and Multi-level Inverter. The Multi-level inverter further classified has advanced bus Clamped-VSI, flying capacitor, or Diode Clamped-VSI and Neutral point Clamped VSI [3].

The FPGA and DSP based Controlling strategies assist different digital AC-Motor drives to control the motor. The DSP based approaches provide great features like simple circuitry, flexibility to use for many applications and software-based control strategies. But failed to maintain the bandwidth performance because of the high sampling rate used in the current control loop and PWM generation. The current control loop and PWM generation use most of the computational resources in the controller and only limited time is available to control other specified functions in the Drive. To fulfill the limitations of DSP based Controlling strategies, The FPGA /ASIC based PWM controlling strategies are used for AC motor drives. There is rapid development in FPGA, which offers faster circuit response continuously and supports parallel processing rather than sequential execution and more excellent speed control capability of 3Φ-induction motors [4-5]. The FPGA is a reprogrammable device that contains logical interconnection and logical Blocks. The logical blocks are designed with the help of logical operations, combinational, and sequential circuits. The FPGA is user friendly, avoids the high NRE cost, inflexibilities, and development cycles than the conventional ASIC approach [6-9].

In this article, an efficient Digital space vector (DSV) Pulse width Modulation (PWM) is designed and implemented on low-cost Artix-7 FPGA. It provides high performance and less resource utilization on hardware and is suitable for real-time power electronics applications. The proposed design uses simple logical operations and combinational circuits to design the DSV-PWM, which assures low power consumption, less execution time, and high reduction in chip area in FPGA and also reduces the complexity in 3 Φ-VSI.

Section II describes the existing approaches of different SV-PWM techniques from different application viewpoints and also analyzes limitations. The proposed DSV-PWM module is described in detail with the basic principle and hardware architecture in section III. Section IV discusses the simulation results and performance analysis of the DSV-PWM method and also compares with the existing SV-PWM technique with constraints improvements. Finally, the section V concludes the Overall work with improvements and suggest the future scope.

II. Related Works

This section discusses the existing approaches of different space vector (SV)-PWM techniques for different application usage. Janik et al. [10] present the SV-PWM technique without the usage of trigonometric operation to improve the hardware performance in real-time scenarios. The module investigates the multilevel converter (MLC) using independent voltage modulation. The MLC Module supports both DSP and FPGA operations on a single MLC interface unit with other supporting peripheral devices. Holtz et al. [11] present the SV-Modulator for Higher switching frequency control operation with the help of three-level Silicon-Carbide (SiC) Inverter. The design uses a two-level SV approach for switching time generation, and the Modulation module is designed based on logic decisions. The 3-Level Inverter is designed based on Circuit topologies, High-frequency PWM, Neutral point Potential (NPP) control, and 3-level Modulator. Kassas et al. [12] discuss the Look-Up Table (LUT) based SV-PWM, which is designed and implemented using simple 8-bit Microcontroller and also modeled using Simulink. The LUT used to improve the time assignment computation in 3-Φ VSI. The LUT based SV-PWM has three main modules includes PWM Timer, High, and low-priority interrupt modules for total harmonic distortion (THD) and error Calculation.

Celik et al. [13] present the real-time SV-PWM signal generation using FPGA Device. The PWM signals used to control the VSI are achieved on FPGA prototype circuit. The SV-PWM design results are verified using a time-domain simulation using Matlab Simulink. Liang et al. [14] present the SV-PWM based control algorithm on FPGA and also verify the FPGA design with a Software-based SV-PWM approach using Matlab Simulink. The SV-PWM Control algorithm includes the sector determination for any space vector, time-domain voltage space vector function generation, switching time, and conduction module. Lotfi et al. [15] discuss the SV-PWM module for voltage inverter-based AC machines. The module includes Voltage reference calculation, Sector determination, Switching time calculation, and also a series of PWM Pulses. The design uses more LUTs, consumes more chip areas and not significant for real-time applications.

The SV-PWM computations are designed for multi-level inverter by Salem et al. [16] and also for open end-winding induction machine (OEWIM) using Dual 3-level T-type converter. The operation of a 3-level dual T-Type converter is designed using vector diagrams and switching states, Sector and region identification, time interval calculation, and switching pattern modules. The design also analyzes the Voltage and current THD against different switching frequencies. Pu et al. [17] describe the Random SV-PWM technique for 3Φ- VSI on FPGA, which includes a randomization algorithm for faster and flexible realization hardware. The Random –SV-PWM has a waveform generation module, Pseudo random signal generation module, and PWM generation module. The random SV-PWM results are useful in terms of standard Line Voltage, wave filtering of Line Voltage, FFT of Line Voltage than conventional PWM approaches. Khlavi et al. [18] present the reconfigurable PWM Generator on FPGA for power electronics appliances. The module is implemented on both ASIC and FPGA, which can be configured easily with different PWM strategies. Garcia et al. [19] discuss the SV-PWM control module for 2 levels 3Φ-inverter on Matlab GUI and FPGA. The design also supports a simplified education platform for teaching SV-based PWM techniques. Suma et al. [20-21] present the FPGA controller for an Induction Motor drive and also SV-PWM based design for 3-level Inverter on non-volatile FPGA. Chinmaya et al. [22] present the analyses of different SV-PWM methods like Conventional SV-PWM, Vector space decomposition-based, vector classification based, and Common mode voltage injection-based SV-PWM approaches for dual 3Φ induction motor drive.

The proposed work overcomes the conventional SV-PWM methods by considering the performance and accuracy of the FPGA system. The present work reduces overall system cost and chip area for 3Φ-VSI on FPGA, and also present work ensures the great flexibility in FPGA for usage in real-time Power electronics applications.

III. Proposed Work

In this section, the Digital space vector-based Pulse width Modulation (DSV-PWM) designed for 3Φ-Voltage source Inverter (VSI) is explained with basic principle and its hardware architectures.

A. Basic Principle of SV-PWM

The Digital SV-PWM provides the low-current –ripple and DC-Link Voltage with better resource utilization, Lower harmonic content, Wider linear modulation index range than the conventional-SV-PWM approaches. The SV-PWM is implemented using Voltage equations in the form of the abc reference frame. This frame is converted to the d-q reference frame, and it has a horizontal (d) and vertical (q) axis. It can be determined by the Reference voltage space vector (V_{ref}), which is rotating in a circular position that constitutes a sinusoidal waveform, and it is represented in Fig. 1.

The Reference Voltage space vector estimates the combination of 8- switching forms (V_0 to V_7) in the PWM Technique. The V_1 to V_6 active (Switching) vectors are divided into 6 sectors in a hexagonal plane, and each of the sectors is arranged in 60 degrees. The V_0 and V_7 are null vectors. The V_{ref} is calculated using two null vectors and any two actives (Switching) vectors, and it is represented in Fig. 2.

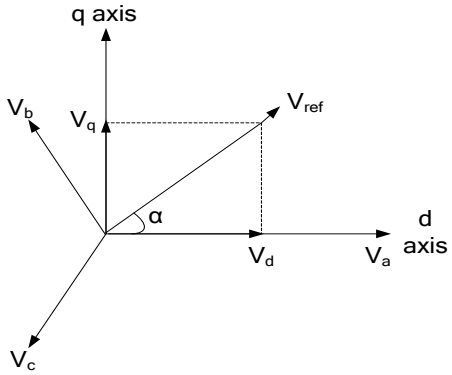


Fig. 1. Reference Voltage Space Vector.

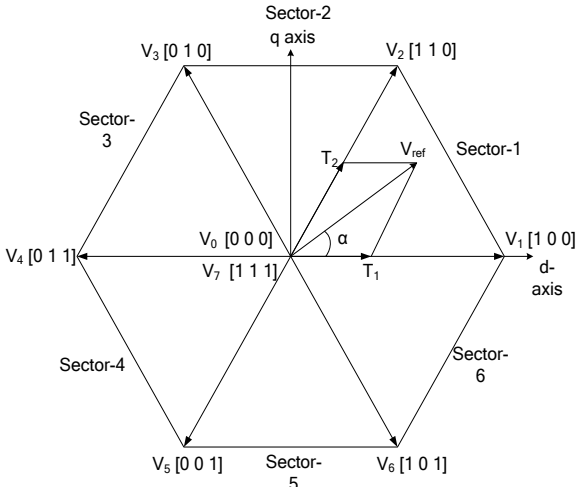


Fig. 2. Fundamental Sectors and Switching Vectors.

The three-phase voltage vector [V_a, V_b, V_c] and switching variable vector relationships are described as follows [9]:

$$\begin{bmatrix} V_d \\ V_q \end{bmatrix} = \frac{2}{3}\begin{bmatrix} 1 & -1/2 & -1/2 \\ 0 & \sqrt{3}/2 & -\sqrt{3}/2 \end{bmatrix}\begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix} \quad (1)$$

The reference voltage and angle are derived using eq (2) and eq (3) as follows:

$$\left|V_{ref}\right| = \sqrt{V_d{}^2 + V_q{}^2} \quad (2)$$

$$\alpha = \tan^{-1}\left(\frac{V_q}{V_d}\right) = \omega_s t = 2\pi f_s t \quad (3)$$

Where f_s is the fundamental frequency.

B. Proposed Hardware Architecture of DSV-PWM

The Hardware architecture of the DSV-PWM method is represented in Fig. 3. The DSV-PWM mainly consists of Xdq Module, Sector generation module (SGM), Switching Time Module (STM), Square root module, Carry save Adder (CSA) module, and PWM generation module. These modules are constructed using simple Logical and combinational circuits, which provide higher performance keeping the hardware constraints into consideration.

The Xdq Module is designed using equation (1), and it is simplified using intermediate vectors X_d and X_q by equation (4) and equation (5) respectively, as follows:

$$X_d = 2V_a - V_b - V_c \quad (4)$$

$$X_q = V_b - V_c \quad (5)$$

The X_{dq} Module simulation results are represented in Fig. 4. The X_{dq} Module uses simple multiplier and subtractors to generate the 8-bit X_d and X_q outputs by using the 8-bit V_a, V_b, and V_c inputs.

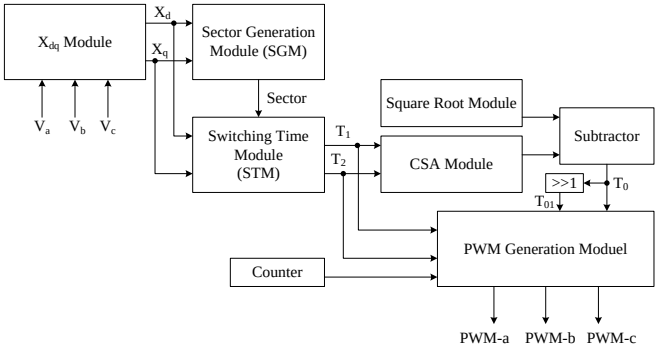


Fig. 3. Proposed Hardware Architecture of DSV-PWM Module.

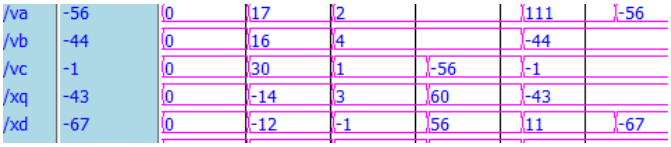


Fig. 4. Simulation Results of X_{dq} Module.

The switching variable vectors V_d and V_q are expressed by using the X_d and X_q in equation (6) as follows:

$$V_d = \frac{1}{3}\left[X_d\right] \text{ And } V_q = \frac{1}{\sqrt{3}}\left[X_q\right] \quad (6)$$

The sector generation module (SGM) is designed by using three conditions (a) sign of X_d, (b) Sign of X_q, and (c) $\left|X_d\right| = \left|X_q/2\right|$. The design uses simple combinational circuits to design these conditions. The one-bit right shift operation achieves the $X_q/2$. The sector generation is based on the following 6 rules, and it is tabulated in Table I.

The simulation results of SGM is represented in Fig. 5 using 8-bit X_d and X_q inputs along with 3-bit select line. The SGM generates the 3bit sector outputs. The SGM Hardware architecture, uses two signed Shifter modules, three multiplexers, comparators, encoder and Bit-wise logical operations (NOT and AND-gates). These modules are combinational circuits and simple to design which consume little amount of power.

TABLE I. SECTOR GENERATION

Rules	Condition	Sector
Rule-1	If (($X_d > 0$) & ($X_q > 0$) & ($X_d > X_q/2$))	1
Rule-2	If (($X_d > 0$) & ($X_q > 0$) & ($X_d < X_q/2$))	2
Rule-3	If (($X_d < 0$) & ($X_q > 0$) & ($X_d > X_q/2$))	3
Rule-4	If (($X_d < 0$) & ($X_q < 0$) & ($X_d > X_q/2$))	4
Rule-5	If (($X_d > 0$) & ($X_q < 0$) & ($X_d < X_q/2$))	5
Rule-6	If (($X_d > 0$) & ($X_q < 0$) & ($X_d > X_q/2$))	6

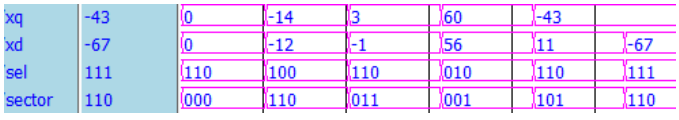


Fig. 5. Simulation Results of Sector Generation Module.

The switching time module is designed based on Table II for the generation of T_1, T_2 and T_0 using 3-bit sector generation module inputs. The T_1 and T_2 are modeled using simple right shifter, addition, and subtraction operations by using Xd and X_q.

The $\sqrt{3}$ square root module is designed using a comparator, simple multiplier, and subtractor operations. The simulation results of square root module are represented in Fig. 6.

The Switching time T_0 is generated by subtracting $\sqrt{3}$ from (T_1+T_2). The (T_1+T_2) operation is designed using an 8-bit carry-save adder (CSA). The CSA contains seven 2-bit Ripple carry adders (RCA) and three 6:3 Multiplexers. The carry-in is assumed as zero. The simulation results of 8-bit CSA are represented in Fig. 7.

The PWM module generation is designed by using the counter, Switching time and Sectors. The PWM module generation for sector -1 is tabulated in Table III. The Counter range is fixed to $2T_0 + 2T_1 + 2T_2$. The PWM pulse of a, b, and c is in the form of '0' or '1'. The T_{01} is used in Table III, which is generated using a one-bit right shifting of T_0, and it is represented in Fig. 3. Similarly, by using Sector 2-6, with different switching times, the PWM pulses are generated for a, b and c. The PWM module generation architecture uses simple shifters, adders, and multipliers. Overall, the present work uses only combinational circuits and Logical operations to construct the DSV-PWM module.

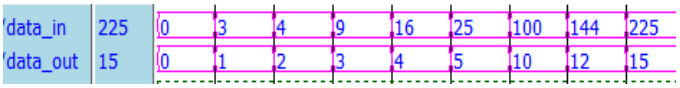


Fig. 6. Simulation Results of Square Root Module.

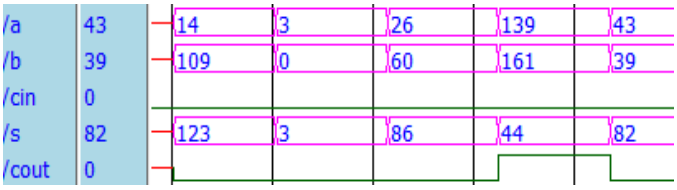


Fig. 7. Carry Save Adder (CSA) Module Simulation Results.

TABLE II. SWITCHING TIME GENERATION

Sector	T_1	T_2	T_0
1	X_d - (X_q >>1)	X_q	$\sqrt{3}$ - T_1 + T_2
2	X_d + (X_q >>1)	(X_q >>1) - X_d	$\sqrt{3}$ - T_1 + T_2
3	X_q	X_d - (X_q >>1)	$\sqrt{3}$ - T_1 + T_2
4	-X_d + (X_q >>1)	-X_q	$\sqrt{3}$ - T_1 + T_2
5	-X_d - (X_q >>1)	X_d – (X_q>>1)	$\sqrt{3}$ - T_1 + T_2
6	-X_q	X_d + (X_q >>1)	$\sqrt{3}$ - T_1 + T_2

TABLE III. PWM GENERATION FOR SECTOR-1

Sector =1			
Counter	*PWM-a*	*PWM-b*	*PWM-c*
T_{01}	1	0	0
T_{01}+ T_1	1	1	0
T_{01}+ T_1 +T_2	1	1	1
T_0 + T_1 +T_2	1	1	1
T_0 + T_1 +T_2 + T_{01}	1	1	0
T_0 + T_1 +2T_2 + T_{01}	1	0	0
T_0 +2T_1 +2T_2 + T_{01}	0	0	0

IV. RESULTS AND DISCUSSION

The proposed Digital Space Vector PWM (DSV-PWM) module is designed and implemented on Artix-7 FPGA for 3Φ-VSI applications. The DSV-PWM is designed using Xilinx ISE 14.7 Environment with Verilog HDL and simulated on Modelsim 6.5f Simulator. The DSV-PWM is implemented on Artix-7 FPGA with the device of XC7A100T-3 CSG 324. The DSV-PWM Module provides low-cost FPGA implementation with high performance. The simulation results of the DSV-PWM Module are represented in Fig. 8.

Once clock (clk) is activated with active-low reset (rst) signal, The DSV-PWM operation starts. The 3-bit select line (sel) is used in the Sector generation module to perform the signed shifting operation. The 3-phase va, vb, and vc are 8-bit voltage inputs and generate the DSV-PWM outputs like xd and xq by Xqd Module 8-bit outputs, 3-bit selection generation output, t0, t1, and t2 are 8-bit switching time outputs and final pa, pb, pc are PWM outputs.

The PWM outputs are generated based on 3-Phase voltage inputs along with sector generation and switching time module. The Resource utilization of DSV-PWM is generated after synthesis operation in Xilinx Tool. The DSV-PWM resource utilization in terms of Chip area, Timing analysis, and power Utilization are tabulated in Table IV.

The DSV-PWM utilized slice register of 27, slice LUT's of 337, and LUT-FF pairs of 27 on Artix-7 FPGA. The DSV-PWM module works at 597.83MHz with a minimum period of 1.67ns and a combinational delay of 2.17ns. The power utilization is analyzed for the DSV-PWM module using the Xilinx Xpower analyzer, and it consumes the total power of 110mW, which includes the static power of 82mW and dynamic power of 28mW. The overall power utilization of DSA-PWM is 100mW, which is quite less and suitable for real-time power electronics applications.

The DSV-PWM module contains four main sub-modules, namely, Xdq Module, Sector generation module (SGM), Square root Module (SRM), and CSA module. The slice – LUT's and the Combination delay of DSV-PWM sub-modules are tabulated in Table V. The X_{dq}, SGM, SRM, CSA utilize 24, 42, 8, and 12 slice LUTs respectively. Similarly, the Xdq, SRM and CSA have a combinational delay of 2.1ns, 1.74ns, and 2.52ns, respectively. The remaining Slice –LUT's of DSV-PWM Module are utilized by the Switching time module and PWM generation Module.

The proposed DSV-PWM is compared with similar existing SV-PWM technique with better improvements in design and hardware constraints. The DSV-PWM module is compared with existing SV-PWM [23] and is tabulated in Table VI for resource constraints.

The DSV-PWM improves around 69%, 14%, and 57% of less overhead in Slice registers, Slice LUTs, and Maximum frequency (MHz), respectively than the existing SV-PWM [23] technique. Similarly, the DSV-PWM dynamic power utilization by using different clock frequency is compared with existing SV-PWM [23] with better improvements is tabulated in Table VII, and graphical representation is showed in Fig. 9.

The DSV-PWM consumes the dynamic power of 7, 14, 21, and 28mW, which is quite less compared to the existing SV-PWM [23] technique with 23, 28, 33, and 37 mW against clock frequency of 25, 50, 75 and 100 MHz respectively.

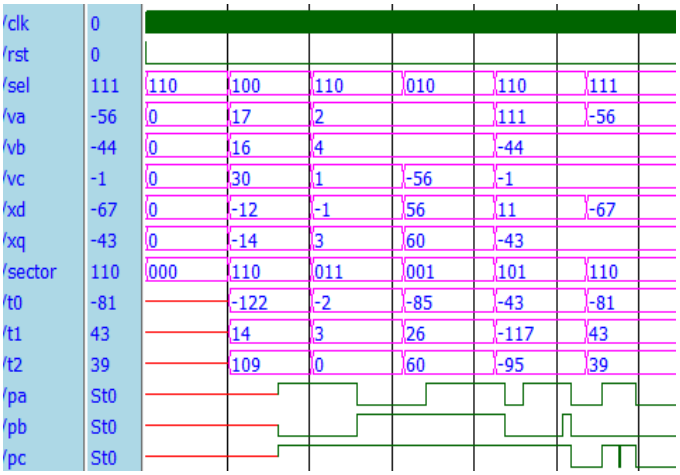


Fig. 8. Simulation Results of DSV-PWM Module.

TABLE IV. HARDWARE RESOURCE UTILIZATION OF DSV-PWM MODULE ON ARTIX-7 FPGA

Resources	**DSV-PWM**
Chip Area	
Slice Registers	27
Slice LUTs	337
LUT-FF pairs	27
Time	
Minimum period (ns)	1.673
Max. Frequency (MHz)	597.836
Combinational Delay (ns)	2.17
Power	
Dynamic Power (W)	0.028
Total Power (W)	0.11

TABLE V. SLICES LUT'S AND COMBINATIONAL DELAY (NS) OF DSV-PWM- SUB MODULES

Sub Modules	**Slice LUTs**	**Combinational Delay (ns)**
X_{dq} Module	24	2.106
Sector Generation Module	42	NA
Square Root Module	8	1.747
CSA Module	12	2.523

TABLE VI. RESOURCE COMPARISON OF DSV-PWM MODULE WITH [23]

Resources	**Ref [23]**	**Proposed**
FPGA Device	Cyclone II	Artix-7
Slice Registers	88	27
Slice LUT's	392	337
Max. Frequency (MHz)	253.85	597.836

TABLE VII. DYNAMIC POWER (MW) COMPARISON OF DSV-PWM MODULE WITH [23]

Clock Frequency (MHz)	**Dynamic Power (mW)**	
	Ref [23]	*Proposed*
25	23	7
50	28	14
75	33	21
100	37	28

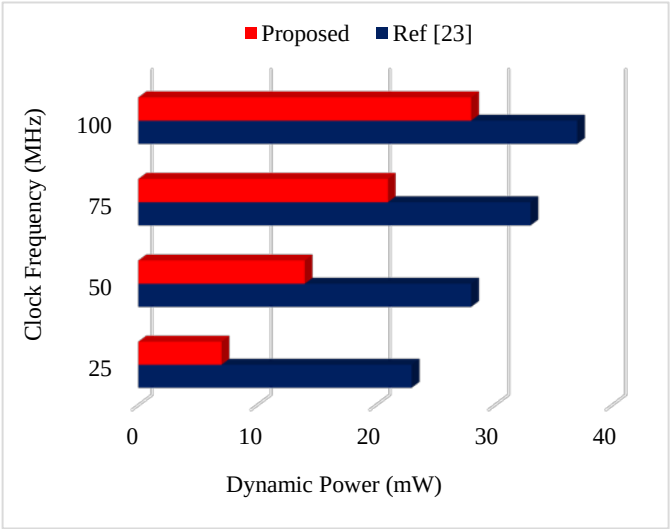


Fig. 9. Graphical Representation of DSV-PWM Dynamic Power (mW) by Concerning [23].

V. CONCLUSION AND FUTURE WORK

The DSV-PWM module is designed and implemented in Artix-7 FPGA. The DSV-PWM Module contains Xdq Module, Sector generation module (SGM), Switching time module, and PWM Generation Module. These modules are designed using simple logical operators and combinational circuits, which improve the hardware constraints like chip area, Power and frequency along with the performance. The DSV-PWM uses less chip area (< 1%), 110 mW total power and works at 597.836 MHz operating frequency on the FPGA device. The DSV-PWM method is also compared with the existing SV-PWM method with more considerable improvement in chip area, Frequency, and Dynamic power. In

the future, The DSV-PWM will be used in real-time power electronics applications and also to optimize the hardware constraints.

REFERENCES

[1] Yu, Zhenyu, Arefeen Mohammed, and Issa Panahi. "A review of three PWM techniques." In Proceedings of the 1997 American Control Conference (Cat. No. 97CH36041), vol. 1, pp. 257-261. IEEE, 1997.

[2] Tzou, Ying-Yu, and Hau-Jean Hsu. "FPGA realization of space-vector PWM control IC for three-phase PWM inverters." IEEE Transactions on power electronics 12, no. 6 (1997): 953-963.

[3] Patil, Sumit R., and V. N. Kalkhambkar. "Hybrid space vector pulse width modulation voltage source inverter-a review." In 2017 International Conference on Data Management, Analytics, and Innovation (ICDMAI), pp. 200-204. IEEE, 2017.

[4] Pongiannan, R. K., and N. Yadaiah. "FPGA based space vector PWM control IC for three-phase induction motor drive." In 2006 IEEE International Conference on Industrial Technology, pp. 2061-2066. IEEE, 2006.

[5] Rajendran, R., and N. Devarajan. "FPGA implementation of space vector PWM technique for voltage source inverter fed induction motor drive." In 2009 Second International Conference on Computer and Electrical Engineering, vol. 2, pp. 422-426. IEEE, 2009.

[6] Alvarez, Jacobo, Oscar Lopez, Francisco D. Freijedo, and Jesús Doval-Gandoy. "Digital parameterizable VHDL module for multilevel multiphase space vector PWM." IEEE Transactions on Industrial Electronics 58, no. 9 (2010): 3946-3957.

[7] Wang, Yingnan, and Uwe Schaefer. "Real-time simulation of an FPGA based space vector PWM controller." In SPEEDAM 2010, pp. 833-838. IEEE, 2010.

[8] Somlal, Jarupula, M. Venu Gopala Rao, and S. Prabhakar Karthikeyan. "Experimental investigation of an indirect current-controlled Fuzzy-SVPWM based Shunt Hybrid Active Power Filter." In 2016 IEEE Region 10 Conference (TENCON), pp. 801-806. IEEE, 2016.

[9] Rait, Aishanou Osha, and Praveen Bhosale. "FPGA implementation of space vector PWM for speed control of 3-phase induction motor." In 2011 International Conference on Recent Advancements in Electrical, Electronics and Control Engineering, pp. 221-225. IEEE, 2011.

[10] Janik, Dusan, Tomas Kosan, Jaroslav Sadsky, and Zdenek Peroutka. "Implementation of SVPWM algorithm without trigonometric functions." In 2014 International Conference on Applied Electronics, pp. 131-134. IEEE, 2014.

[11] Holtz, Joachim, Markus Höltgen, and Jens Onno Krah. "A space vector modulator for the high-switching frequency control of three-level SiC inverters." IEEE Transactions on Power Electronics 29, no. 5 (2013): 2618-2626.

[12] Kassas, Mahmoud, and Naseer Ahmed. "Simulation and Implementation of Space Vector PWM Using Look-Up Table." Arabian Journal for Science and Engineering 39, no. 6 (2014): 4815-4828.

[13] ÇELIK, Hakan, Mahmut Temel ÖZDEMIR, and Eyyüp ÖKSÜZTEPE. "Generation of Space Vector PWM Signals Based on the FPGA in Real Time." International Journal of Natural and Engineering Sciences 9, no. 3 (2015): 54-59.

[14] Liang, Xuexiu, Min Luo, Bo Zhao, Yanwei Yuan, and Zhi Chen. "Research and implementation of SVPWM control algorithm based on FPGA." In 2016 IEEE International Conference on Mechatronics and Automation, pp. 22-26. IEEE, 2016.

[15] Lotfi, Elhabib, Mustapha Elharoussi, and Elhassane Abdelmounim. "VHDL Design and FPGA Implementation of the PWM Space Vector of an AC Machine Powered by a Voltage Inverter." In Proceedings of the Mediterranean Conference on Information & Communication Technologies 2015, pp. 41-47. Springer, Cham, 2016.

[16] Salem, Aboubakr, Frederik De Belie, and Jan Melkebeek. "A novel space-vector PWM computations for a dual three-level T-type converter applied to an open end-winding induction machine." In 2016 Eighteenth International Middle East Power Systems Conference (MEPCON), pp. 633-638. IEEE, 2016.

[17] Pu, Tianyu, Feifei Bu, Wenxin Huang, and Lin Zhu. "Implementation of random SVPWM strategy for three-phase voltage source inverter based on FPGA." In 2017 20th International Conference on Electrical Machines and Systems (ICEMS), pp. 1-4. IEEE, 2017.

[18] Khiavi, A. Moallemi, J. Sobhi, Z. Daie Koozehkanani, and M. Farhadi Kangarlu. "FPGA-based reconfigurable PWM generator for power electronic converter applications." Journal of Control, Automation, and Electrical Systems 28, no. 4 (2017): 516-531.

[19] García, Raymundo Cordero, Igor Esdras Silva Ono, Vitória dos Santos Fahed, and João Onofre Pereira Pinto. "Simplified educational platform for SVPWM control of a two-level three-phase inverter using Matlab GUI and FPGA." In 2017 Brazilian Power Electronics Conference (COBEP), pp. 1-6. IEEE, 2017.

[20] Sumam, M. J., and G. Shiny. "Rapid prototyping of high-performance FPGA controller for an induction motor drive." In 2018 8th International Conference on Power and Energy Systems (ICPES), pp. 76-80. IEEE, 2018.

[21] Sumam, M. J., and G. Shiny. "Prototype Non-Volatile FPGA SVPWM Controller for 3-Level Inverter." In 2018 8th IEEE India International Conference on Power Electronics (IICPE), pp. 1-6. IEEE, 2018.

[22] Chinmaya, K. A., and Girish Kumar Singh. "Experimental analysis of various space vector pulse width modulation (SVPWM) techniques for dual three-phase induction motor drive." International Transactions on Electrical Energy Systems 29, no. 1 (2019): e2678.

[23] Rashidi, Bahram, and Mehran Sabahi. "High-performance FPGA based digital space vector PWM three-phase voltage source inverter." International Journal of Modern Education and Computer Science 5, no. 1 (2013): 62.

Lung Cancer Detection using Bio-Inspired Algorithm in CT Scans and Secure Data Transmission through IoT Cloud

C.Venkatesh[1]
Research Scholar, Department of ECE
Koneru Lakshmaiah Education Foundation
Green Fields, Vaddeswaram, Guntur-522502
Andhra Pradesh, India

Polaiah Bojja[2]
Professor, Department of ECE
Koneru Lakshmaiah Education Foundation
Green Fields, Vaddeswaram, Guntur-522502
Andhra Pradesh, India

***Abstract*—Primary recognition of pulmonary cancer nodules eloquently increases the odds of survival, also leads it solider problem to resolve, as it often relies on a tomography scan filmic examination. By increasing the possibility of effective treatment, earlier tumor diagnosis decreases lung cancer mortality. Radiologists usually diagnose lung cancer on medical images by a systematic analysis that consumes more time and is unreliable often, because of the substantial improvement in the transmission of data in the healthcare sector, the protection and integrity of medical data has been a huge problem for healthcare applications. This study utilizes computational intelligence techniques. For detection and data transmission, a novel Hybrid model is therefore proposed in this paper. Two steps are involved in the proposed method where diverse image processing procedures are used to detect cancer in the first step using MATLAB and data transfer to authorized persons via the IoT cloud in the second stage. The simulated steps include pre-processing, segmentation by Otsu thresholding along with swarm intelligence algorithm, extraction of features by local binary pattern and classification using the support vector machine (SVM). This work demonstrates the dominance of swarm-intelligent framework over the conventional algorithms in terms of performance metrics like sensitivity, accuracy and specificity as well as training time. The tests carried out show that the model built can achieve up to 92.96 percent sensitivity, 93.53 percent accuracy and 98.52 percent specificity.**

***Keywords*—Pulmonary; mortality; carcinogenic; swarm intelligence; IoT**

I. Introduction

A malicious tumor characterized by uninhibited cell evolution in lung tissues is lung cancer. Carcinomas are the majority of cancers that originate in the lungs. Most of the patients are diagnosed at an advanced stage due to no apparent early cancer symptoms [1], which typically results in high costs and a worse prognosis. In medical diagnosis and treatment, medical imagery has become important. These images playa extensive part in clinical applications since medical professionals expose attention in exploring the interior structure [2]. Several procedures have been established based on cross-sectional images, such as magnetic resonance imaging (MRI) or computed tomography (CT) or other topographic modes [3,4,5]. The application of medical image processing has played an important role in both technological and clinical aspects in helping to identify and examine anomalies by making it easier for medical practitioners to work with more scientific and sophisticated approaches to solve the problem [6]. A CT Scan obtains images of an organ that cannot be seen on a regular x-ray that results in earlier diagnosis [7]. The biggest issue with lung cancer is that these cases of cancer are later diagnosed, making treatments more complicated and decreasing the probability of survival subsequently [8]. It is therefore important to recognize a modern, robust method for diagnosing lung cancer at an earlier stage [9]. For cancer diagnosis, CT scan images are being used; they are analyzed by radiologists to recognize and identify nodules into malignant and benign nodules [10]. These techniques, require highly trained radiologists who are not in particular, accessible to people in remote regions. In addition, in manual testing, there seems to be a significant chance of human error, and therefore optimization-based systems are required that can assist radiologists in diagnosing and help minimize the incidence of false results [11]. To detect the nodules, their form, scale, and other characteristics from CT scans, digital image processing techniques can be used. In order to design specialist support systems for the diagnosis of various diseases such as lung cancer identification, medical image processing has been widely and rapidly implemented. In addition, the existence of nodules that define a patient's destiny is also very complex, as their shape and size differ from slice to slice. They are often connected, such as arteries or bronchioles, to other pulmonary structures [12]. It can also vary the color in which they appear on CT scans. These variables contribute to the difficulty of defining them.

In this work an efficient framework is proposed to decipher the lung cancer at an early stage and also data transmission to medical practitioners. Detection stage involves pre-processing, separation of nodules with optimization, feature extraction and classification. Transmission stage involves transmission of statistical parameters through IoT as well as MATLAB IoT cloud Thing speak. As direct data transmission is not possible, thingspeak module has been considered for effective transmission. The structure of this paper contains Section II: related work, discusses about the

previous works, Section III proposed methodology represents the methods, block diagram and corresponding algorithms, Section IV shows segmentation with optimization concepts, Section V is the Extraction by LBP method, Section VI is Classification by SVM and Section VII presents the Simulation results, provides output images, statistical values and corresponding thingspeak plots.

II. Related Work

Malayil Shanid et al. [13] in 2020 presented a pulmonary cancer detection system with SE (slap elephant) optimization and deep learning techniques. By this work authors gained 96 percent accuracy.

Noor Khehrah et al. [14] in 2020 presented a pulmonary nodule detection system with thresholding and statistical features techniques. By this work authors gained 93.75 percent sensitivity.

Shankar et al. [15] in 2019 documented an Alzheimer's identification technique that uses the gray-level run-length matrix and scale-invariant conversion to extract different features. By this framework 96.23 percent accuracy is gained

K.Senthil Kumar et al. [16] in 2019 recognized a lung cancer detection scheme by GCPSO. By this model 95 percent meticulousness is acquired.

C.Venkatesh et al. [17] in 2019 projected a detection scheme by genetic approach .By this approach 90 percent precession.

Vijh et al. [18] in 2019 proposed a detection procedure using whale optimization algorithm and SVM By this work authors gained 95 percent accuracy.

Preethijoon et al. [19] in 2019 projected a respiratory cancer recognition strategy with the SVM classifier using fuzzy c & k-mean partition methodologies. By this model less than 93 percent accuracy is gained.

S.Perumal et al. [20] in 2018 documented an enhanced ABC optimization for cancer detection and classification. 92 percent proficiency is accomplished by using this procedure.

Uc-ar et al. [21] in 2019 recommended a detection model by Laplacian and Gaussian filter model with CNN architecture. In this method 72.97 percent precision is attained.

In all the above conventional (existing) techniques, the accuracy is lower. In this paper, therefore, an assorted approach is projected where PSO has been used for segmentation to obtain greater accuracy along with SVM classifier and LBP for feature extraction.

III. Proposed Methodology

Fig. 1 shows a detailed view of the proposed system where it involves two phases. In first phase lung cancer is identified from CT images using the optimization method of swarm intelligence. In second phase the data transferred through thingspeak and IoT. Initially, the CT input images of lung cancer are read from private and public databases. The attained CT images typically encompass a noise [22]. In pre-processing step by the use of median filter the noise is condensed. Then, output image of filter is segmented by swarm optimization with the Otsu thresholding technique. The partitioned image then endures an extraction process by LBP to excerpt textural topographies. Then, the extracted topographies fed to classification stage to detect whether image is normal or abnormal. If the image is anomalous the attributes are determined and transferred to the medical persons via Thingspeak or IoT.

A. Pre-processing

Optimal reliability inspection is improved by image pre-processing. All images probably contain noise, so the image has to be pre-processed by median filtering to suppress the noise [23]. It improves the aesthetic value and accuracy of the image.

B. Median Filtering

This filter reduces the noise of salt and pepper and also retains the image edges. The random bit error in a communication channel generates salt and pepper noise. The median filter is a basic regional sliding kernel that swaps the kernel's centre point with the kernel's average of all the pixels [24].

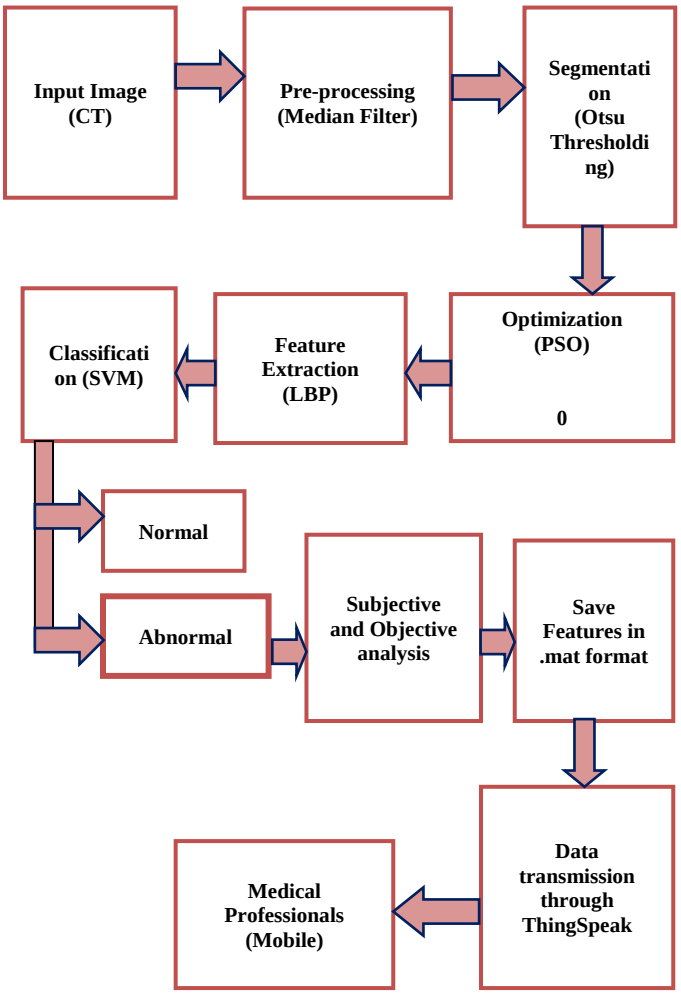


Fig. 1. Proposed Framework.

IV. Segmentation with Optimization

A. Segmentation

Segmentation tends to slice the image pixels into sections that are directly connected to the objects of the image. Typically, it is the basic step for all the computer vision systems [25]. Usually, segmentation algorithm relies on pixel intensities. All the algorithms entail certain threshold parameter to be set. The appropriate threshold results in greater segmentation. The threshold value is set bestowing to the intensity values [26]. To achieve best threshold value in this work otsu thresholding techniques is used.

B. Otsu Thresholding

Otsu thresholding is based on the basic idea of identifying a threshold that mitigates the weighted variance in the class, which is the same as optimizing the variance among classes [27]. It works directly on the gray-level bimodal histograms. Also no other entity structure description and regional continuity is required. It has set numbers, but can be modified to suit locally [28].

C. Algorithm

1: Estimate Histogram and likelihood of every intensity value

2: Initialize xi(0) and μi(0)

3: Move across all feasible points t=1 Upgrade wi and μi

4: Determine σb2 (t) and considered it as preferred threshold

5: Measure two maxima'sσb12 is the higher limit and the value limit is greater or equal to σb22

6: optimal threshold = (Th1+Th2)/2)

D. Particle Swarm Optimization

It is a metaheuristic process used effectively in the study of medical images [29]. It mimics the communal movements of food-seeking birds [30]. Because of simplicity and generality, this algorithm has been effectively used for cancer detection. PSO falls quickly, however, into the local optimal solution. The argument and alliance of information is a good basic principle of PSO. In this process every particle has a preliminary position and velocity [31]. Each particle's position signifies a probable solution and has a fitness value calculated by its fitness function. The position and speed are altered based on the fitness value and gets updated. After updating with a group of random particles, the procedure pursuits for optima. The equations to update position and speed are as follows [32]:

$$S(t+1) = S(t) + a_1 c_1[P_{best}(t) - m(t)] + a_2 c_2[G_{best}(t) - m(t)] \quad (1)$$

m(t+1)=m(t)+V(t+1)

Where c_1, c_2 are random values and a_1, a_2 are acceleration values

The optimistic concert relies on the fitness function. The equation of fitness function is given below:

$$f = \sum_{k=1}^{n} \frac{Interband\ distance}{Intraband\ distance} \quad (2)$$

Where k is the number of bands

The weight of inertia enables global searching which enhances the rate of convergence and also reduces the iterations, while a small weight of inertia enables local searching [1].

$$S(t+1) = W * S(t) + a_1 c_1[P_{best}(t) - m(t)] + a_2 c_2[G_{best}(t) - m(t)] \quad (3)$$

Where w is weight of inertia the values of constant and random inertia are 0.7 & 0.5+rad()/2 respectively.

Algorithm

1. Initialize with some random position and velocity vectors for the particles.
2. For all particles in the group calculate fitness value.
3. When fitness(p) is better compared to fitness(P_{best}) and P_{best} =P.
4. Assign G_{best} as the best particle value.
5. Determine each particle speed velocity is measured for each particle.
6. Update speed and position of particles.

V. Feature Extraction by Local Binary Pattern (LBP)

In diverse fields, the LBP method has been used. It is a texture description operator based on symbols of variances between central and adjacent pixels [33]. In this technique a binary cypher of every pixel is gained by thresholding its surrounding pixel with the centre pixel. If the value of adjacent pixel is superior or equal to threshold value it assigns 1 otherwise 0. First, to evaluate the frequency values of binary patterns, a histogram is constructed [34]. The likelihood of a binary pattern contained in the image is represented by texture characteristics. The equation of LBP is as follows.

$$LBP(m_c, n_c) = \sum_{k}^{L-1} 2^k f(g_k - g(m_c, n_c)) \quad (4)$$

Where g (m_c, n_c) is the grey values of center and surrounding pixels, f(x) is the function whose value is 0 if g<0 and 1 if g≥0.Finally, the LBP value is the center pixel (m_c, n_c).

VI. Classification by SVM

SVM was formerly used to categorize linearly detached data for binary classification. The preliminary purpose is to discover an optimum hyper plane [35,36]. The Hyper plane is a two-class frontier. It not separates two classes but also enhances the boundary between two classes. The boundary is the major distance in each class among the hyper plane and the nearest data [37]. The ideal boundary is attained by maximizing the distance between the support vector and the hyper plane Let m=(x_1, x_2) and W(m,-1) then for each class hyper plane can be expressed as y(t) and the equations can be written as follows:

$$y(t) = mx + n \quad (5)$$

$$mx_1 - x_2 + n = 0 \quad (6)$$

$$w.x + n = 0 \quad (7)$$

If hyperplane is defined then based on assumptions the hypothesis function can be written as follows:

$$h_f(x_i) = \begin{cases} +1 \; if \; w.x + n \geq 0 \\ -1 if \; w.x + n < 0 \end{cases} \quad (8)$$

From the above equation if the point is above the plane then it is categorized as +1 otherwise -1.The data set used in the proposed method is as shown in Fig. 2.

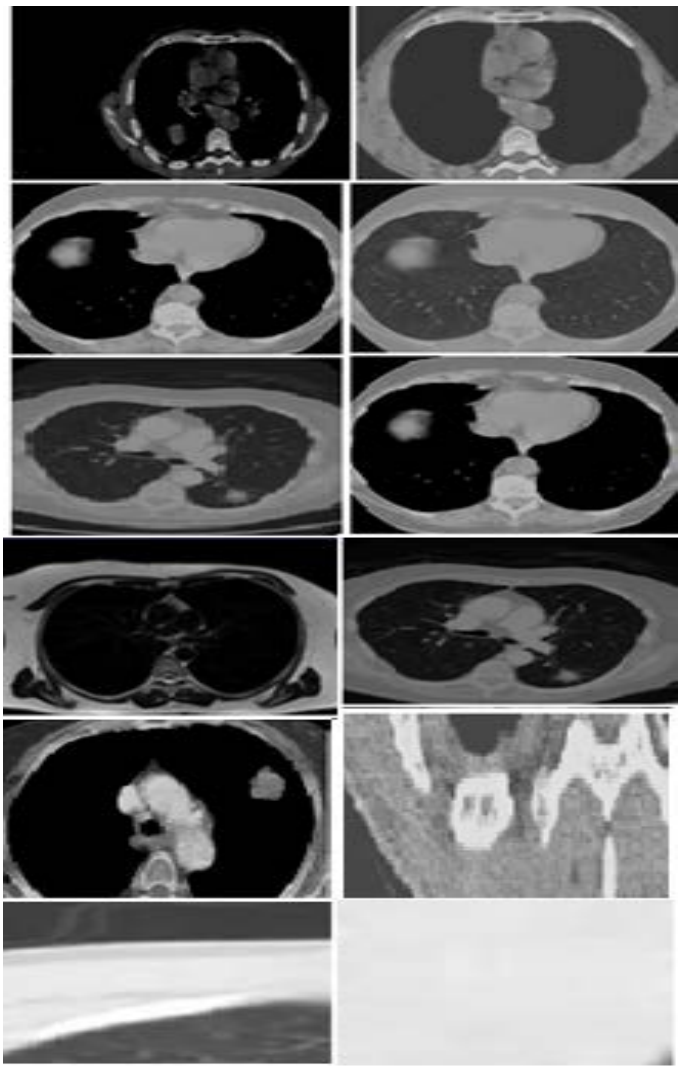


Fig. 2. Sample Data Set (Courtesy: Satyam Diagnostic Center, Anantapuramu).

VII. Simulation Results

In this work, firstly by using MATLAB, the lung cancer is detected and secondly, the attributes attained are transferred to doctors using Thingspeak. The CT images are collected from private hospital in Ananthapuramu. In this approach the problem of thresholding is considered as an optimization issue and can be resolved by the particle swarm principle. In this work algorithms were implemented using MATLAB (R2017b) on an Intel Core i5 PC at 1.80Ghz with a total physical memory of 8GB RAM.

A. Detection Phase using MATLAB

In Fig. 3 and 4, the input and median filter output of CT lung cancer images are shown. The CT image commonly has noise with less falsification. The input image is processed into a median filter to abolish noise and falsification in the image.

Fig. 3. CT Input Lung Image.

Fig. 4. Median Filter Output.

To discern mutant in CT image, together with optimization technique, the filter output image is partitioned with Otsu thresholding. At first, the CT image is partitioned through simple Otsu thresholding in which the segmented classes, "all category variance," are greatly increased. By refining by particle swarm optimization, the results acquired from the thresholding method should be optimized.

In PSO, by following the maximum particles present, the potential solutions, called particles, pass through the space of the problem. In Fig. 5 and 6, the segmented output images are shown. Fig. 7 shows the classifier output. The features of image are extracted by feature extraction with LBP after classification.

1) Statistical Results of Existing Method: The traits attained from the proposed model method are shown in Table I.

In the abovementioned table, the proposed method is proved as best to obtained the less MSE at 0.186 also high PSNR at 42.729 and high accuracy at 96.550% as compared to conventional systems.

B. Data Transmission

Finally, the obtained result is plotted as graph in the ThingSpeak and is shared to the authorized personnel. Fig. 8 to 13 shows the ThingSpeak plots which are shared to medical professionals.

Fig. 5. Segmentation Output by Thresholding.

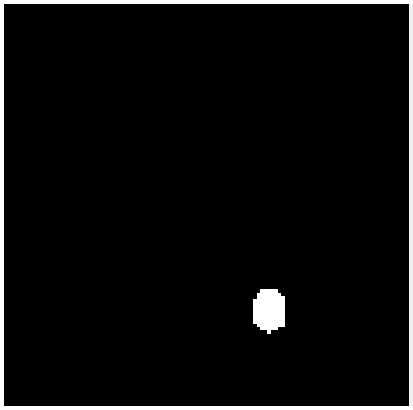


Fig. 6. Segmented Output by PSO.

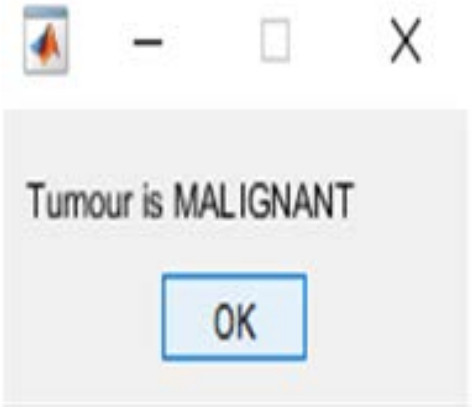


Fig. 7. Classifier Output.

TABLE I. ATTRIBUTES OBTAINED FROM PROPOSED METHOD

PARAMETERS	VALUES
MSE	0.186
PSNR	42.729
Specificity(%)	92.180
Sensitivity(%)	97.210
Accuracy (%)	96.550
Entropy	0.610
Correlation	0.723
Processing Time	20.56

Fig. 8. Accuracy Data Plot of ThingSpeak.

Fig. 9. Specificity Data Plot of ThingSpeak.

Fig. 12. Entropy Data Plot of ThingSpeak.

Fig. 10. Mean Square Error Data Plot of ThingSpeak.

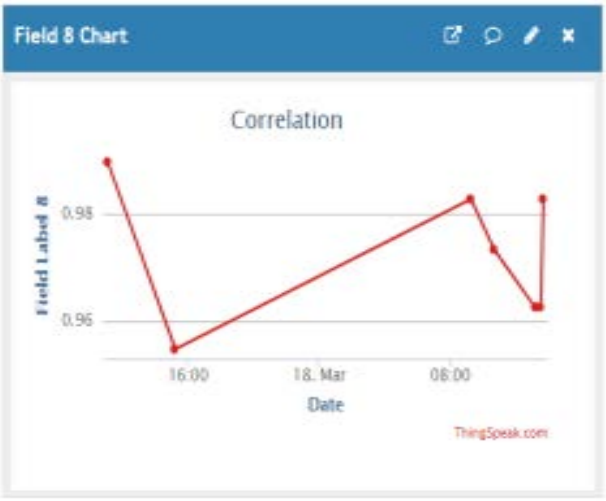


Fig. 13. Correlation Data Plot of ThingSpeak.

Fig. 11. PSNR Data Plot of ThingSpeak.

VIII. Conclusion

In this paper, a new strategy to early detection, prediction and diagnosis has been introduced in order to improve patient safety and mitigate the risks also the data is transferred to medical professionals through MATLAB IoT cloud called Thingspeak. The image pre-processing and segmentation procedures are used for partition the lung nodule along with particle swarm algorithm. The several features are extracted by LBP to the study of statistical information that assists in the decision-making process by SVM whether the tumour is malicious or non-malicious. The proposed approach outperforms by providing an accuracy of 96.5 percent. The interpretation of the obtained results are limited with accuracy and specificity parameters due to convergence of local optima in the algorithm. Further, these results can be analyzed more effectively using deep learning techniques and advanced hardware processors in near future.

REFERENCES

[1] A. A. Brindha, S. Indirani, and A. Srinivasan, “Lung cancer detection using SVM algorithm and optimization techniques,”Journal of Chemical and Pharmaceutical Sciences, vol. 9, no. 4,2016.

[2] T. Saba, S. Al-Zahrani, and A. Rehman, "Expert system for offline clinical guidelines and treatment," Life Sci J, vol. 9, pp. 2639-2658, 2012.

[3] A. Norouzi, M. S. M. Rahim, A. Altameem, T. Saba, A. E. Rad, A. Rehman, et al., "Medical image segmentation methods, algorithms, and applications," IETE Technical Review, vol. 31, pp. 199-213, 2014.

[4] R. S. Kapse, S. Salankar, and M. Babar, "Literature survey on detection of brain tumor from MRI images," IOSr Journal of electronics and communication engineering, vol. 10, pp. 80-86, 2015.

[5] M. Majid, A. Abidin, N. Anuar, K. Kadiran, M. Karis, Z. Yusoff, et al., "A comparative study on the application of binary particle swarm optimization and binary gravitational search algorithm in feature selection for automatic classification of brain tumor MRI," Journal of Fundamental and Applied Sciences, vol. 10, pp. 486-498, 2018.

[6] A. Kashyap, V. K. Gunjan, A. Kumar, F. Shaik, and A. A. Rao, "Computational and Clinical Approach in Lung Cancer Detection and Analysis," Procedia Computer Science, vol. 89, pp. 528-533, 2016.

[7] J. L. Prince and J. M. Links, Medical imaging signals and systems: Pearson Prentice Hall Upper Saddle River, NJ, 2006.

[8] K.Raveendra, P. V. N Reddy, P. V. V Kishore, “A Review on Signature and Logo Identification and Extraction using Automatic Logo Based Document Image Retrieval Methods “ Helix, Volume:08, Issue :01, 2018, Pg.:2726-2729.

[9] B. Rani, A. K. Goel, and R. Kaur, “A modified approach for lung cancer detection using bacterial forging optimization algorithm,” International Journal of Scientific Research Engineering and Technology, vol. 5, no. 1, 2016.

[10] Ning, J.; Zhao, H.; Lan, L.; Sun, P.; Feng, Y. A Computer-Aided Detection System for the Detection of Lung Nodules Based on 3D-ResNet. Appl. Sci., Volume :09, Issue: 5544, 2019.

[11] Zhang, G.; Yang, Z.; Gong, L.; Jiang, S.; Wang, L.; Cao, X.; Wei, L.; Zhang, H.; Liu, Z. An Appraisal of Nodule Diagnosis for Lung Cancer in CT Images. J. Med. Syst., Volume:43, Issue: 181, 2019, 43, 181.

[12] Narayanan, B.N.; Hardie, R.C.; Kebede, T.M Performance analysis of a computer-aided detection system for lung nodules in CT at different slice thicknesses. J. Med. Imaging, Volume :5, 2018, Pg.No.:5–10.

[13] Malayil Shanid, Anitha A, “Lung Cancer Detection from CT Images Using Salp-Elephant Optimization Based Deep Learning”, Biomedical Engineering: Applications, Basis and Communications, Volume: 32, Issue No:1, 2020, Pg.No:2050001-1 to 2050001-14.

[14] Noor Khehrah, Muhammad Shahid Farid, Saira Bilal, Muhammad Hassan Khan, “Lung Nodule Detection in CT Images Using Statistical and Shape-Based Features”, Journal of Imaging, Volume:06, Issue:06, 2020, Pg.No:1-14.

[15] Shankar K, Lakshmanaprabu SK, Khanna A, Tanwar S, Rodrigues JJ, Roy NR (2019) Alzheimer detection using Group Grey Wolf Optimization based features with convolutional classifier. Comput Electr Eng 77:230–243.

[16] K. Senthil Kumar , K. Venkatalakshmi , K. Karthikeyan, “Lung Cancer Detection Using Image Segmentation by means of Various Evolutionary Algorithms”Hindawi, Computational and Mathematical Methods in Medicine, Research article , 2019,Pg.No:1-16 Article ID 4909846.

[17] C.Venkatesh, Polaiah Bojja, “A Novel Approach for Lung Lesion Segmentation Using Optimization Technique”, Helix the scientific explorer , Volume:9, Issue No: 1, 2019,Pg. 4832-4837,E-ISSN: 2319-5592, DOI 10.29042/2019-4832-4837.

[18] Vijh S, Gaur D, Kumar S (2019) An intelligent lung tumor diagnosis system using whale optimization algorithm and support vector machine. Int J Syst Assur Eng Manag 11:1–11.

[19] PreetiJoon, ShaliniBhaskar Bajaj and AmanJatain, “Segmentation and Detection of Lung Cancer Using Image Processing and Clustering Techniques” Springer Nature Singapore, Progress in Advanced Computing and Intelligent Engineering, Advances in Intelligent Systems and Computing , Volume No:1, 2019, Pg.No:13-23.

[20] S. Perumal , T. Velmurugan, “Lung cancer detection and classification on CT CT scan images using enhanced artificial bee colony optimization”International Journal of Engineering & Technology, Volume No:7, Issue No:2.26, 2018, Pg. No.: 74-79.

[21] Uc¸ar M, Uc¸ar E (2019) Computer-aided detection of lung nodules in chest X-rays using deep convolutional neural networks.Sakarya Univ J Comput Inf Sci 2(1):41–52.

[22] Durga Prasad Tripathi, P. Pardhasaradhi , B.T.P.Madhav, “Statistical parameters-based image enhancement techniques in pure and nanodispersed 6O.O8 liquid crystalline compounds”, Phase Transitions, Volume:91, Issue :08, 2018, Pg.:820-832.

[23] B.T.P. Madhav, P. Pardhasaradhi, R.K.N.R. Manepalli, P.V.V. Kishore & V.G.K.M.Pisipat, “Image enhancement using virtual contrast imagefusion on Fe3O4 and ZnO nanodispersed decyloxybenzoic acid”, Liquid Crystals, Taylor & Francis, Volume:42, Issue:9, 2018,Pg.:1329-1336.

[24] C.Venkatesh, Polaiah Bojja, “Development of Qualitative Model for Detection of Lung Cancer Using Optimization”, International Journal of Innovative Technology and Exploring Engineering, Volume: 08, Issue: 09, July 2019, Page No.:.3143-3147.

[25] Md Zia Ur Rahman, B. M. K Reddy. “Efficient SAR Image Segmentation Using Bias Field Estimation”, Journal of Scientific & Industrial Research, volume :76, Issue:06, 2017, Pg.:335-338.

[26] Syed Inthiyaz, B.T.P.Madhav, P.V.V.Kishore, “ Flower image segmentation with PCA fused colored covariance and gabor texture features based level sets”, Ain Shams Engineering Journal, Volume:9, Issue: 4,2018, Pg.:3277-3291.

[27] Otsu N (1979) A threshold selection method from gray-level histograms.IEEE Trans Sys Man Cyber 9(1):62–66.

[28] Aparna Puvvadi, Polurie Venkata Vijay Kishore, “An Efficient Medical Image Watermarking Technique in E-healthcare Application Using Hybridization of Compression and Cryptography Algorithm”, Journal of Intelligent Systems, Volume:27, Issue:1,2017, Pg.:No:115-133.

[29] B. Malakonda Reddy, Md. Zia Ur Rahman, “SAR Electromagnetic Image Conditioning Using a New Adaptive Particle Swarm Optimization”, ACES JOURNAL, Volume:33, Issue :12, 2018, Pg.:1439-1446.

[30] Ahmed Sultan Al-Hegami , Abeer Saleh Hamdi Bin-Ghodel , “ A Particle Swarm based Approach for Classification of Cancer based on CT Scan” International Journal of Computer Applications Volume:178, Issue :12, 2019, Pg.No:26-31.

[31] P.V.V.Kishore, Ch. Raghava Prasad, “Computer vision based train rolling stock examination”, Optik, Elsevier, Volume:13, 2017, Pg.:427-444.

[32] Prasanna, N. Merrin; Bojja, Polaiah, “Optimization of Rotary Kiln in Cement Industry Using Conventional Control Systems” Helix, Volume:9, Issue:1, Pg.:4843- 4849.

[33] T. Ojala, M. Pietik¨ainen and D. Harwood, “A comparative study of texture measures with classification based on feature distributions” Pattern Recognition vol. 29, 1996.

[34] C.Venkatesh,K.Bhagyalakshmi,L.Sivayamini, “Detection of Diverse Tumefactions in Medial images by Various Cumulation Methods” International Research Journal of Engineering and Technology(IRJET), ,Volume: 04, Issue: 08,Aug2017, Pg.No:1195-1200.

[35] V. Ratna Bhargavi,V. Rajesh, “Computer Aided Bright Lesion Classification in Fundus Image Based on Feature Extraction”, International Journal of Pattern Recognition and Artificial Intelligence, Volume: 32, Issue :11, 2018,

[36] E.Kiran Kumar, P.V.V.Kishore, A.S.C.S.Sastry, M.Teja Kiran Kumar, D.Anil Kumar, “Training CNNs for 3-D Sign Language Recognition With Color Texture Coded Joint Angular Displacement Maps” IEEE Signal Processing Letters, Volume:25, Issue:5, 2018,Pg.:645-649.

[37] Puvvadi Aparna, Polurie Venkata Vijay Kishore, “Biometric-based efficient medical image watermarking in E-healthcare application”, IET Image Processing, Volume:13, Issue :03, 2019, Pg.:421-428.

Liver Tumor Segmentation using Superpixel based Fast Fuzzy C Means Clustering

Munipraveena Rela[1], Suryakari Nagaraja Rao[2], Patil Ramana Reddy[3]
Department of Electronics and Communication Engineering[1, 2, 3]
Jawaharlal Nehru Technological University Anatapur, Ananthapuramu, Andhra Pradesh, India[1]
G. Pulla Reddy Engineering College (Autonomous), Kurnool, Andhra Pradesh, India[2]
JNTUA College of Engineering (Autonomous), Ananthapuramu, Andhra Pradesh, India[3]

***Abstract*—In computer aided diagnosis of liver tumor detection, tumor segmentation from the CT image is an important step. The majority of methods are not able to give an integrated structure for finding fast and effective tumor segmentation. Hence segmentation of tumor is most difficult task in diagnosing. In this paper, CT abdominal image is segmented using Superpixel-based fast Fuzzy C Means clustering algorithm to decrease the time needed for computation and eradicate the manual interface. In this algorithm, a superpixel image with perfect contour can be obtain using a Multiscale morphological gradient reconstruction operation. Superpixel is pre-segmentation algorithm and is employed to obtain segmentation accuracy. FCM with modified object is used to obtain the color segmentation. This method is examined on 20 CT images gathered from liveratlas database, results shows that this approach is fast and accurate compared to most of segmentation algorithms. Statistical parameters which include accuracy, precision, sensitivity, specificity, dice, rfn and rfp are calculated for segmented image. The results shows that this algorithm gives high accuracy of 99.58% and improved rfn value of 8.34% compared with methods discussed in the literature.**

Keywords*—CT scan image; image segmentation; fuzzy c mean clustering; liver mask; superpixel image

I. Introduction

Liver cancer causes third most deaths in world-wide [1]. Computed tomography is commonly used modality for liver cancer diagnosis. The cancer is treated effectively providing it's detected in early stage. In order to diagnose liver tumors, such as hepatocellular carcinoma and metastatic carcinomas, computer aided diagnosis (CAD) is used. Importance of Segmentation algorithms are to separate cancerous portion which is essential for medical diagnosis of cancer [2]. Structure of liver changes with age, sex, body shape, hence the tumor detection is complicated using CAD system.

There are two sorts of segmentation algorithms that are supervised and unsupervised. Unsupervised does not depend on training data samples and labeling of data, these include GraphCut [3], watershed transform (WT) [4], fuzzy entropy [5], clustering [6], etc. Training data and labelling images are required for supervised segmentation, and these include neural networks [7-8]. In these methods, features are used for learning and to achieve segmentation.

In unsupervised segmentation, clustering is the common form of segmentation utilized for color and grayscale images [9]. In this method, Segmentation depends on Fuzzy c mean clustering (FCM) with minimized objective function.

II. Related work

Deepesh Edwin et al used fuzzy logic based thresholding and Shannon's entropy function for tumor segmentation [10]. Amita Das et al utilized adaptive thresholding and FCM for segmentation. The tumor is classified using Multi-Layer Perceptron (MLP) and C4.5 decision tree classifiers [11]. Ramin Ranjbarzadeh et al used the Kirsch filter for extracting organ edges, then calculated the concave and convex points, the mean shift is implemented to obtain uniform images, and eventually FCM is carried out to segment the liver [12].

Muthuswamy J suggested a method in which pre-processing is carried out using median filter and neutrosophic (NS) domain with FCM thresholding for segmentation, then liver contour is obtain using morphological operations, The classifier of support vector machines is used to categorize the tumor [13]. Munipraveena Rela et al used region growing algorithm for tumor segmentation. The tumor is categorized by the area and perimeter of the tumor as benign or malignant. [14].

Jing Zhao et al mentioned a technique to reduce image noise. Here, initially neutrosophic image is obtained, then FCM and particle swarm optimization is utilized to improve the global search [15]. Souhil Larbi Boulanouar et al proposed an algorithm based on the FCM and Modified Fuzzy Bat Algorithm (MFBA) to get better initial cluster centers [16]. Xiaofeng Zhang et al discussed a segmentation method using patch-weighted distance and FCM. Initially patch weighted distance is used to find the correlation between adjacent pixels, and then the influence of neighboring information in fuzzy algorithms is replaced with the pixel correlation, hence this method enhances segmentation[17].

FCM based segmentation algorithms discussed in the literature are sensitive to noise and requires more computation time. These methods require to select the cluster which has liver region, then liver portion is extracted using morphological operations.

Since Superpixel is used in computer vision to enrich the information extracted from an image and group pixels with similar visual properties. Also, a convenient image representation that is reason for computationally efficient. A superpixel picture can give better neighborhood spatial data

than a neighboring window of fixed size and shape, mean shift [18], Simple Linear Iterative Clustering (SLIC) [19] and WT [20], usually regarded as a pre-segmentation algorithm used to improve the segmentation results generated by clustering Algorithm.

Hence, the Superpixel based Fast fuzzy c mean clustering (SFFCM) method of segmentation is performed at low computational complexity, with high accuracy, and fully automatic.

A. Research Problem

The segmentation of Liver Tumor is important step in CAD system, and this is a difficult task due to following reasons:

- Liver region is overlapped with other organs and intensity difference between other organ and liver is very less.
- Only FCM does not segment the tumor accurately, FCM should be used with the pre segmentation algorithms.
- Some algorithms require prior information of region of interest.

Hence the SFFCM clustering is proposed for efficient segmentation, and can portion liver, tumor, and other organ effectively. In this method, a pre-segmentation algorithm called superpixel image is used along with the FCM for accurate segmentation. Hence, Multiscale morphological gradient recon-struction operation and WT (MMGR-WT) is utilized to acquire superpixel image. This algorithm is executed in less time compare to other superpixel algorithms. The performance of algorithm is not sensitive to parameters.

B. Contribution

Majority of the segmentation algorithms mentioned in the literature are noise sensitive and additionally also requires more computation time and human interference. This paper discusses the SFFCM clustering algorithm for tumor segmentation, and is fully automatic segmentation method and additionally shows that this technique offers high accuracy in comparison to the strategies discussed in the literature.

III. Research Method

In this paper, the tumor is segmented from CT liver image utilizing distinctive segmentation methods such as connected component labelling (CCL), K-means clustering (KMC), FCM clustering, SFFCM. Tumor segmentation using the SFFCM method is computed in less time and with less human interaction. In this method, The CT liver image is divided into various regions, then tumor region is extracted based on intensity value selected from the histogram of segmented image. The general block diagram representation of tumor segmentation is shown in Fig. 1.

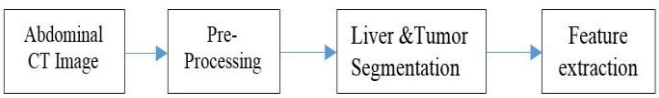


Fig. 1. Block Diagram of CT Liver Tumor Segmentation.

- Abdominal CT image

The CT image containing liver is obtained from liveratlas database. The SFFCM algorithm is implemented on 20 abdomen CT images. These images are converted to JPEG format. Usually CT liver images are low contrast, and liver is overlapped with other organ. So it is important to segment the liver and therefore tumor region.

- Pre-processing

In this step, the noise in the image will be eliminated by using filters such as wiener filter, median filter, and bilateral filters etc. contrast of the image is increased in this process.

- Liver and tumor segmentation

The segmentation of liver is very important step for tumor identification and classification. Segmentation means divide an image into its foreground and background. There are different techniques are available for segmentation. But these techniques don't give any fruitful result. In this paper, the SFFCM clustering method is applied for tumor segmentation. The result obtained using this methodology is compared with other strategies such as CCL, KMC, and FCM clustering.

- Feature Extraction

The ground truth images of tumor are utilized to compare with tumor regions obtained from the different segmentation methods and measured the parameters such as Accuracy, Sensitivity, Fmeasure, Precision, MCC, Dice, Jaccard, Specificity, rfp, rfn. These parameters are important to measure efficiency of tumor segmentation methods.

A. CCL

In CCL, pixel connectivity is used to group the pixels into components in the given input image. After grouping, a gray level is used to assign the pixel in the corresponding component. CCL examines the pixel by pixel so as to identify adjacent pixels with the same intensity values in the region of the image. CCL can be implemented on binary or gray level images and it uses 4- or 8-connectivity. Label connected components can be obtain using the keyword bwlabel in MATLAB.

Fig. 2 shows the block diagram representing the method to segment the tumor region. CT image is given as input. All connected components are identified using CCL. Liver region is selected from connected components, then this region is used as mask to extract only liver area. Again CCL is used to obtain connected components in the liver region, then extracted tumor mask. Tumor mask is used to extract the tumor region from original image.

B. KMC

KMC distribute the data among K number of clusters. If K=2, then two centers will be selected randomly. The data points are allocated to one of the cluster depending upon the distance between the data point and cluster center. Then cluster center will be updated by taking mean of the distance between data points and the center in the cluster.

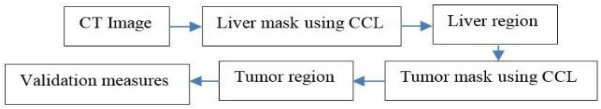


Fig. 2. The Tumor Segmentation from CT Image of Liver using CCL.

Fig. 3 shows tumor segmentation using KMC. The CT image is clustered using K=4, selected the cluster which has liver region. Liver is the largest connected component in that cluster and it very well may be extracted by utilizing CCL and also tumor region from liver.

C. Contrast Limited Adaptive Histogram Equalization (CLAHE)

Adaptive histogram equalization (AHE) causes noise amplification, whereas CLAHE is different from AHE. In CLAHE, The noise amplification can be reduced by clipping limit, i.e., the histogram is clipped at predefined value before calculating cumulative distributed function [21]. Block Size (BS) and Clip Limit (CL) are two main parameters in CLAHE. BS is important to divide the input image into non overlapping sections called blocks. If CL increases, the histogram becomes flatter so that it increases brightness of input image. When BS increases, the dynamic range of pixel intensity becomes larger. In CLAHE, histogram equalization is applied to each non overlapping block separately. The CL is employed to clip the original histogram and clipped intensity values are redistributed. The redistributed histogram is dissimilar from original histogram, because CL is utilized to restrict the pixel intensity [22]. Generally Medical images are low contrast, CLAHE is applied for enhancing medical images [23-25].

Syntax: J = adapthisteq (I, param1, val1, param2, val2...), J is contrast enhanced grayscale image of I. In CLAHE, the image is divided into tiles. The 'Distribution' parameter is utilized to specify the histogram shape, so that histogram of output region should match with it to obtain enhanced contrast of each tile. The artificially boundaries are induced while combining neighboring tiles, the tiles are combined by using bilinear interpolation to eliminate these boundaries. Noise amplification in homogeneous region can be avoided by contrast limiting.

Fig. 4 shows the tumor segmentation using KMC and CLAHE. Here CT image is enhanced by using CLAHE. This gives better results compared to KMC method of segmentation.

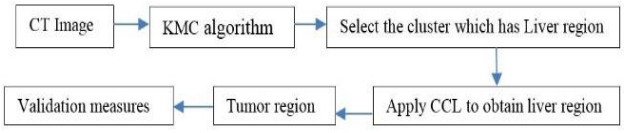


Fig. 3. Tumor Segmentation using KMC.

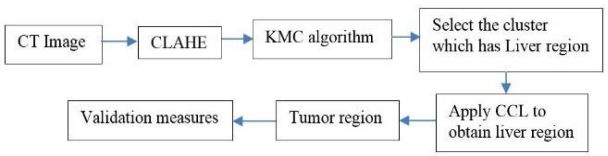


Fig. 4. Tumor Segmentation using KMC and CLAHE.

D. FCM Clustering [9]

Clustering analysis refers to subdivide data set X into c subsets which are called clusters, these clusters are dissimilar, and nonempty. These clusters on combining produce data set X. This is called nonfuzzy clustering. According to Zadeh's, membership function is used to denote the similarity among data point and cluster, it has some value called membership which ranges between [0,1]. Each data point has some similarity, called membership in every cluster. If membership is one, then there is high similarity between data point and the cluster. If membership is near to zero, then there is very low similarity between data point and the cluster. A fuzzy c-partition of X describes the membership function to denote the membership of a data point with all clusters. The sum of memberships for any data point must be one [26, 27].

Let $X = \{x_1, x_2, ..., x_n\}$ be a sample of N observations in 2-dimensional Euclidean space; x_k is the $k-th$ feature vector; x_{kj} the $j-th$ feature of x_k. If c is an integer, $2 \le c \le n$, A conventional c-partition of X is a c-tuple $(X_1, X_2, ..., X_c)$, subsets of X that satisfies three conditions as shown in (1), (2), and (3):

$$X_i \neq \phi; 1 \le i \le c; \tag{1}$$

$$X_i \bigcap X_j = \phi; i \neq j; \tag{2}$$

$$\bigcup_{i=1}^{c} X_i = X \tag{3}$$

The FCM algorithms is represented by modifying conditions in matrix-form as shown in Eq. (1). Let U be a real $c \times N$ matrix, $U = [u_{ik}]$. U is the matrix representation of the partition $\{X_i\}$ in (1), it is denoted as in (4), (5), and (6).

$$u_i(x_k) = u_{ik} = \begin{cases} 1; & x_k \in X_i \\ 0; otherwise \end{cases} \tag{4}$$

$$\sum_{i=1}^{N} u_{ik} > 0 \quad \text{for all } i; \tag{5}$$

$$\sum_{i=1}^{N} u_{ik} = 1 \quad \text{for all } k \tag{6}$$

The generalized least-squared errors method is proposed for detecting fuzzy c-partitions in X. The least squared errors is given in (7).

$$J_m(U, v) = \sum_{k=1}^{N} \sum_{i=1}^{c} (u_{ik})^m \left\| x_k - v_i \right\|_A^2 \tag{7}$$

Where $X = \{x_1, x_2, ..., x_n\} \subset R^n$ = the data, c =the number of clusters in Y ; $2 \le c < n$, $m =$ weighting exponent; $1 \le m < \infty$, U = fuzzy c-partition of X ; $v = (v_1, v_2, ...v_c)$ = vectors of centers; $v_i = (v_{i1}, v_{i2}, ..., v_{in})$ =centers of cluster i ,

A =positive-definite $(n \times n)$ weight matrix, $\| \|_A$ =induced A -norm on R^n , d_{ik}^2 is the squared distance between x_k and v_i as given in (8).

$$d_{ik}^2 = \|x_k - v_i\|_A^2 = (x_k - v_i)^T A (x_k - v_i) \tag{8}$$

$(U, \hat{v})$ Pairs are used to define optimal fuzzy clustering of X that locally minimize J_m . For $m > 1$, if $x_k \neq v_j$ for all j and k , $(U, \hat{v})$ may be locally optimal for J_m only if v_i and u_{ik} are as shown in (9) and (10).

$$v_i = \sum_{k=1}^{N} (u_{ik})^m y_k \Big/ \sum_{k=1}^{N} (u_{ik})^m ; 1 \le i \le c; \tag{9}$$

$$u_{ik} = \left(\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{2/(m-1)} \right)^{-1} ; 1 \le k \le N; 1 \le i \le c; \tag{10}$$

Fig. 5 represents tumor segmenatation using FCM. Here input image is grouped into four cluster. A cluster which has liver region is selected, then CCL is implemented to that cluster to attain liver and tumor regions.

E. SFFCM Clustering

Majority of image segmentation algorithms based on improved FCM clustering require more execution time and incapable to give desired results because of two reasons [28]. The primary reason is that the immoderate computational complexity within a local neighboring window is due to the repeated distance calculation between clustering centers and pixels. Second reason is that these techniques aren't able to offer preferred results due to constant size and shape of neighboring window. Hence, A SFFCM clustering algorithm requires less execution time and accurate for color image segmentation [29].

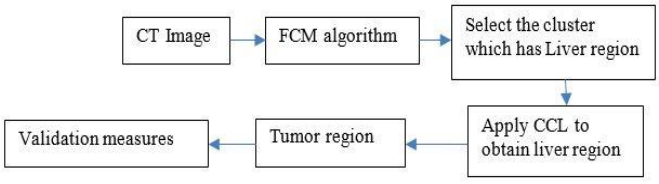


Fig. 5. Tumor Segmentation using FCM.

Local spatial information in a superpixel image is obtained by using MMGR operation. The adaptive local spatial neighborhoods that's helpful for segmentation is provided by superpixel image. The histogram of superpixel picture can be obtained without any problem. Lastly, Segmentation result can be obtained by implementing FCM with histogram parameters on the superpixel image. Improved segmentation results can be obtain by combining MMGR-WT and FCM, since MMGR-WT is utilize to accomplish the local feature of an image whereas global features can be obtained using FCM. A SFFCM algorithm is developed by means of considering adaptive local spatial information into the objective function of FCM [30].

SFFCM's objective function is defined using the MMGR-WT based superpixel image as shown in (11).

$$J_m = \sum_{l=1}^{q} \sum_{k=1}^{c} S_l u_{kl}^m \left\| \left(\frac{1}{S_l} \sum_{p \in R_l} x_p \right) - v_k \right\|^2 \tag{11}$$

where l is the color level, $1 \le l \le q$, q is the number of regions of the superpixel image, $l, q \in N^+$, S_l is the number of pixels in the l th region R_l , and x_p is the color pixel within the l th region of the superpixel image obtained by MMGR-WT. In this process, to substitute of color pixel in the original picture, the mean value of the color pixels within the corresponding region of the superpixel picture is used, so the number of regions within the superpixel image is same as the number of color levels. Thus, the complexity of this algorithm is efficiently decreased due to $l \ll N$.

The cluster center and fuzzy c partition is obtained using following (12) and (13):

$$v_k = \frac{\sum_{l=1}^{q} u_{kl}^m \sum_{p \in R_l} x_p}{\sum_{l=1}^{q} S_l u_{kl}^m} \tag{12}$$

$$u_{kl} = \frac{\left\| \left(\frac{1}{S_l} \sum_{p \in R_l} x_p \right) - v_k \right\|^{-2/(m-1)}}{\sum_{j=1}^{c} \left\| \left(\frac{1}{S_l} \sum_{p \in R_l} x_p \right) - v_j \right\|^{-2/(m-1)}} \tag{13}$$

The algorithm of SFFCM is shown in Fig. 6.

Fig. 7 shows tumor segmentation using the SFFCM. CT image is converted to Superpixel image using MMGR-WT, then FCM with proposed objective function is utilized to acquire segmented image. Histogram of segmented image is obtained to select the gray level of tumor region. Finally, gray level thresholding is implemented to obtain tumor region.

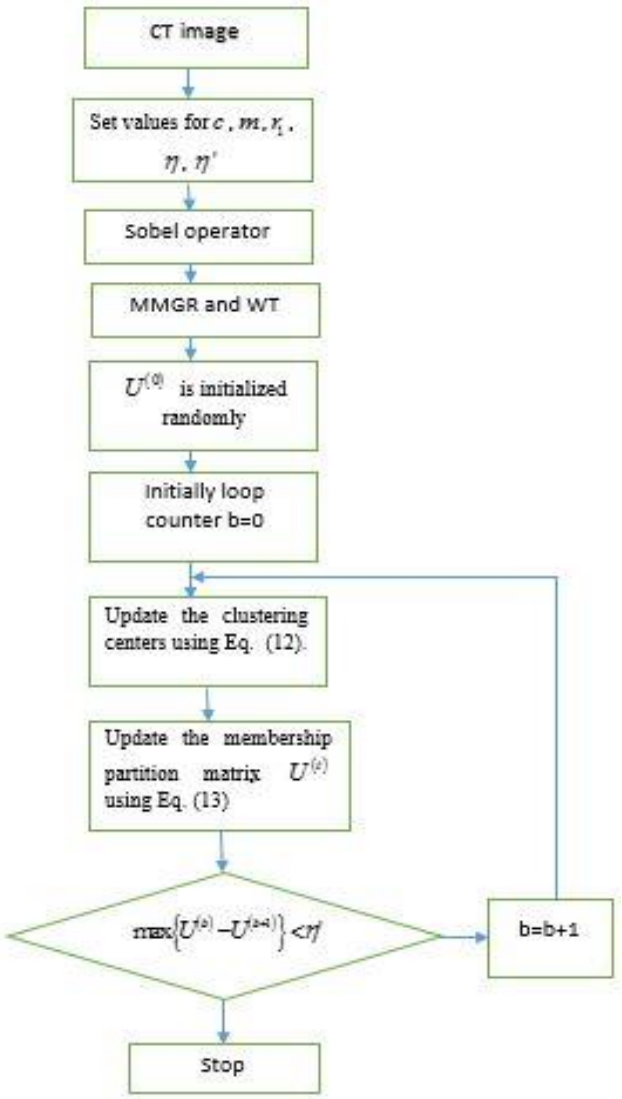


Fig. 6. Algorithm of SFFMC.

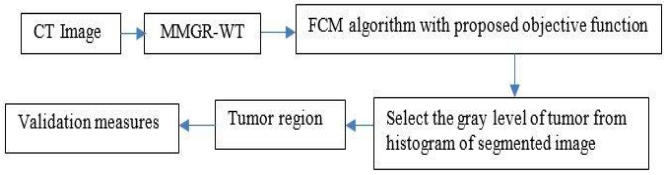


Fig. 7. Tumor Segmentation using the SFFCM.

IV. Results and Discussions

Here 20 CT abdominal images from liveratlas database are considered for analysis. The proposed work carried out using Intel CORE i5 processor and MATLAB R2015a version. In this section, the tumor segmentation obtained by applying different segmentation methods are discussed.

A. Validation Measures

The following parameters are used to compare ground truth and segmented image. These parameters are defined by (14) – (22).

Accuracy (14) is the measure of correctly predicted observations with reference to the total observations.

$$Accuracy = \frac{TP + TN}{FN + FP + TP + TN}; \quad (14)$$

Sensitivity as given in (15), measures which frequently a diagnostic test is positive, given that the person has the disease.

$$Sensitivity = \frac{TP}{TP + FN}; \quad (15)$$

Precision is the quantitative relation of properly expected positive observations to the full expected positive observations, as shown in (16).

$$\Pr ecision = \frac{TP}{TP + FP}; \quad (16)$$

Matthews Correlation Coefficient (17).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{\left((TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)\right)}}; \quad (17)$$

Dice Coefficient (18) is the proportion of overlap between the two sets, which is a number between 0 and 1.

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN}; \quad (18)$$

The Jaccard Index (19) is utilized in understanding the similarities between sample sets.

$$Jaccard = \frac{Dice}{2 - Dice}; \quad (19)$$

Specificity (20) is the chance that a diagnostic check is negative, provided that the individual doesn't has the illness.

$$Specificity = \frac{TN}{TN + FP}; \quad (20)$$

The false positive rate (rfp) (21) is defined as the likelihood of falsely rejecting the null hypothesis

$$rfp = \frac{FP}{FP + TN}; \quad (21)$$

The false negative rate (rfn) (22) is the proportion of the individuals with a known positive condition for which the test outcome is negative.

$$rfn = \frac{FN}{FN + TP}; \quad (22)$$

TP - true positive, FP - false positive, TN - true negative, FN - false negative

B. Segmented Images

The input image is color image, hence it is converted to grayscale image. Grayscale image is enhanced by using histogram equalization. Thresholding is utilized to get binary image. bwlabel is employed to label the connected components in binary image. Liver mark is extracted from connected components. Using liver mark, liver is obtained. From liver region, Tumor region is extracted using morphological operations.

C. Parameter Setting

In CCL, histogram equalization is used in preprocessing stage, threshold with level=0.67 is used to obtain binary image, total number of connected components are 116, the largest connected component is 11 which is liver, then morphological operations are used to extract the liver region, thresholding with level 0.5 is used to obtain tumor region.

In FCM, median filter is used as preprocessing. Divided the image into four clusters, selected the cluster which has liver region, then used morphological operations to extract the tumor.

In SFFCM, No. of clusters=8, structuring element for MMGR=3, exponent for the partition matrix U=2, max. Number of iteration=50, min. amount of improvement=1e-5.

D. Results

Fig. 8(a) shows the output of KMC. Fig. 8(b) shows clustering using FCM. From these clustering, the cluster which contains liver region is selected, that is in Fig. 8(a), cluster3 is selected, whereas in Fig. 8(b) cluster4 is selected. Then CCL is applied to this cluster to obtain largest connected component, which is nothing but liver. Then the tumor is identified using morphological operation.

Fig. 9(a), 9(b), 9(c), 9(d), and 9(e) show tumor segmentation using CCL, KMC, and CLAHE enhanced KMC, FCM and SFFCM. Fig. 10 shows the tumor segmentation using SFFCM clustering for different CT images. The SFFCM gives color segmented image. From the color segmented image, pixel value of tumor region is utilized in gray level slicing to get the tumor mask. Then tumor region is extracted by multiplying tumor mask with original image.

Table I lists the comparison of statistical parameters measured for tumor segmented images of CT1, CT2 and CT3 with ground truth images. Fig. 11 shows the plot of statistical parameters for tumor segmented images. Table II gives the comparison of rfp, rfn values with reference to different methods of segmentation. It is observed that segmentation using SFFCM has improved rfn. SFFCM is used to acquire segmented image directly, it is additionally less tedious and fully automatic. The SFFCM method of segmentation gives accuracy of 99.58%, sensitivity of 87.77%, Precision of 96.37%, MCC of 91.61%, Specificity of 99.93%, Jaccard of 84.64%, and Dice of 91.54%.

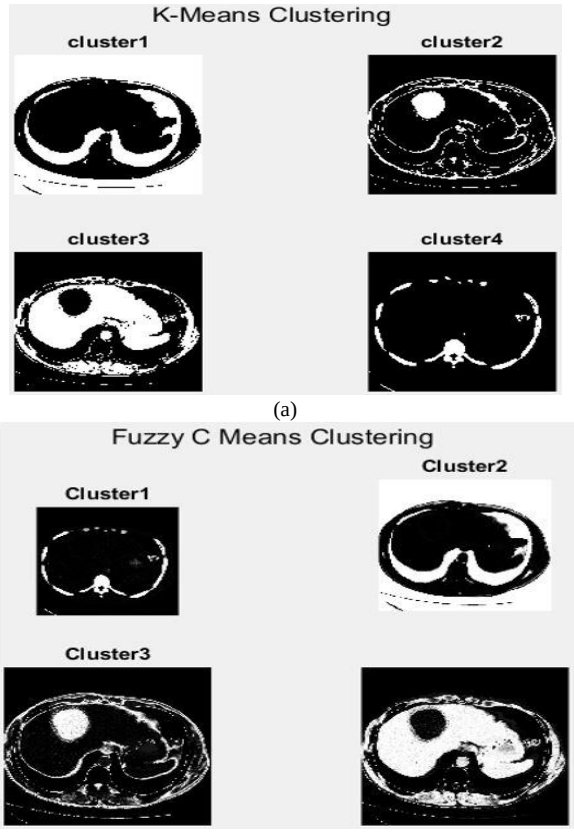


Fig. 8. Results of Clustering using (a) KMC (b) FCM.

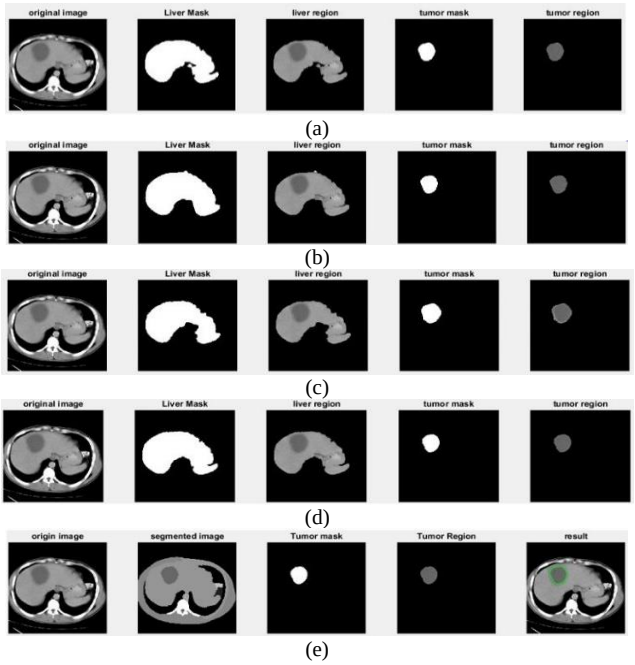
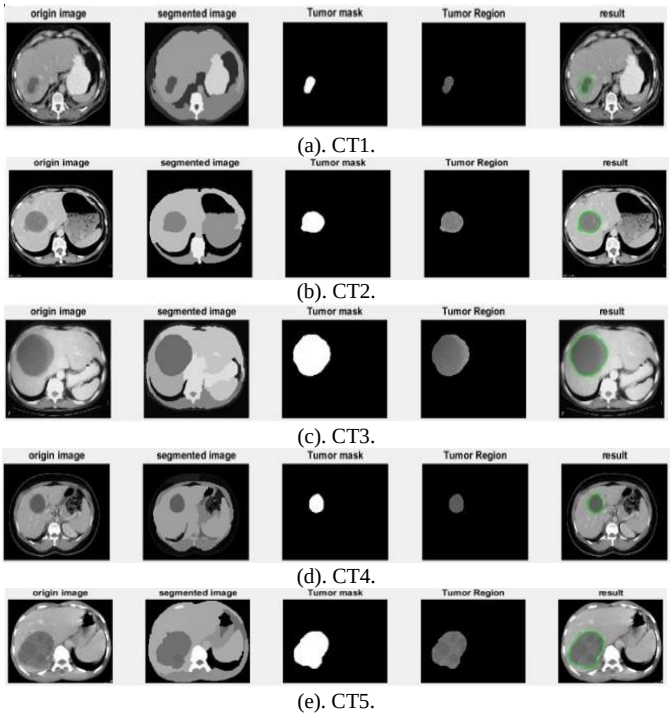


Fig. 9. Tumor Segmentation (a).CCL, (b).KMC, (c). CLAHE and KMC, (d). FCM, (e).SFFCM.



(a). CT1.

(b). CT2.

(c). CT3.

(d). CT4.

(e). CT5.

Fig. 10. Tumor Segmentation using SFFCM.

TABLE I. STATISTICAL PARAMETERS FOR IMAGE SEGMENTATION USING SFFCM

INPUT	CT1	CT2	CT3	CT4	CT5	Avg
Accu	0.998	0.992	0.997	0.999	0.993	0.9958
Sens	0.918	0.765	0.9504	0.97764	0.97217	0.917
Prec	0.920	1	0.9713	0.9898	0.97096	0.970
MCC	0.918	0.8713	0.9593	0.98334	0.96791	0.940
Spec	0.999	1	0.9989	0.99978	0.99626	0.9988
Jaccard	0.850	0.765	0.9244	0.96789	0.9447	0.8904
Dice	0.919	0.867	0.9607	0.98368	0.97157	0.9403
rfp	0.08	0	0.0281	0.0101	0.029	0.0294
rfn	0.082	0.235	0.0496	0.0224	0.0278	0.0834

Comparison of accuracy for SFFCM against the methods in literature is listed in Table III. It is shown that, compared with the methods in the literature, the SFFCM provides high accuracy.

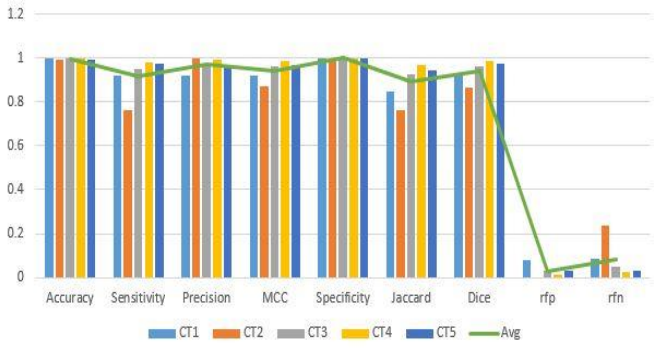


Fig. 11. Statistical Parameter of SFFCM for different Input Images

TABLE II. COMPARISON OF SEGMENTATION METHODS USING *rfp* AND *rfn*

Segmentation method	rfp	rfn
CCL	0	0.19596
KMC	0	0.21861
CLAHE-KMC	0.47912	0.05777
HPF-KMC	0	0.22631
FCM	0.05377	0.06594
SFFCM	0	0.0834

TABLE III. COMPARISON ON ACCURACY FOR SFFCM AGAINST THE METHODS IN LITERATURE

S.No.	Author	Accuracy
1	Grzegorz Chlebus[32]	90%
2	Dong Yang[31]	95%
3	Sultan Almotairi[28]	98.8%
4	Lamia N. Mahdy[33]	96.86%
5	SFFCM	99.58%

Comparison of computational time of various methods used to obtain superpixel image in shown in Table IV. The results shows that MMGR-WT requires less execution time compare to the methods discussed.

TABLE IV. COMPARISON OF COMPUTATIONAL TIME(IN SECONDS) OF METHODS USED TO OBTAIN SUPERPIXEL IMAGES

Algorithms	Execution time(in seconds)
SLIC	3.86
mean-shift1	1.02
mean-shift2	2.66
MMGR-WT	0.32

V. CONCLUSION

In this paper, a SFFCM clustering algorithm is employed for liver tumor segmentation. In order to obtain a good superpixel image, the MMGR operation is presented and the color histogram is integrated into the objective function to achieve fast image segmentation. The experimental findings show that the SFFCM is superior to state-of-the-art clustering algorithms because it offers the best results for segmentation and has the shortest runtime.

The SFFCM-based Liver tumor segmentation is automatic. Therefore, relative to other unsupervised segmentation approaches, human intervention is less. This method has achieved accuracy of 99.5%, Precision of 96.37%, Specificity of 99.93% for tumor segmentation. The results shows that the SFFCM method of segmentation gives high accuracy and improved compared to the methods in literature.

For color image segmentation, our algorithm is very fast because pre-segmentation method called superpixel image is used before segmentation, but it has limitations in realistic applications, similar to other k-means clustering algorithms,

because the number of clusters must be set before-hand. We will combine frequency domain approach and deep learning approach to explore fast clustering algorithms in future work, which automatically estimate the number of clusters.

REFERENCES

[1] Jacques Ferlay, Hai-Rim Shin, Freddie Bray, David Forman, Colin Mathers and Donald Maxwell Parkin, 'Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008', International Journal of Cancer, 127, pp. 2893–2917, 2010.

[2] Munipraveena Rela, S. Nagaraja Rao,and P. Ramana Reddy, 'Efficient Image Enhancement Techniques Applied on Medical Imaging-A State-of- The Art Survey', International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-7, Issue-6S4, April 2019.

[3] Jizhou Ma, Shuai Li, Hong Qin, Aimin Hao, 'Unsupervised Multi-Class Co-Segmentation via Joint-Cut Over L1 -Manifold Hyper-Graph of Discriminative Image Regions', IEEE Transactions on Image Processing, vol. 26, no. 3, pp. 1216-1230, Mar. 2017.

[4] Min Bai, Raquel Urtasun, 'Deep Watershed Transform for Instance Segmentation', 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, July 2017, pp. 2858-2866.

[5] ShibaiYin, Yiming Qian, Minglun Gong, 'Unsupervised hierarchical image segmentation through fuzzy entropy maximization', Pattern Recognition, Vol. 68, pp. 245-259, Aug. 2017.

[6] Basel Abu - Jamous Rui Fa Asoke K. Nandi, 'Integrative Cluster Analysis in Bioinformatics', Copyright © 2015 John Wiley & Sons, Ltd, 20 March 2015, Print ISBN:9781118906538 ,Online ISBN:9781118906545, |DOI:10.1002/9781118906545.

[7] Alex Krizhevsky , Ilya Sutskever , Geoffrey E. Hinton, 'Imagenet classification with deep convolutional neural networks', Advances in Neural Information Processing Systems (NIPS), New York, NY, 2012, pp. 1097-1105.

[8] Evan Shelhamer, Jonathan Long, Trevor Darrell, 'Fully convolutional networks for semantic segmentation', IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 640-651, Apr. 2017.

[9] Marcelo Pereyra, Steve McLaughlin, 'Fast Unsupervised Bayesian Image Segmentation With Adaptive Spatial Regularization', IEEE Transactions on Image Processing, Vol. 26, Issue: 6 , pp. 2577-2587, June 2017.

[10] Deepesh Edwin and S. Hariharan,' Liver and tumour segmentation from abdominal CT images using adaptive threshold method', Int. J. Biomedical Engineering and Technology, Vol. 21, No. 2, 2016, pp. 190-204.

[11] Amita Das, Priti Das, S. S. Panda, and Sukanta Sabut, 'Detection of Liver Cancer Using Modified Fuzzy Clustering and Decision Tree Classifier in CT Images', Pattern Recognition and Image Analysis, 2019, Vol. 29, No. 2, pp. 201–211.

[12] Ramin Ranjbarzadeh, and Soroush Baseri Saadi, 'Automated liver and tumor segmentation based on concave and convex points using fuzzy c-means and mean shift clustering', Measurement, Volume 150, January 2020.

[13] Muthuswamy Jayanthi, 'Extraction and Classification of Liver Abnormality Based on Neutrosophic and SVM Classifier', Advances in Intelligent Systems and Computing, vol 713. Springer, Singapore, 2019.

[14] Munipraveena Rela, S Nagaraja Rao, and P Ramana Reddy, 'Extraction and Classification of tumor in CT Liver Image', TEST Engineering and Management, ISSN: 0193-4120, pp. 8747 – 8753, January-February 2020.

[15] Jing Zhao, Xiaoli Wang and Ming Li, 'A Novel Neutrosophic Image Segmentation Based on Improved Fuzzy C-Means Algorithm (NIS-IFCM)', International Journal of Pattern Recognition and Artificial Intelligence, Vol. 34, No. 5 (2020).

[16] Souhil Larbi Boulanouar and Chaabane Lamiche, 'A New Hybrid Image Segmentation Method Based on Fuzzy C-Mean and Modified Bat Algorithm', International Journal of Computing and Digital Systems, ISSN (2210-142X), No.4 (July-2020).

[17] Weiwei Wu, Shuicai Wu, Zhuhuang Zhou, Rui Zhang, and Yanhua Zhang, '3D Liver Tumor Segmentation in CT Images Using Improved Fuzzy C-Means and Graph Cuts', Hindawi BioMed Research International, Vol. 2017, pp. 1-12, Sep. 2017.

[18] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol.24, no. 5, pp. 603-619, May 2002.

[19] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods" IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 11, pp. 2274-2282, Nov. 2012.

[20] Z. Hu, Q. Zou and Q. Li, "Watershed superpixel," in Proc. IEEE Int. Conf. Image Process. (ICIP), Quebec City, QC, 2015, pp. 349-353.

[21] Xiaofeng Zhang, Muwei Jian, Yujuan Sun, Hua Wang & Caiming Zhang, 'Improving image segmentation based on patch-weighted distance and fuzzy clustering', Multimedia Tools and Applications, Vol. 79, pp. 633–657, 2020.

[22] Pisano, E.; Zong, S.;Hemminger, B.; DeLuca,M.; Johnston, R; Muller, K.; Braeuning, M.; Pizer, S., 'Contrast Limited Adaptive Histogram Equalization Image Processing to Improve the Detection of Simulated Spiculations in Dense Mammograms', Journal of Digital Imaging, 11, pp. 193–200, 1998.

[23] Jinxiang Ma, Xinnan Fan, Simon X. Yang, Xuewu Zhang and Xifang Zhu," Contrast Limited Adaptive Histogram Equalization Based Fusion for Underwater Image Enhancement", International Journal of Pattern Recognition and Artificial IntelligenceVol. 32, No. 07, 1854018 (2018).

[24] Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla, 'SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, Issue: 12, pp. 2481 – 2495, Dec. 2017.

[25] Yue Zhang, Jiong Wu, Benxiang Jiang, Dongcen Ji, Yifan Chen, Ed X. Wu, and Xiaoying Tang, 'Deep Learning and Unsupervised Fuzzy C-Means Based Level-Set Segmentation for Liver Tumor', 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, Iowa, USA , pp. 1193-1196, April 3-7, 2020.

[26] James C. Bezdek, Robert Ehrlich, and William Full, 'FCM: THE FUZZY c-MEANS CLUSTERING ALGORITHM', Computers & Geosciences, Vol. 10, No. 2-3, pp. 191-203, 1984.

[27] Shan Zeng, Xiuying Wang, Hui Cui, Chaojie Zheng, David Feng, 'A Unified Collaborative Multikernel Fuzzy Clustering for Multiview Data', IEEE Transactions on Fuzzy Systems, Vol. 26, No. 3, pp. 1671-1687, Jun. 2018.

[28] Sultan Almotairi, Ghada Kareem, Mohamed Aouf, Badr Almutairi and Mohammed A.-M. Salem, 'Liver Tumor Segmentation in CT Scans Using Modified SegNet', Sensors, 20, 1516, 2020.

[29] Maoguo Gong, Hao Li, Xiang Zhang, Qiunan Zhao, Bin Wang, 'Nonparametric Statistical Active Contour Based on Inclusion Degree of Fuzzy Sets', IEEE Transactions on Fuzzy Systems, Vol. 24 , Issue: 5 , pp. 1176-1192, Oct. 2016.

[30] Tao Lei, Xiaohong Jia, Yanning Zhang, Shigang Liu, Hongying Meng, Asoke K. Nandi, 'Superpixel-Based Fast Fuzzy C-Means Clustering for Color Image Segmentation', IEEE Transactions on Fuzzy Systems, Vol. 27, No. 9, pp. 1753-1766, Sept. 2019.

[31] Dong Yang, Daguang Xu, S. Kevin Zhou, Bogdan Georgescu, Mingqing Chen, Sasa Grbic, Dimitris Metaxas and Dorin Comaniciu, 'Automatic Liver Segmentation Using an Adversarial Image-to-Image Network', arXiv, cs.CV, eprint:1707.08037, 2017.

[32] Grzegorz Chlebus, Hans Meine, Jan Hendrik Moltz, Andrea Schenk, 'Neural Network-Based Automatic Liver Tumor Segmentation With Random Forest-Based Candidate Filtering', arXiv, cs.CV, eprint. 1706.00842, 2017.

[33] Lamia N. Mahdy, Kadry A. Ezzat, Mohamed Torad, Aboul E. Hassanien, 'Automatic segmentation system for liver tumors based on the multilevel thresholding and electromagnetism optimization algorithm', International journal of imaging systems and technology, DOI: 10.1002/ima.22432, Mar 2020.

RHEM: A Robust Hybrid Ensemble Model for Students' Performance Assessment on Cloud Computing Course

Sapiah Sakri[1], Ala Saleh Alluhaidan[2]
Department of Information Systems, College of Computer and Information Science
Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia

***Abstract*—Creating tools, such as a prediction model to assist students in a traditional or virtual setting, is an essential activity in today's educational climate. The early stage towards incorporating these predictive models using techniques of machine learning focused on predicting the achievement of students in terms of the grades obtained. The research aim is to propose a robust hybrid ensemble model (RHEM) that can warn at-risks students (on Cloud Computing course) of their likely outcomes at the early semester assessment. We hybridised four renowned single algorithms – Naïve Bayes, Multilayer Perceptron, k-Nearest Neighbours, and Decision Table – with four well-established ensemble algorithms – Bagging, RandomSubSpace, MultiClassClassifier, and Rotation Forest – which produced 16 new hybrid ensemble classifier models. Hence, we have thoroughly and rigorously built, trained, and tested 24 models all together. The experiment concluded that the Rotation Forest + MultiLayer Perceptron model was the best performing model based on the model evaluation in terms of Accuracy (91.70%), Precision (86.1%), F-Score rate (87.3%), and Receiver Operating Characteristics Area detection (98.6%). Our research will help students identify their likely final grades in terms of whether they are excellent, very good, good, pass, or fail, and, thus, transform their academic conduct to achieve higher grades in the final exam accordingly.**

Keywords*—Academic performance; classification algorithms; cloud computing course; ensemble algorithms; hybrid ensemble classifier model; student academic performance tracking

I. Introduction

Innumerable data are generated and gathered in numerous fields. The big data created need to be collected, organized, and analysed in order to extract useful information. In order to obtain valuable information, real-world environments and industries need to analyse vast quantities of generated data. To do so, Data Mining (DM) techniques are used to create a model that analyses the given dataset and identifies useful trends in the results. DM includes numerical data analysis techniques and the discovery of useful knowledge. One of the most required procedures in big data and data mining is prediction, which has been utilized in different domains to increase efficiency and reduce costs. This usage of algorithms in education is still in progress. To explain, the education provided at the university level is usually connected to the economy and development of a country. However, the quality and output of education at this level depends on the kind of students admitted and whether they are able to complete their studies. The prediction of student academic performance helps in identifying weak students who will struggle with their studies. Science and IT majors are among the hardest at college level [1],[2]. Therefore, the management of computer and IT related institutions take essential steps to detect and correct the way for weak students. Many prediction and data mining algorithms have been used, such as clustering, classification, and association rule techniques, to extract knowledge from student datasets [3],[4]. This paper explores the effect of certain factors on student performance in advanced IT courses, such as cloud computing. Parameters, such as business course, maths course grade, science course grade, and core IT course grade can provide an indication of future students' performance in higher advanced courses. The current approaches have failed to analyse and monitor the progress of the student achievements [5]. Inappropriate methods or investigation procedures can also contribute to failure. This paper attempts to predict the educational performance of students based on motivational and academic factors. It introduces a hybrid prediction framework for measuring student performance in advanced computing courses, such as cloud infrastructure and services.

This paper is organized as follows. Section 2 explains the importance of this research. Section 3 presents the related work of the research. Section 4 elaborate the research approach. Section 5 describes the research preliminaries. Section 6 reports the experiment results. Section 7 discusses the findings in Section 6. Finally, Section 6 draws the conclusion and identifies future work.

II. Research Motivation

Predicting student performance at early stage of the semester would benefit both the university and students, particularly those students in their final year. For the university, high performing students would reflect the university's high quality of education system. For the students, knowing their level of performance at the early stage of the semester would avoid any decreased of grade in any courses which would have detrimental effect on their cumulative grade point average. Hence, we were motivated to propose a robust hybrid ensemble model (RHEM) that harnessed on machine learning algorithms which can predict at-risks students (on Cloud Computing course) of their likely outcomes at the early stage of the semester assessment. The prediction of exam performance through the context of student progress during

the course can increase efficiency and reduce the possibility of failing by offering pertinent advice and taking precautions. The marks obtained by a student in the examinations throughout the course duration can indicate the final exam results. Therefore, it becomes essential to predict whether the student will achieve an excellent, very good, good, pass or fail grade in the course. If the prediction indicates there is a high probability of the student failing the first exam prior to the final exam, then extra effort can be made to pass the exam.

III. Related Work

Machine Learning Algorithms are a group of useful tools that are used to create predictive models of student performance. Assessing the success prediction of the students is a very complex problem and uses different algorithms for this purpose. A systematic literature review to identify and collect the beneficial features for predicting student performance was discussed in [6], as well as the importance of Feature Selection (FS) to eliminate unrelated data that can produce a 10% difference in the prediction accuracy. Filter feature selection algorithms and classification algorithms were examined in [3]. The review shows that a variety of techniques have been used, but that there is no unified method that can be used for prediction in all cases. Specifically, the review uncovered a lack of quality and that there is a real need for more detailed reporting of the methods and results [6]. Referring to a closer topic to this research, student performance prediction by participating in an online discussion was mentioned in [2]. The sample was large as it comprised 76 second-year university students studying a Computer Hardware course. The study design was oriented to answer whether student performance prediction is possible and to compare different algorithms and features using classification and pre-processing techniques. The k-Nearest Neighbour algorithm accurately predicted unsuccessful students (89%). Moreover, students who were unsuccessful at the end of term could be predicted in the first 3 weeks with 74% accuracy.

The data collected by institutions or learning management systems are used in the sense of learning analytics to forecast student performance and recognize important factors that may contribute to the successful completion of a course. As we are interested in estimating student results, we review the related research work in this field. Although different techniques have been implemented in terms of prediction within the education field, it is still possible to improve the current approaches and provide more accurate results in terms of the context in which it is implemented. In the following, we review the current approaches that have been established. Educational Data Mining (EDM) and Learning Analytics (LA) to reveal knowledge from educational data were used to predict student success using data from the various universities in Pakistan [7]. "Learning analytics, discriminative and generative classification models are used to determine whether or not a student will complete his or her degree" [7]. Outcomes reveal better accuracy due to the reference of family expenditure, such as a natural gas, electricity, telephone, water, and accommodation, and students' personal data, such as gender, marital status, and employment, etc. To enhance engineering students' performance, a study by [1] identified the factors that can affect student success in this tough major. The study focused on the use of J48 and REP Tree algorithms to elicit the type of relationship between social parameters and student performance and predicting students' performance in their third semester. Analysis revealed that parents' education influences student performance and that previous semester grades greatly indicate the performance in the third semester. This finding helps in the early prediction of weak students to take the necessary decisions for improving students' performance. In terms of the algorithms used, J48 was more accurate than the REP Tree algorithm. Similarly, within science colleges, the author in [8] examined how the linking of prior knowledge and attitude for first year undergraduate chemistry students can affect their chemistry exam performance. Statistics showed that there are significant differences between the mean scores of students who have prior knowledge in chemistry and those who have not. Analysing the correlation and regression showed that previous knowledge affects the success of examinations. Two predictive models were suggested based on the regression analysis. In a similar vein, the research of [9] focused on how proficiency in certain courses can give an insight into student performance in programming courses. This is an IT concentrated research. The results of courses, such as introductory to physics and maths, can indicate performance in programming courses. Methods, such as Artificial Neural Network data mining, were used for prediction. The findings showed that having a background knowledge of mathematics and physics is vital for proficiency in programming. In looking for the most influential factors in student academic performance prediction, the authors in [10] aimed to present a predictive model for computer science students' study duration based on grades in the first two semesters. Naïve Bayes, decision tree, and Support Vector Machine (SVM) were used. The findings showed no significant difference between Naïve Bayes and decision tree in terms of efficiency, while SVM had the lowest performance. The influencing factors were grades, general subjects' grades, gender, and major subjects' grades.

A new method for prediction using Multi-Input Multi-Output, which relies on the Multi Adaptive Neuro-Fuzzy Inference System with Representative Sets, was introduced in [11]. To explain, authors used both global and a local training set with random parameters in the former, and premise and consequent parameters in the latter. Once the parameters have been refined, for the testing set, Fuzzy k-Nearest Neighbour is used to find which group it belongs to. This MANFIS-S model is validated against ANFIS, MANFIS, OneR, and Random Tree and is found to be more accurate. A dataset was collected from VNU University of Science, Vietnam, and three educational datasets were taken from KDD Cup. Another attempt using the Fuzzy Probabilistic Neural Network was mentioned in [12]. The Probabilistic Neural Network is a 4-layer, feed-forward, which includes an algorithm used for classification and mapping. It is based on Bayes' decision strategy and non-parametric kernel-based estimators of probability density functions. The experiments and results revealed that FPNN takes less time to be trained and the results are more accurate (average of 98.56%). The output consists of a class of three values (Good, Average, and Poor).

MATLAB was used to analyse 760 samples of the training dataset with over 18 factors as inputs (merit, interest, family background, class and study behaviour, interest and belief in learning). Various techniques and features have been designed to predict academic success from the literature reviewed; however, there is still a shortage of work predicting the achievement of higher courses in computer education. Therefore, this study aims to fill the gap by focusing on an advanced level course (highly required from Computer major students as it is the trend and the new approach of hosting and managing databases in the job market). The hybrid algorithm was designed to produce more accurate results.

IV. Methodology

The proposed research approach for this study is as shown in Fig. 1. Four phases are involved and four major experiments will be conducted.

A. Phase 1 – Data Pre-Processing

In phase 1, raw data was pre-processed by performing normalization, replacing missing values, and transforming the raw data into a new clean dataset appropriate for the experiment's requirements. The dataset was split 70% for training and 30% for testing [13][14]. These two sets of data were used to train the models in the two main experiments: without hybridization and with hybridization.

B. Phase 2 – Train Models without Hybridisation

Phase 2 involved two parts. The first part was the building, training, and testing of ensemble-based models by using Bagging (BAG), Random SubSpace (RNDS), MultiClass Classifier (MCC), and Rotation of Forest (ROF) Algorithms [15]. The second part was the building, training and testing of the base learner or classification-based models using Naïve Bayes (NB), MultiLayer Perceptron (MLP), k-Nearest Neighbour (KNN), and Decision Table (DT) algorithms. The test option for both parts was to use a training set with 10-fold cross-validation during training and using the supplied test set with 10-fold cross-validation during the model testing.

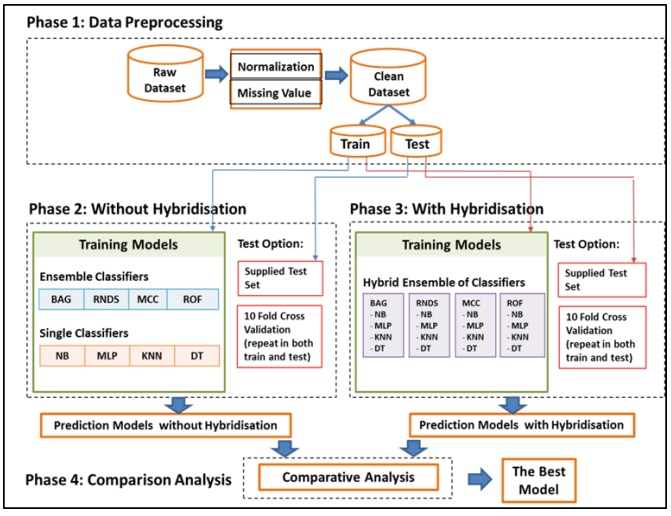


Fig. 1. Proposed Research Approach.

C. Phase 3 – Train Models with Hybridisation

Phase 3 involved the building, training, and testing of all the hybrid ensemble-based models by hybridising ensemble algorithms with classification algorithms as base learners [14][16]. The models were BAG+NB, BAG+MLP, BAG +KNN and BAG+DT. Followed by RNDS+NB, RNDS+MLP, RNDS+KNN, and RNDS+DT. Next were MCC+NB, MCC+MLP, MCC+KNN, and MCC+DT. The last hybrid ensemble-based models were ROF+NB, ROF+MLP, ROF+KNN, and ROF+DT. The test option for both parts was to use a training set with 10-fold cross-validation during training and using the supplied test set with 10-fold cross-validation during the model testing.

D. Phase 4 – Perform Comparison Analysis

Phase 4 involved the comparison analysis of the performance metrics for all the models trained in phase 2 and phase 3. The metrics were in terms of accuracy, precision, recall, F-measure, and ROC area [14][17][18]. The models were the ensemble-based models, classification-based models, and hybrid ensemble-based models.

V. Research Preliminaries

A. Dataset Descriptions

Real data were collected based on more rational attributes that were suggested by the previously conducted relevant research. An online questionnaire, generated using Google forms, was circulated on social media to different groups targeting university students taking the Cloud Computing course to gather the necessary data. A total of 319 students filled out the questionnaire, which was considered an appropriate dataset size to be used in building and training single classifier-based models, ensemble classifier-based models, and classifier-based hybrid ensemble models. The questionnaire was designed to include students' demographic and students' motivational behaviour questions for the course cloud computing. The independent variables can be easily transformed to dependent variables or attributes that may predict the class of final examination results (Excellent, Very Good, Good, Past, Fail). The list of collected attributes is illustrated in Table I.

B. Performance Metric Descriptions

1) Multi class confusion matrix evaluation: The prediction model's method of evaluating fitness was by analysing the confusion matrix. The confusion matrix, as shown in Table II, contained information about the proposed classifier's actual and predicted classification. With the aid of the Academician expert, the proposed model was verified to check the prediction model's accuracy.

2) Accuracy detailed evaluation: The performance metrics that we apply to assess the proposed model's performance were in terms of classification accuracy, recall, precision, F-measure, and ROC area [19]. Table III shows the classification measures representations.

TABLE I. DATASET DESCRIPTION

Attribute	Description	Possible Value
Demographic Information		
Age	Student's Age	18-20, 21-24, > 25
FSize	Student's Family Size	Less or equal than 3, more than 3
HStatus	Student's health status	Healthy, Not Healthy
MStatus	Student's marital status	Single, Married, Other
WStatus	Student's work status	Full-time job, part-time job, online job
Family Highest Education Level	Student's family highest education level	Tertiary, Secondary, Primary
Motivational Behaviour Information		
Level of Self Confidence	Confidence towards this course	High, Medium, Low
Level of interest	Interest towards this course	High, Medium, Low
Level of positive thinking	Positive thinking towards this course	High, Medium, Low
Level of focus in class	Focus in the class	High, Medium, Low
Have Personal Goal	Personal Goal towards this course	Yes, No
Academic Information		
Major Program	Name of program	CS, IS, IT
No. of Absence	Hours of absence in the class	1-2, 3-4, > 4
Current GPA	Current GPA	1.0 – 1.9, 2.0 – 2.9, 3.0 – 3.9, 4.0 – 5.0
Grade acquire in First Exam	Exam 1 grade (20%)	Excellent, V. Good, Good, Pass, Fail
Data Class	Categories	Excellent, V. Good, Good, Pass, Fail

TABLE II. MULTI-CLASS CONFUSION MATRIX EVALUATION

Actual		Predicted				
		Very Good	Excellent	Fail	Pass	Good
		a	b	c	d	e
Very Good	a	**TP**	Error $_{ab}$	Error $_{ac}$	Error $_{ad}$	Error $_{ad}$
Excellent	b	Error $_{ba}$	**TP**	Error $_{bc}$	Error $_{bd}$	Error $_{bd}$
Fail	c	Error $_{ca}$	Error $_{cb}$	**TP**	Error $_{cd}$	Error $_{cd}$
Pass	d	Error $_{da}$	Error $_{db}$	Error $_{dc}$	**TP**	Error $_{dd}$
Good	e	Error $_{ea}$	Error $_{eb}$	Error $_{ec}$	Error $_{ed}$	**TP**

TP – True Positive E- Error

TABLE III. ACCURACY DETAILED EVALUATION

Accuracy Detailed	Explanation
Classification Accuracy	Accuracy measures the proportion of instances that are a correctly classified. Formula: (TP + TN)/(TP + FP + FN + TN)
Sensitivity (Recall)	Sensitivity is the proportion of positive factors that the classifier correctly identifies as positive. Formula: TP/(TP + FN)
Precision	This is a calculation of the relevant collected instances Formula: TP/(TP + FP)
F-Score / F-Measure	The F1 Score is needed to seek a balance between Precision and Recall. Formula: F1 = 2 (Precision * Recall)/(Precision + Recall)
ROC Area	The AUC-ROC curve is a classification problem quality calculation at various threshold settings. ROC is the curve of probability and AUC is the degree metric. It indicates how much a model between classes can be differentiated. The higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. Formula: FP /(FP + TN)).

C. Single Algorithm Descriptions

1) Naïve Bayes Algorithm (NB): NB algorithm is a supervised method of learning as well as a quantitative approach of classification proposed by Thomas Bayes [21]. This algorithm assumes a proactive inhibition model, which results in doubt about the system by specifying the likelihood of outcomes in theory. Diagnostic and predictive problems can be addressed. The Bayesian classification provides realistic algorithms for learning and prior knowledge incorporating observed data.

2) Multilayer Perceptron (MLP): The feed-in class to the artificial neural network is an MLP. The MLP contains at least three layers of nodes: an input layer, a hidden layer, and an output layer [21]. The node, excluding the input nodes, is a neuron that uses a nonlinear activation function. MLP uses the guided learning method for instruction, called back propagation. MLP differentiates from a linear perceptron by its multiple layers and non-linear activation. Data that cannot be separated linearly can be differentiated.

3) k-Nearest Neighbour Algorithm (kNN): kNN is an algorithm of gradation that is widely used in the identification of statistical patterns [21]. Every class has a few sample models and a set of pattern vectors. When a vector must be named, it will be one of the model vectors that is its nearest k neighbour. The majority rule is the tag category. To prevent relations to class and overlap areas, the value of k should be odd. This rule is sophisticated but plain, and, in practice, has a low error rate.

4) Decision Table Algorithm (DT): DT are classification models used in forecasting, such as Decision Trees and neural networks [22]. They are induced by algorithms for machine learning. A table of decisions includes a hierarchy of the values in which each entry in a higher table is broken down to form a table by the values of a pair of additional attributes.

D. Ensemble Algorithms Description

1) Bagging Algorithm (BAG): An Ensemble Meta-Stimator for BAG is an ensemble that fits the base classifiers into a random subset of the original dataset and then aggregates its individual forecasts (by vote or by means of an average of) into a final prediction [22]. Normally, such a meta-estimator can be used as a means of reducing the variance of a blackbox estimator (e.g. the decision tree), randomizing its design process, and then creating an ensemble from it.

2) Random Subspace Algorithm (RNDS): The base classifier model is based on a set constructed from the initial set of functionalities using the RNDS approach proposed by Ho [23]. Through a simple majority vote procedure, the outcomes of the individual graders are merged in a final decision.

3) Multiclass Classifier Algorithm (MCC): MCC is a metaclassifier with 2-class classifiers for managing multi-class datasets [22]. This can also add error to correct a metaclassifier output code in order to improve accuracy.

4) Rotation Forest Algorithm (ROF): ROF is a way to produce classifier assemblies based on the extraction of features [22]. The feature set is randomly divided into K (K is an algorithm parameter), and Principal Component Analysis (PCA) is applied to each subset to create training data for a base categorizer. In order to preserve variation information in the results, all the principal components are retained. Therefore, the K axis rotation forms the new features for a simple classification system.

VI. Experimental Results

In this study, four main experiments were conducted sequentially with the aim to assess the students' performance using cloud computing course dataset by training various single, ensemble and hybrid ensemble classifiers. Followed by conducting comparative analysis to detect any performance improvement in all the different types of models. These experiments eventually identify the best performing model in predicting student's performance on cloud computing course [20].

A. Experiment 1: Training Models without Hybridisation

The aim of this experiment is to observe the effect of the four ensemble classifiers and the four single classifiers without the process of hybridisation between the two classifiers type. In total, eight models were evaluated in this experiment. Fig. 2 shows the results of the evaluation which indicate that each classifiers have achieved their highest performance for different metrics. ROF model achieved the highest accuracy value at 90.90% and also the highest ROC metric value at 98.10%. MCC model obtained the highest precision value at 83.8% and also the highest F-score value at 86.10%. Whereas, RNDS model achieved the highest recall value at 94.30%.

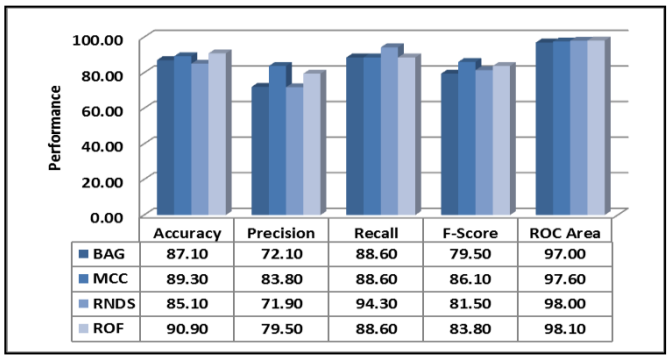
Fig. 3 shows the experiment results for the single-based model evaluation. It shows that MLP out-performed the rest of the models by obtaining the highest accuracy value at 90.50%, the highest precision value at 81.4%. The highest F-score value at 84.90% and the highest ROC value at 97.60%. However, in terms of the recall metric, NB and DT models achieved the highest value at 91.40%.

B. Experiment 2: Training Models with Hybridisation

The aim of this experiment is to hybrid the ensemble classifiers with the single classifiers as the base learners. In this experiment, we thoroughly evaluated 16 hybrid ensemble models. The results were shown in Fig. 4, Fig. 5, Fig. 6 and Fig. 7.

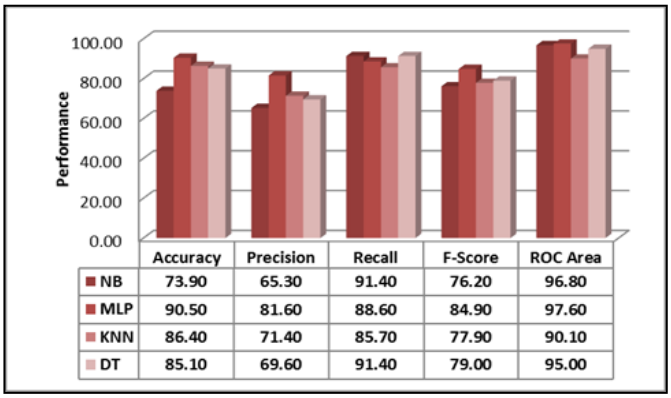
Fig. 4 shows the evaluation results of the hybrid BAG-based models which indicated that BAG+MLP model achieved the highest performance in all the evaluation metrics. This model obtained the highest accuracy metric (89.30%), the highest precision metric (76.20%), the highest F-score metric (83.90%) and the highest ROC metric (98.30%). However in terms of the recall metric, this model shared the highest value with BAG+NB, and BAG+DT models at 91.40%.

Fig. 5 shows the experiment results for the hybrid RNDS-based models' evaluation. The results shows that, RNDS+KNN model obtained the highest accuracy value at 87.60%, the highest precision value at 76.20% and the highest ROC value at 98.60%. However, in terms of recall metric, RNDS+DT model achieved the highest value at 100% and RNDS+MLP model demonstrated the highest F-Score value at 78.50%.



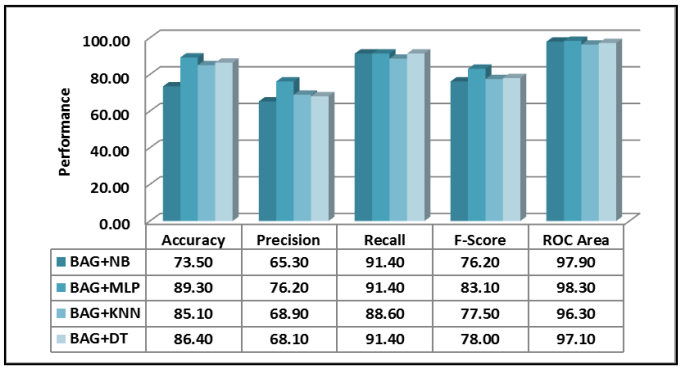
	Accuracy	Precision	Recall	F-Score	ROC Area
BAG	87.10	72.10	88.60	79.50	97.00
MCC	89.30	83.80	88.60	86.10	97.60
RNDS	85.10	71.90	94.30	81.50	98.00
ROF	90.90	79.50	88.60	83.80	98.10

Fig. 2. Ensemble-Based Models Evaluation.



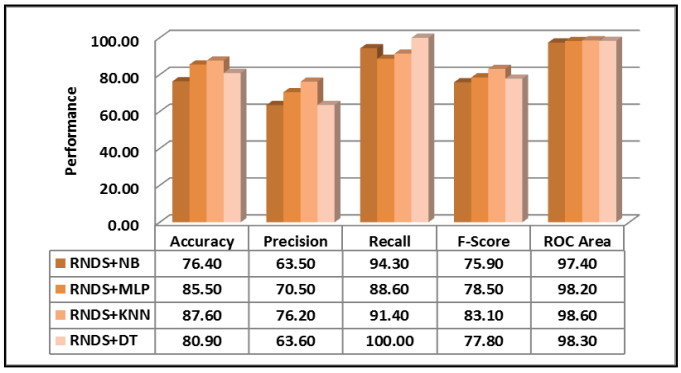
	Accuracy	Precision	Recall	F-Score	ROC Area
NB	73.90	65.30	91.40	76.20	96.80
MLP	90.50	81.60	88.60	84.90	97.60
KNN	86.40	71.40	85.70	77.90	90.10
DT	85.10	69.60	91.40	79.00	95.00

Fig. 3. Single-Based Models Evaluation.



	Accuracy	Precision	Recall	F-Score	ROC Area
BAG+NB	73.50	65.30	91.40	76.20	97.90
BAG+MLP	89.30	76.20	91.40	83.10	98.30
BAG+KNN	85.10	68.90	88.60	77.50	96.30
BAG+DT	86.40	68.10	91.40	78.00	97.10

Fig. 4. Hybrid BAG-Based Models Evaluation.

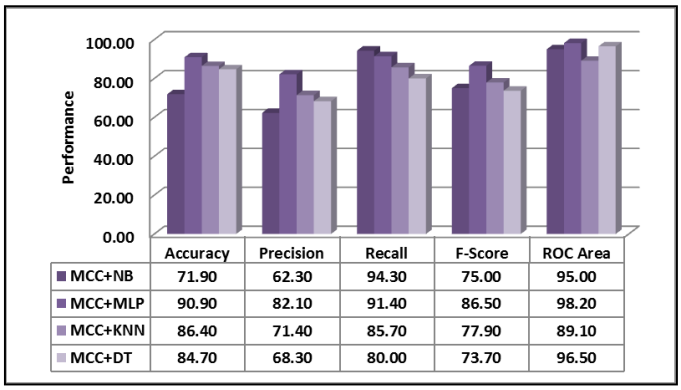


	Accuracy	Precision	Recall	F-Score	ROC Area
RNDS+NB	76.40	63.50	94.30	75.90	97.40
RNDS+MLP	85.50	70.50	88.60	78.50	98.20
RNDS+KNN	87.60	76.20	91.40	83.10	98.60
RNDS+DT	80.90	63.60	100.00	77.80	98.30

Fig. 5. Hybrid RNDS-Based Models Evaluation.

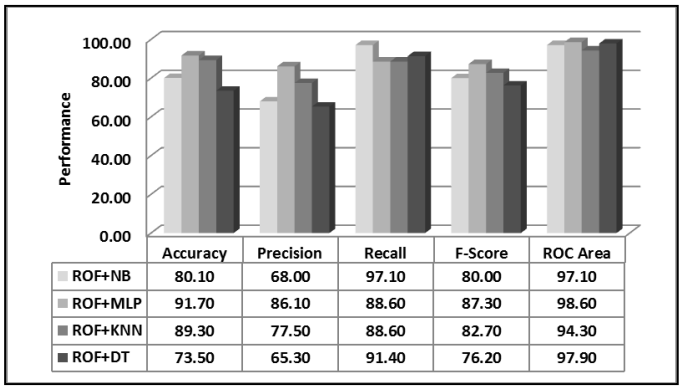
Fig. 6 presents the results for the hybrid MCC-based models' evaluation. The results clearly demonstrated that MCC+MLP model out-performed the rest in terms of the accuracy metric (90.90%), the precision metric (82.10%), the F-score metric (86.50%) and the ROC metric (98.20%). However, in terms of the recall metric, MCC+NB model achieved the highest value at 94.30%.

Fig. 7 shows the experiment results for the hybrid ROF-based models' evaluation which highlight the best performance of ROF+MLP model in terms of accuracy (91.70%), precision (86.10%) and ROC (98.60%). However, in terms of recall and F-score metric, ROF+NB model achieved the highest value at 97.10% and 88.00%, respectively.



	Accuracy	Precision	Recall	F-Score	ROC Area
MCC+NB	71.90	62.30	94.30	75.00	95.00
MCC+MLP	90.90	82.10	91.40	86.50	98.20
MCC+KNN	86.40	71.40	85.70	77.90	89.10
MCC+DT	84.70	68.30	80.00	73.70	96.50

Fig. 6. Hybrid MCC-Based Models Evaluation.



	Accuracy	Precision	Recall	F-Score	ROC Area
ROF+NB	80.10	68.00	97.10	80.00	97.10
ROF+MLP	91.70	86.10	88.60	87.30	98.60
ROF+KNN	89.30	77.50	88.60	82.70	94.30
ROF+DT	73.50	65.30	91.40	76.20	97.90

Fig. 7. Hybrid ROF-Based Models Evaluation.

C. Experiment 3: Comparative Analysis

In this analysis, our aim is to observe any performance improvement in the single-based, ensemble-based and hybrid ensemble-based models by comparing the evaluation results of the hybrid and the non-hybrid models. The first experiment is to compare between the non-hybrid models. The second experiment is to compare between the hybrid models. The third experiment is to evaluate the confusion matrix and AUC ROC that corresponds to the best-fitted model.

1) Experiment 3-1: Comparative Analysis of the Non-Hybrid Models: The aim of this experiment is to observe the performing achievement between the non-hybrid models. The experiment results, as shown in Fig. 8 demonstrated that ensemble-classifiers out-performed the single classifiers-based model. ROF-based models perform better in accuracy metric (90.9%) and in ROC metric (98.10%). Whereas, MCC-based model achieved the highest precision value at 83.8%. However for the recall metric RNDS-based model achieved the highest value at 94.3%. Nevertheless, MLP-based model which represent the single-classifier, achieved the highest F-score value at 84.9%.

2) Experiment 3-2: Comparative Analysis of the Hybrid Models: The aim of this experiment is to identify the best performing hybrid model by evaluating and comparing the hybrid models' performance accordingly. Due to the complexity of the experiments, the results representation were divided into three parts as shown in Fig. 9, Fig. 10 and Fig. 11. In Fig. 9, it was observed that ROF+MLP performed better in the accuracy metric (91.7%) and in the precision metric (86.10%). In other words, this model can predict student performance for the excellent class with 91.7% accuracy as compared to the rest of the hybrid models. The result in terms of the precision metric can be interpreted as the model's ability to precisely predict that 86.1% of the data were relevant to the 'excellent', 'very good', 'good', 'pass', and 'fail' class. The results clearly indicate that the hybrid ensemble-based model improves the accuracy and precision of the prediction model.

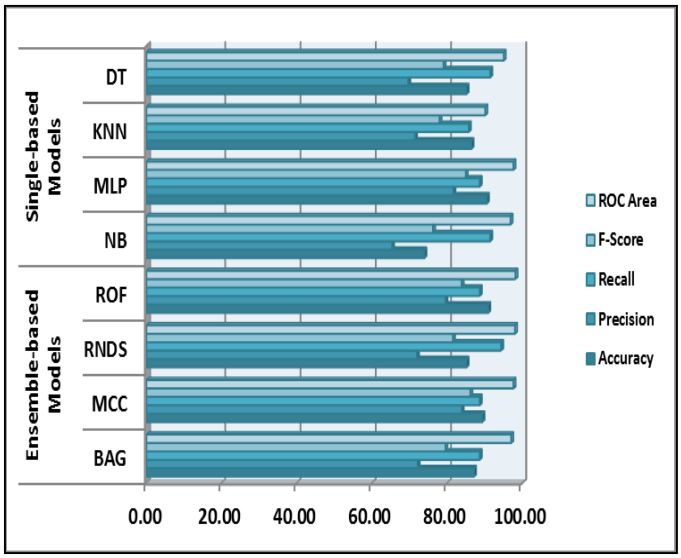


Fig. 8. Comparison between Non-Hybrid Models.

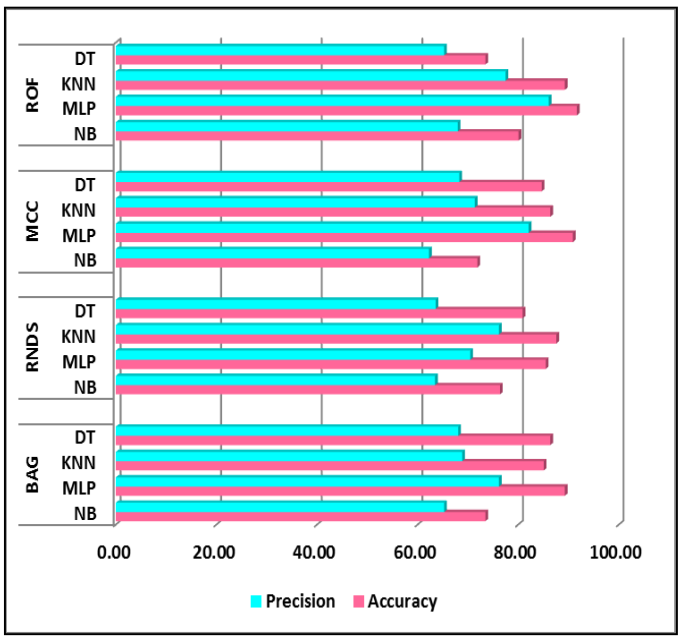


Fig. 9. Comparison of Precision and Accuracy.

Fig. 10 observed the comparison analysis between the hybrid ensemble-based models in terms of recall and f-score metric. The results indicate that RNDS+DT model achieved the highest recall value of 100%. Whereas, ROF+MLP model achieved the highest F-score value at 87.3%.

The following experiment is to evaluate and compare in terms of ROC area metric as shown in Fig. 11. The results demonstrated that ROF+MLP and RNDS+KNN model have the highest ROC value of 98.6%. In other words, by analogy, the higher the ROC, the better the model is at distinguishing between students' grades which were classified as 'excellent', 'very good', 'good', 'pass', or 'fail'.

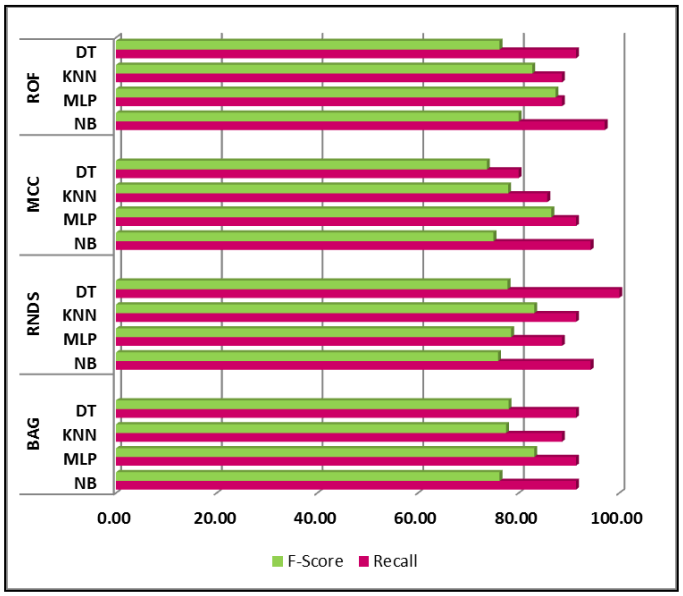


Fig. 10. Comparison of F-Score and Recall.

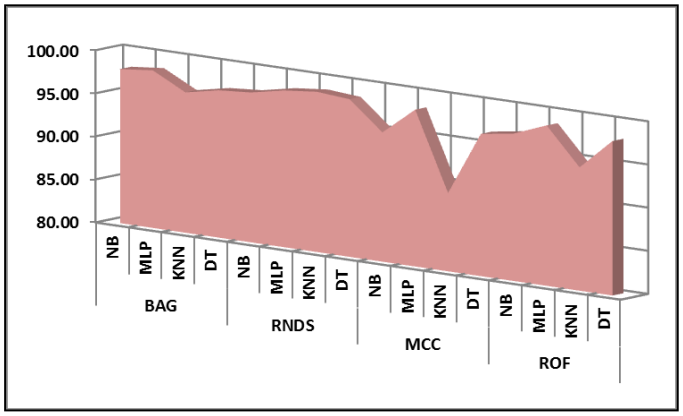


Fig. 11. Comparison of the AUC ROC Area.

VII. Discussion and Analysis

After comparing all the models, we have sufficient evidence to show that the hybridised ensemble models out-performed the non-hybridised ensemble-based models and also the single-based models. Table IV shows the summary of the comparative analysis.

A. Experiment 4: The Confusion Matrix of ROF+MLP

Based on Table IV, there is clear evidence that ROF+MLP model is the best fitted model for predicting student academic performance in the cloud computing course. Thus, this experiment aim to prove that ROF+MLP as the best performing model by using the confusion matrix and observing the ROC area for all the classes. Confusion Matrix results as shown in Table V, confirms the above findings.

1) Confusion Matrix for ROF+MLP Model: Table V shows the confusion matrix for the ROF+MLP model. The results indicate that the model can correctly predicts 31 students or 93% are 'Excellent' students. The model also correctly predicts the rest of the class as follows: 30 students or

81% are 'Very Good' students, 78 students or 95% are 'Good' students, 27 students or 96% are 'Good' students, and 54 students or 95% are 'Fail' students. The confusion matrix indicates that the ROF+MLP model has excellent ability to correctly predict student performance with less than a 19% error.

TABLE IV. THE SUMMARY OF THE COMPARATIVE ANALYSIS

Performance Metrics	Single-based Model	Ensemble-based Model	Hybrid Ensemble-based Model
Classification Accuracy	MLP = 90.5%	ROF=90.9%	ROF+MLP = 91.7%
Precision	MLP = 81.6%	MCC=83.8%	ROF+MLP = 86.1%
Recall	NB = 91.4% DT = 91.4%	RNDS=94.3%	RNDS+DT = 100%
F-Score	MLP = 84.9%	MCC=86.1%	ROF+MLP = 87.3%
ROC (AUC)	MLP = 97.6%	ROF=98.1%	ROF+MLP = 98.6% RNDS+KNN = 98.6%

TABLE V. CONFUSION MATRIX FOR ROF+MLP MODEL

Actual		**Predicted**				
		V. Good	Excellent	Fail	Pass	Good
		a	b	c	d	e
Very Good	a	**30**	3	0	0	1
Excellent	b	3	**31**	0	0	1
Fail	c	0	0	**54**	2	0
Pass	d	2	1	2	**27**	2
Good	e	2	1	1	1	**78**

2) ROC for Each Class in ROF+MLP Model: The aim of this experiment is to observe the performance of ROF+MLP model in distinguishing the value between classes in the model. Fig. 12 shows the experiment results with regards to the Area under ROC or the threshold curve for the individual classes in the ROF+MLP model. The results indicate that all the classes have a high value of 'Area Under ROC'. In other words, ROF+MLP model is good at distinguishing between class = "Excellent", class = "Very Good", class = "Good", class = "Pass", and class = "Fail" students. The highest value is obtained by the class = "Fail" with 99.9%. Followed by class = "Good" with 99.6%. The third place is class = "Excellence" with 97.95%. The fourth place is class = "Very Good" with value of 97.56%. The lowest value under ROC is class = "Pass" with value 86.7%.

Class = Excellent, ROC = 97.95%

Class = Very Good, ROC = 97.56%

Class = Good, ROC = 99.6%

Class = Pass, ROC = 86.47%

Class = Fail, ROC = 99.96%

Fig. 12. AUC ROC for Each Classes in ROF+MLP Model.

VIII. CONCLUSION AND FUTURE WORK

Cloud computing is considered to be a very tough course for most students. Hence early warning of the assessment outcome would be beneficial to at-risk students who have problems in sustaining their grades in that course through-out the whole semester. A robust hybrid ensemble model (RHEM) is highly useful in the prediction of assessment course outcome, assisting the students in deciding to continue or to drop the course at early semester. Based on the summary of the comparative analysis depicted in Table IV, it clearly demonstrated that the hybrid ensemble classifiers were able to improvise the ensemble and also the single classifiers. After many iterations of thorough and rigorous training that were carried out using all 24 models, the analysis indicated that Rotation Forest ensemble classifier hybrid with Multilayer Perceptron classifier as the base learner (ROF+MLP), appears to be the best robust hybrid ensemble model or RHEM that out-performed the rest of the models to predict students' performance in cloud computing course at early stage of the semester.

A logical extension of this work would be the creation of a meta-analysis system for future study, which can be regarded as a decision support method based on the model that will achieve the highest efficiency and effectiveness.

ACKNOWLEDGMENT

This research was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University through the Fast-track Research Funding Program.

REFERENCES

[1] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, "Student academic performance prediction using supervised learning techniques," Int. J. Emerg. Technol. Learn., vol. 14, no. 14, p. 92, 2019.

[2] G. Akçapınar, A. Altun, and P. Aşkar, "Using learning analytics to develop early-warning system for at-risk students," Int. J. Educ. Technol. High. Educ., vol. 16, no. 1, 2019.

[3] M. Zaffar, M. Ahmed, K. S. Savita, and S. Sajjad, "A study of feature selection algorithms for predicting students academic performance," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 5, 2018.

[4] P. Nuankaew, "Dropout Situation of Business Computer Students, University of Phayao," Int. J. Emerg. Technol. Learn, vol. 14, no. 19, p. 115, 2019.

[5] A. J. P. Delima, Ariel, and Ruji, "Variable reduction-based prediction through modified genetic algorithm," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 5, 2019.

[6] A. Hellas, P. Ihantola, A. Petersen, V.V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom, and S.N. Liao, "Predicting academic performance: a systematic literature review," in Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education - ITiCSE 2018 Companion, 2018.

[7] A. Daud, N. R. Aljohani, R. A. Abbasi, M. D. Lytras, F. Abbas, and J. S. Alowibdi, "Predicting student performance using advanced learning analytics," in Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion, 2017.

[8] J. Ross, E. Guerra, and S. Gonzalez-Ramos, "Linking a hierarchy of attitude effect to student engagement and chemistry achievement," Chem. Educ. Res. Pr., vol. 21, no. 1, pp. 357–370, 2020.

[9] N. Sael, T. Hamim, and F. Benabbou, "Implementation of the analytic hierarchy process for student profile analysis," Int. J. Emerg. Technol. Learn., vol. 14, no. 15, p. 78, 2019.

[10] E. Abou Gamie, S. Abou El-Seoud, and M. A. Salama, "Comparative analysis for boosting classifiers in the context of higher education," Int. J. Emerg. Technol. Learn., vol. 15, no. 10, p. 16, 2020.

[11] L. H. Son and H. Fujita, "Neural-fuzzy with representative sets for prediction of student performance," Appl. Intell., vol. 49, no. 1, pp. 172–187, 2019.

[12] X. Zhang, R. Xue, B. Liu, W. Lu, and Y. Zhang, "Grade prediction of student academic performance with multiple classification models," in 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2018.

[13] O. W. Adejo and T. Connolly, "Predicting student academic performance using multi-model heterogeneous ensemble approach," J. appl. res. high. educ., vol. 10, no. 1, pp. 61–75, 2018.

[14] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining approach," J. Med. Syst., vol. 43, no. 6, p. 162, 2019.

[15] A. Alhassan, B. Zafar, and A. Mueen, "Predict Students' Academic Performance based on their Assessment Grades and Online Activity Data," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 4, 2020.

[16] P. Sokkhey and T. Okazaki, "Hybrid machine learning algorithms for predicting academic performance," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 1, 2020.

[17] A. Almasri, E. Celebi, and R. S. Alkhawaldeh, "EMT: Ensemble meta-based tree model for predicting student performance," Sci. Program., vol. 2019, pp. 1–13, 2019.

[18] R. Hasan, S. Palaniappan, S. Mahmood, K. U. Sarker, and A. Abbas, "Modelling and predicting student's academic performance using classification data mining techniques," Int. J. Bus. Inf. Syst., vol. 34, no. 3, p. 403, 2020.

[19] X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms," Comput. Human Behav., vol. 98, pp. 166–173, 2019.

[20] A. I. Adekitan and E. Noma-Osaghae, "Data mining approach to predicting the performance of first year student in a university using the admission requirements," Educ. Inf. Technol., vol. 24, no. 2, pp. 1527–1543, 2019.

[21] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining," Knowl. Based Syst., vol. 200, no. 105992, p. 105992, 2020.

[22] O. Sagi and L. Rokach, "Ensemble learning: A survey," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 8, no. 4, p. e1249, 2018.

[23] D. Ndirangu, W. Mwangi, and L. Nderu, "Improving multiclass classification and outlier detection method through ensemble technique," in Proceedings of the 4th International Conference on Communication and Information Processing - ICCIP '18, 2018.

An Ontology-Based Predictive Maintenance Tool for Power Substation Faults in Distribution Grid

Moamin A. Mahmoud[1], Alicia Y.C. Tang[2]
Institute of Informatics and Computing in Energy
Universiti Tenaga Nasional
Kajang, Malaysia

Nur Liyana Law Mohd Firdaus Law[4]
Mathuri Gurunathan[5]
College of Computing and Informatics
Universiti Tenaga Nasional, Kajang, Malaysia

Kuganesan Kumar[3]
Graduate School of Business
Universiti Kebangsaan Malaysia
Bangi, Malaysia

Durkasiny Ramachandran[6]
College of Engineering
Universiti Tenaga Nasional
Kajang, Malaysia

***Abstract*—Recent advances in Power Grid (PG) technology pose an important problem of measuring the effectiveness of power grid configurations. Current assessment models are not adequate to mitigate the setup issues due to the absence of a high-fidelity evaluation framework that can consider diverse scenarios based on the market interest. Consequently, we develop a highly flexible Ontology-based Evaluation System that can accommodate and assess different scenarios. The use of ontology as middleware is the best approach to produce an efficient, semantically aware, and operationally accurate system environment for managing flexibility in evaluation. The evaluation is made by predicting the failure intensity and subsequently generate a maintenance report of a particular configuration. The selection of the best configuration is made by comparing the maintenance report of different configurations. The developed evaluation system consists of three main components which are Configuration Generator Tool (GCT), Failure Prediction Model (FDM), and Hybrid Simulation Platform (HSP). The GCT is a knowledge-based system that provides a powerful tool for engineers to generate alternative configurations. The GCT data were collected from literature, validated by experts, and modeled using Web Ontology Language (OWL). While the HSP was developed using several modelings and ontology-based tools such as blender 3D modeling, unity 3d, asp.net, my sql, and apache Jena fuseki. Finally, the FDM was developed based on the impact and relationship of odd events to power grid components and the impact of a failed component to other components, the prediction is modeled using two methods Poisson Model and Likelihood Estimation Method.**

Keywords*—*Predictive maintenance; ontology; power substation faults; distribution grid

I. Introduction

The Energy Power Grid Ontology development proposed in this paper is expected to allow the sharing of ontology among applications and stakeholders. Utilizing a similar lexicon, the normal issues that happen during application interoperability will be anticipated and fathomed. Ontologies development that covers domain and application characteristics can be utilized to not as it bolstered systems integration by utilizing standardized vocabularies but to system advancement by reusing these ontologies. As such, the use of ontology for the energy power grid is very important in order to allow for applications to inter-operate and stakeholders to share knowledge [13] [14] [21].

Ontologies have been effective in integrating the knowledge needed for tackling complex issues, for example, energy management issues. The ontology usage as middleware is the most ideal approach to manage produce a semantically aware, productive, and operationally accurate program atmosphere for directing flexibility in the distribution system. A typical lexicon and information model helps ease the exertion needed for programming software antiques to communicate viably with others within the energy management framework [11] [12]. Another application is ontology in agent communication language in complex systems [23] [24] [25] [26] [27].

Fault occurrences in the distribution systems are due to multiple reasons, such as human mistakes, components failure, lightning strikes, or infrastructure aging. These marvels are basic and proactive activity. Hence, exact data of the fault area plays a crucial part in speeding up the reclamation process, after being exposed to any sort of fault in power distribution grids [1]. Techniques for fault detection, location, and classification, are in transmission lines and distribution networks have been intensively considered over the long term. With the ideas related to power grid drawing in developing concern among analysts, the significance of building an intelligent fault monitoring and analysis framework fit for characterizing and finding various sorts of faults cannot be exaggerated.

Numerous techniques have been created since the past to find and identify faults in distribution systems with the dispersed era. The strategies can be separated into two classifications, artificial intelligence techniques and conventional. Traveling-wave method [2] and impedance-based method [3] are included in Conventional techniques while Artificial Neural Network (ANN) [4], Support Vector Machine (SVM) [5], Fuzzy Logic [6], Genetic Algorithm (GA) [7], and matching approach [8], are included in artificial

intelligence techniques. In any case, the fault area utilizing intelligent strategies are challenging since they require preparing information for processing and are time-devouring.

In the distribution system about the fault data, can be acquired at the operation center utilizing end-client data or protective device operation. For fault identification, power utilities have been practicing conventional techniques in the past. The foremost common conventional strategy is based on a visual inspection and trial-and-error switching. For a little zone, a foot patrol is practiced to look at the conceivable fault location whereas, for a bigger scale zone, a vehicle or helicopter is commonly utilized. This methodology of fault location through the visual inspection is appropriate for overhead lines. In any case, for underground cables, the fault line isn't noticeable. Too, the trial and error method could be a manual cycle of changing the hand-off to on/off condition until the circuit breaker trips. It relies upon the organization administrator's fault-finding encounter to find the faulted segment. In any case, this preparation is time-devouring and in the long run, will harm the performance of cables. Due to these issues, different fault area strategies have been presented for the reason of assisting the method of locating issues.

On the other hand, ontologies contain definitions for objects and sorts of objects just as their semantics and relations in a formal, machine-readable way constraining a shared understanding of a few domain areas [9]. In differentiate to a simple taxonomy, which speaks to a hierarchically organized lexicon of non-specific and specialized concepts, an ontology expands this thought by implies of concept relations or limitations in order to upgrade semantic translation. Primary components of ontologies are a chain of command of concepts representing sorts of substances, relations between concepts, limitations on relations, and instances [10].

Several models have been developed to detect or predict failure using AI techniques [15] [16] [17] [18] [19], however, using ontology for this purpose is quite new. According to Ullah et al., [15], The main anomalies arise due to lopsided current, minor breaks in insulators, contact issues, and increments and fall of voltage levels, and other comparative related issues. Their emphasis on recognizing the increment in temperature, which helps in the cure of breakdown even before it happens. Hence, in this article, for non-destructive defect investigation and prevention in power substations, they utilize the computer vision approach and machine learning to identify the issue in the early stages of hardware breakdown by misusing and taking advantage of the infrared thermal pictures.

Another study by Bhattacharya & Sinha [16] to analyze the fault automatically using machine learning by monitoring the grid and after that within the case of fault decide the type of the fault as and when it happens. To develop an artificially intelligent system that can analyze the grid network data at any given time and decide the health of the network through the utilization of advanced formal models and novel machine learning methods like recurrent neural systems. The scope of this paper emphasized on Predictive Maintenance Tool for Power Substation Faults to evaluate the Distribution Grid configuration since the ontology part is already presented in another paper published by the author [20]. The term "power grid" mentioned in this paper is referring to the portion of the power grid that starts from the substation where the power is stepped-up until the substation where the power is stepped-down.

II. Proposed Ontology-based Predictive Maintenance Tool

As shown in Fig. 1, the process starts when engineers generate several alternative configurations using GCT. The GCT tools contained three types of settings which are Environmental, Operational, and Physical Components. Having generated the configurations, each configuration will be evaluated separately using HSP. In HSP, the duration of simulation of a selected configuration is specified by the user. The next step is to run the simulation, the simulation will receive odd events and communicate with the ontology base to understand the direct impact and indirect impact of each event to the power grid components and subsequently predict failure. The direct impact such as lightening impact to component "A", and indirect impact when we have dependent component such as lightening impact to component "A: and component "A" impact to component "B". During runtime, the simulation can report the status of power grid components, incoming odd events (e.g. rain, wind, lightning), failures, communication activities between the simulation engine and the ontology base, and operational stability level. After the simulation is ended, a maintenance report that shows the failure intensity will be generated. From the failure intensity of each configuration, a decision can be made to choose the best configuration.

Since this paper focuses on the predictive maintenance tool only, the failure prediction model is formulated based on [22]. In this simulation, we use the term odd event which the event that would potentially cause failure. Predicting the failure probability of Component A using Poisson Model, X is the random event.

$$P(X = k) = (e^{-u})(u^{k})\,/k! \quad k = 1,2,3$$

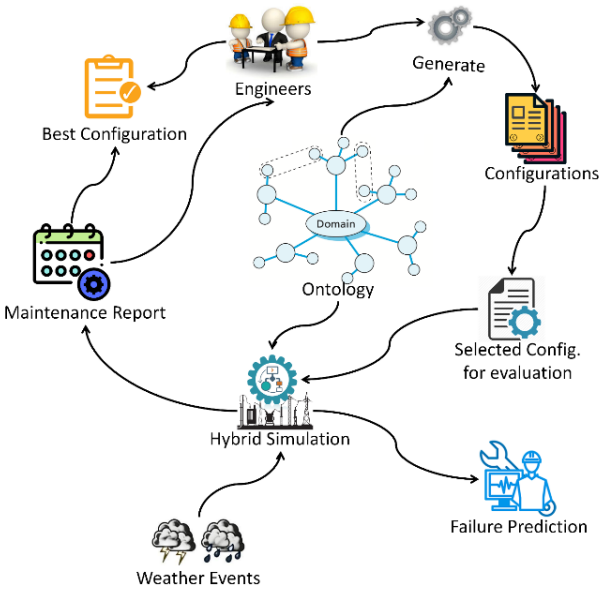


Fig. 1. The Proposed Model.

Given the length of the period and the average events per time to identify u. If no odd event, the life span of a component is,

$$u = \frac{\text{events}}{\text{time}} * \text{T}$$

The failure probability per unit time is

$$P(X \geq 1) = 1 - (e^{-u})$$

If P>0.60 then failure occurred in that component [22]

Thus, the non-failure probability is $1 - P$

If an odd event is considered, Lightning is L and wind speed is W, using the polynomial regression R

$$u = \frac{\text{events}}{\text{time}} * \text{T}$$

According to Shuai et a., 2018

$$R_{LW} = -13.9225 + 3.1331L - 0.0791LW - 0.2645\,L^2 + 0.0727W^2$$

$$R_{LW} = \ln(u)$$

then

$$u = e^{R_{LW}}$$

The failure probability based on events inputs per unit time is

$$P(X \geq 1) = 1 - (e^{-u}).$$

If P>0.10 then failure occurred.

Thus, the non-failure probability is $1 - P$.

III. Substation Ontology

The components of the grid identified are categorized based on their usefulness. Based on the data assembled from open sources and recommendations grid specialists as discussed in [20], the team has compiled a set of standard components of the grid. The Power Grid components are isolated into two primary categories: Transmission and Substation. The components, sub-components, properties, and information values are spoken to by diverse colors to appear the relationship of these components. Categorizing the power grid components will too ease the representation of the components data in an Ontology-based environment.

The substation data is represented in ontology utilizing the Protégé computer program. Based on the discussion with the power grid professional, it is suggested that the study should be conducted on the distribution substation. This can be since the number of substation components is relatively lesser and the experience mastered during the information portrayal cycle will facilitate the future portrayal processes. Other than this, beginning with a more modest scale information representation is continuously empowering. The master too focused that, among power grid components, substations' components are more inclined to expensive failure compared to those transmission components, and hence the group has more practical information to assess the success of the portrayal. The representation begins with the choice of ontology improvement stage. For this project, Protégé was chosen as the development stage. Fig. 2 shows a sample of the visualization extracted from the Protégé development platform that shows the representation of substation's components using OWL in protégé. This portrayal is based on the power grid key components chart. The connection between the most components, sub-components, and the component's properties are clearly shown. Protégé too permits the client to present the relationship of the components in a diagrammatic arrangement as shown in Fig. 2.

The configuration and simulation model is developed by using the software listed in Table I. This model is accessing the data stored in the Ontology storage to perform the simulation on each component of the substation. Based on the simulation, the model is able to predict the possibility of component failure based on the geometrical and environmental factors. The model was tested using the historical data and proved that the model can alert the user with the possibility of the components' failure.

The collected data used in developing the tool are presented in Tables II to VIII.

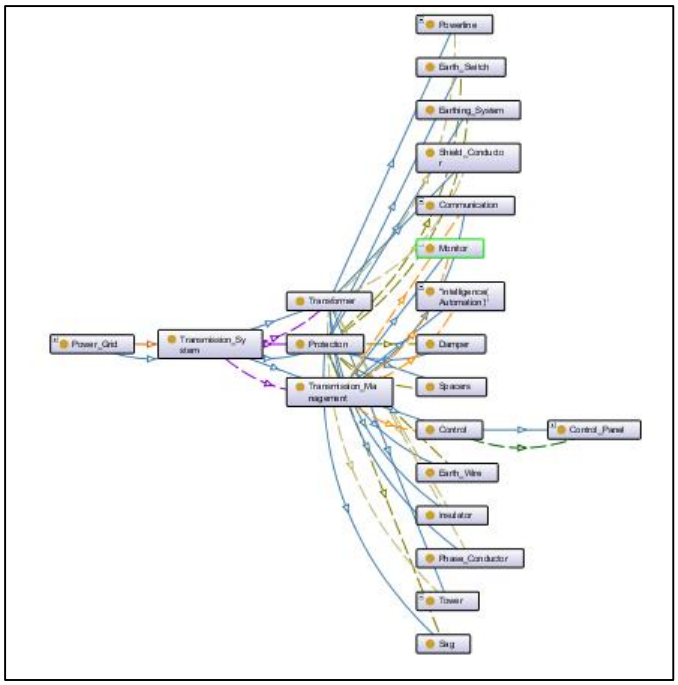


Fig. 2. Diagrammatic Representations of Substation Components.

TABLE I. Software used in the Development

Development	Software used
3D modeling	Blender
3D engine	Unity 3D
Web application	MS SQL
Ontology	Apache Jena, Fuseki-Triplet Database Store
Server	Asp.net
Client	Angularjs

TABLE II. POWERLINE

Components		Specifications				
		Material	Voltage	Phase	distance	Type
Powerline	Overhead AC	• Aluminum • Alloy • Steel • Copper	765kV	Three-phase	3,800km	
	Overhead DC		800kV	Three-phase	4,800km	
	Underground AC	Copper	> 50kV and 10kV - 33kV		80km	
	Underground DC					
	Submarine	• Copper • Fiber optic • Aluminum	• Ac – 30 – 5d5kV • Dc – 250 – 600kV			• Oil linked • Cross-linked polyethylene

TABLE III. TRANSFORMER FOUNDATION

Components			Specifications		
			Types	Foundation Type	Requirement
Transformer	Tower	Foundation	• Steel grillage • Concrete spread footing • Concrete auger/caisson • Pile • Rock • Novel • Undercut pyramid • Concrete auger	• Igneous • Sedimentary rock • Metamorphic soil • Sand/gravel • Cohesive soil	• 1yard cement • 30in x 30in x 6ft (per stub)

TABLE IV. TRANSFORMER CROSS-ARM

Components			Specifications							
			Materials	Braces type	Bolt size	Types	Shape	Bracket type	Voltage	Design
Transformer	Tower	Cross-arm	• Wood • Steel • Fiber glass • Galvanized steel • timber	• Double arming bolt • Pigtail bolt • U bolt	2mm gauge diameter	• Pole top • Two arm • Single arm • Line arm • Side arm	'V' design	• Short aluminum • Short stainless steel • Aluminum • Longest stainless steel • Longest aluminum	• Ac – 35 – 1,000k V / 110 – 1,000k V • Dc – 800kV	• Single post • Single post stayed • Double post stayed

TABLE V. TRANSFORMER STRUCTURE

Components				Specifications						
				Voltage	Height	Angles	Structure	Circuit	width	material
Transformer	Tower	Structure	Lattice steel tower	• 220kV • 500kV	• 110-200ft • 150-515ft • 80-200ft	• L beam • T beam	• Bolt connections • Main structural members • Bracing system	• Double • Single		
			Tubular steel tower	• 110kV / 500kV • 220kV	• 70 – 200ft • 55 – 200ft					• Steel • Steel tube pylons
			h-frame	230kV	60-90ft				100-160ft	wood
			Monopole	Up to 400kV	40m				1.5 -2m	aluminum
			Underground	345kV					40-65m	
			Turning							

TABLE VI. SUBSTATION TRANSFORMER

Components		Specifications				
		Weight	Type	Height	Shape	Material
Substation transformer	Buchholz relay	• 3.9kg • 2,31kg	• BS25 • BS50 • BS80			
	Oil level indicator					
	Breather	5kg	Silicon gel	<1200mm		
	Main tank	100tons			Bell	• Plate steel (base) • Sand blasted • Rus inhibiting primer

TABLE VII. SUBSTATION TRANSFORMER PRIMARY AND SECONDARY WINDING

Components		Specifications				
		Voltage	Material	No. Of turns	Size gauge wire	Height
Substation transformer	Primary winding	2,400 -13,800kV	• Aluminum • Copper	*many	*small	10cm
	Secondary winding	<600kV		*few	*large	

TABLE VIII. SUBSTATION TRANSFORMER

Components		Specifications				
		Type	Material	Voltage	Size	Phase
Substation Transformer	Radiator	• Air natural/self-cooled • Air blast • Oil natural air natural • Oil natural air forced • Oil forced air forced • Oil forced water forced				
	LT & HT Bushes	• Normal • A/F • LV	• Paper • Porcelain • Resin	• 12 – 36kV • 0.66 – 52kV • 1.1 – 3.6kV		
	Tap changer	• No load • On load • Mechanical		Vary		
	Air release plug		M.S.		15mm	• Single phase • Three-phase

IV. SIMULATION AND RESULTS

Fig. 3 shows the landing page for the application. From here the user can navigate to configuration creation page or run a simulation.

As shown in Fig. 3, the user can either enter the simulation from the shortcut in the landing page or from the sidebar menu. There are three steps the user needs to complete before running the simulation, (i) Step 1 – select configuration, (ii) Step 2 – enter the simulation runtime (number of days). Step 3 – confirm the selected settings. Fig. 4 shows the list of created configurations that can be selected for the simulation.

Fig. 5 shows the field where the user must enter the simulation runtime, the number of days the simulation will run.

Fig. 6 shows all the settings, a user has selected; the user expects to confirm these settings before running the simulation.

Fig. 7 shows the entire power grid setup. For this ontology demonstration, we are only focusing on the substation transformer near the housing area. Notice that the menus on the side and top are disabled to prevent the user from accidentally exit the simulation. Users can exit the simulation by clicking on the 'Exit to dashboard' button on the top right corner.

In order to enter the substation transformer simulation, the user has to click on the substation highlighted in Fig. 8.

Fig. 9 shows the substation transformer simulation setup. When the user entered the substation transformer simulation, the simulation will be immediately started. The green lights over the transformer, represent each component, user can hover over it to see the component name as shown in Fig. 10.

Fig. 11 shows the activities in the simulation. The activity is shown by each day followed by the current environmental event if any. As each day pass the simulation will consult ontology bypassing the current environment event, then the ontology will reply back the substation components that might be affected by this environment event.

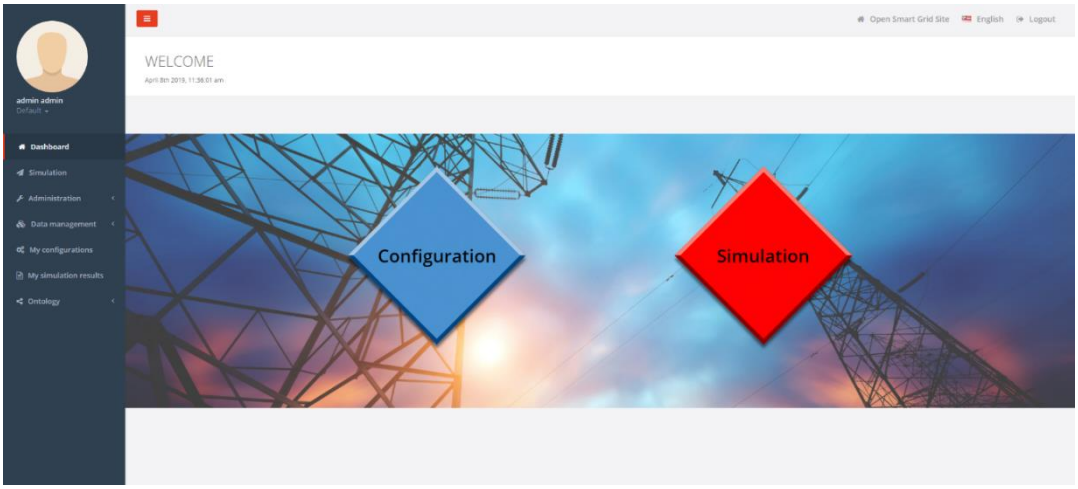


Fig. 3. Landing Page.

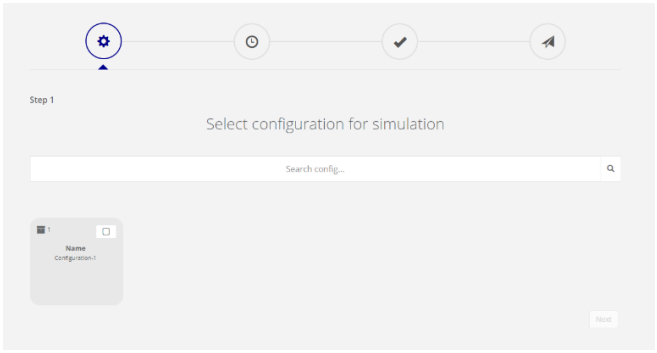


Fig. 4. Select a Configuration for Testing.

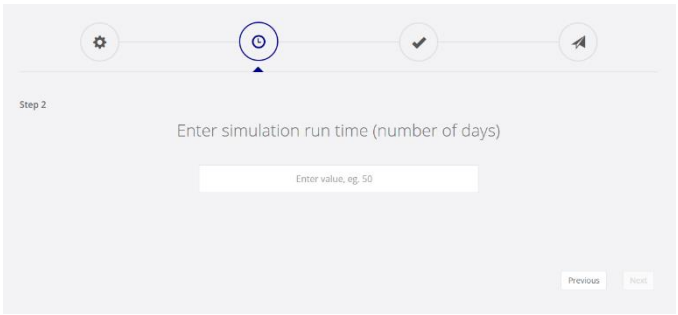


Fig. 5. Enter the Simulation Runtime.

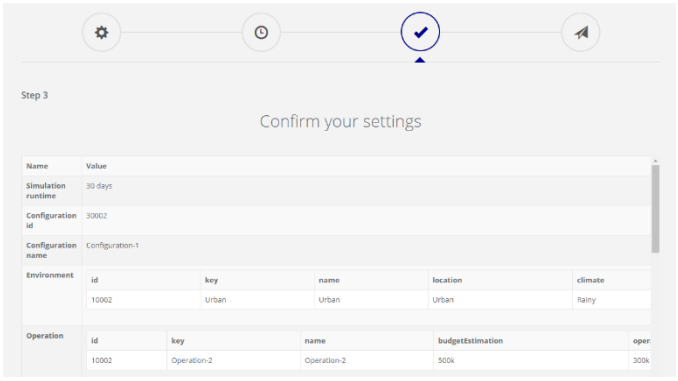


Fig. 6. View All Settings.

Fig. 7. The Entire Power Grid Setup.

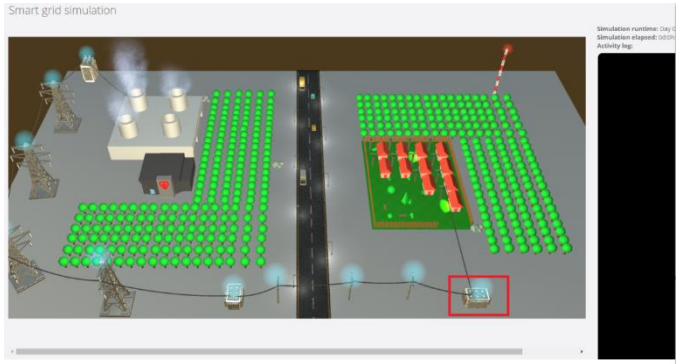


Fig. 8. Substation Transformer.

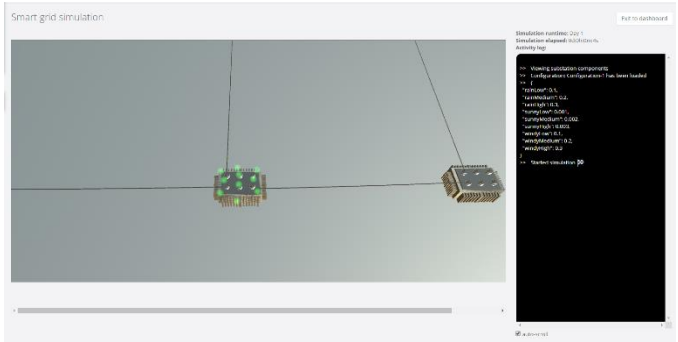


Fig. 9. The Substation Transformer Simulation.

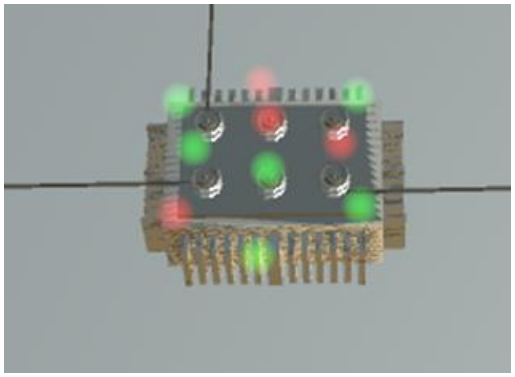


Fig. 10. Shows the Components that Failed/Effected by the Environment Event (Green – Active, Yellow – Predicted Failure, Red – Failure).

Fig. 12 shows the status of the components in more detail. Once the simulation has ended, the user can get the simulation report by going to the report page as shown in Fig. 20.

The simulation tool can analyze the simulation data and display reports of critical information. One of them presenting the failure occurrence and its probability using the Poisson model. As shown in Fig. 13, the probability of the highest probability of failure occurrence in the substation is 63 times with the simulation period. Fig. 14 shows the link to simulation report page. In order to view the report, the user can click on the 'Action' button next to the listed report and see the option to view the report. 'View report' will show a brief readable report, 'View report data' will show the complete data logged in the simulation in raw form. Fig. 14 shows the simulation report. User can also download this report as PDF document by clicking on the 'Download' button on top of the report.

Fig. 15 shows the predicted failure in which day based on the analysis of odd events, based on [22], when the probability exceeds 0.05, the failure would happen.

Fig. 16 Status of the substation components from Day 2 to Day 48 using the Poisson regression model.

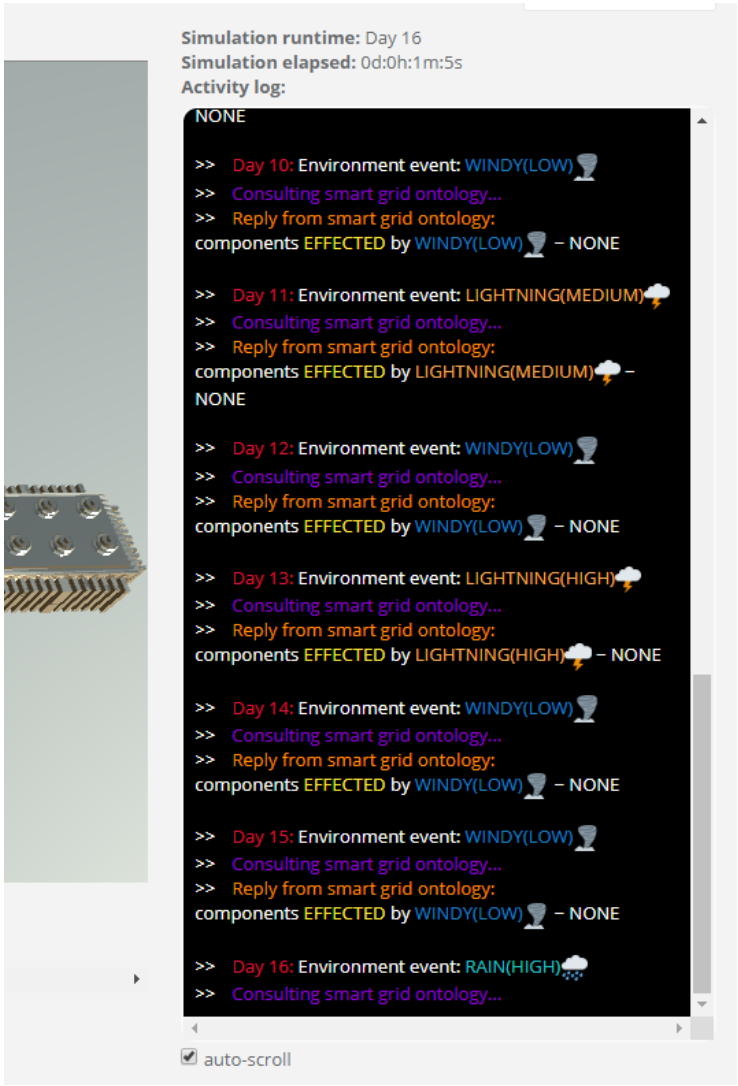


Fig. 11. The Activities in the Simulation.

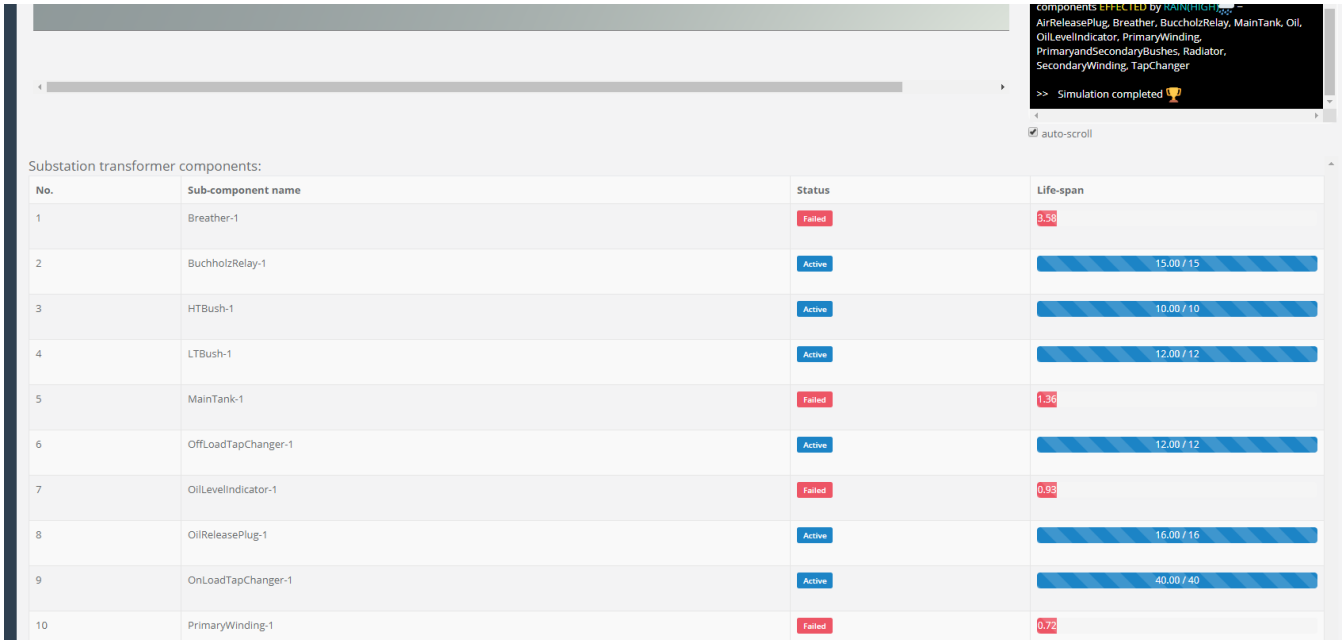


Fig. 12. The Components Status.

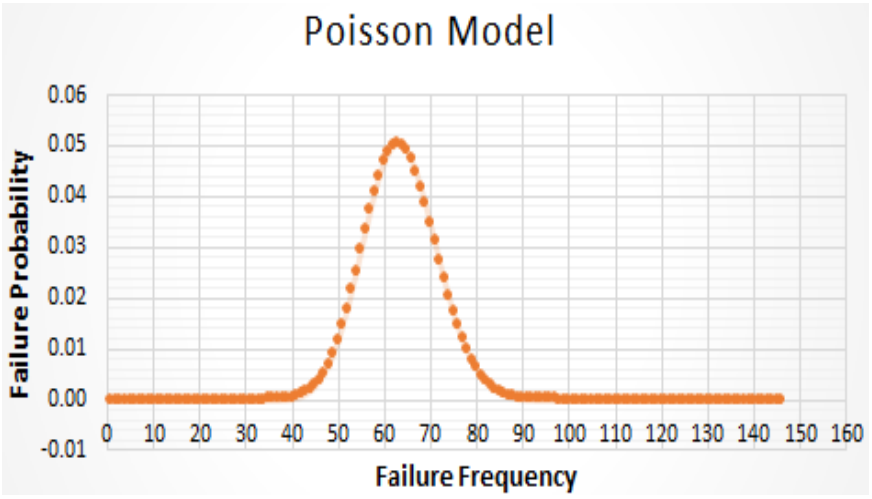


Fig. 13. Failure Occurrence Probability.

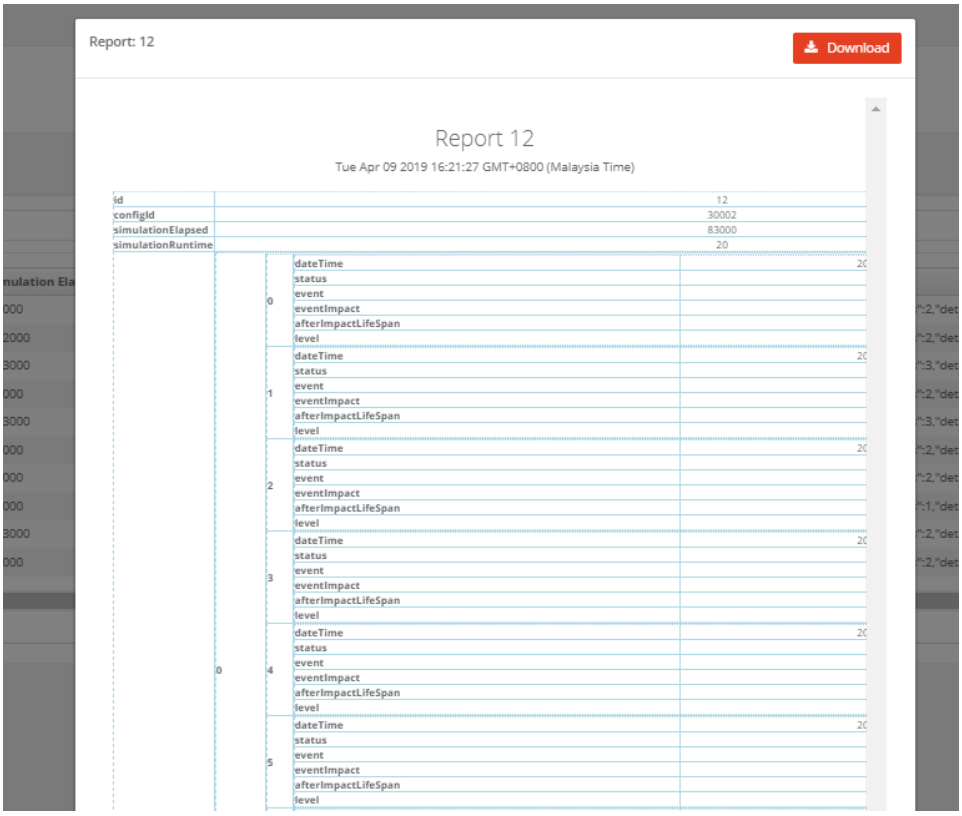


Fig. 14. The Simulation Report.

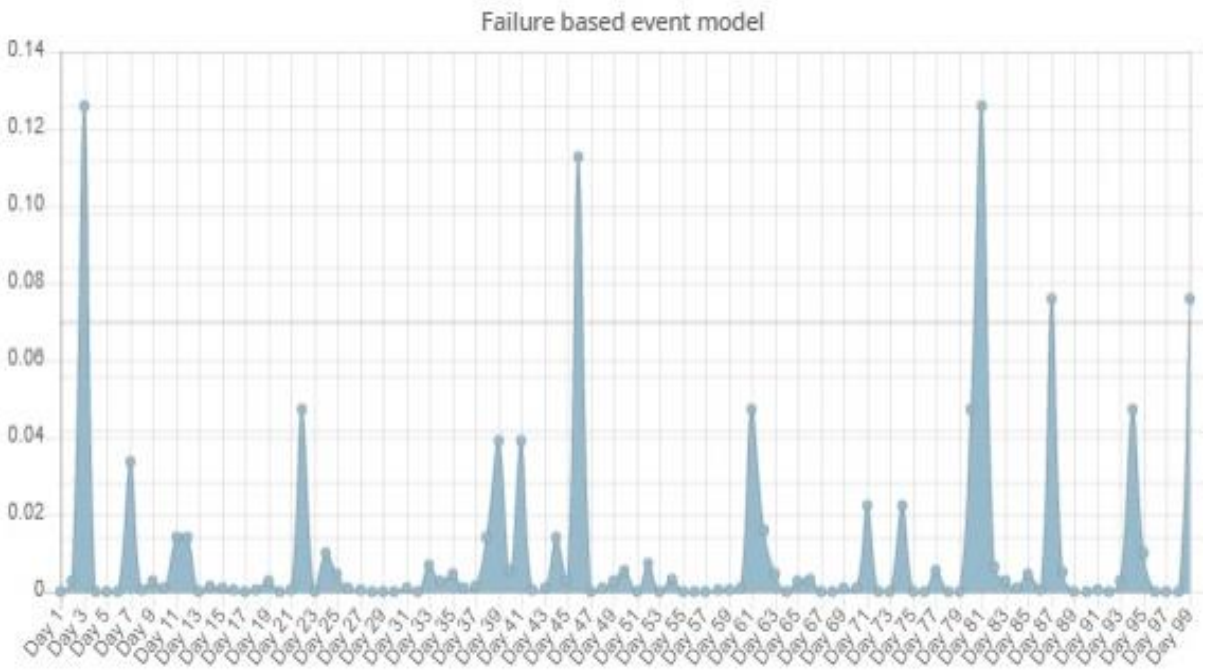


Fig. 15. Failure Prediction.

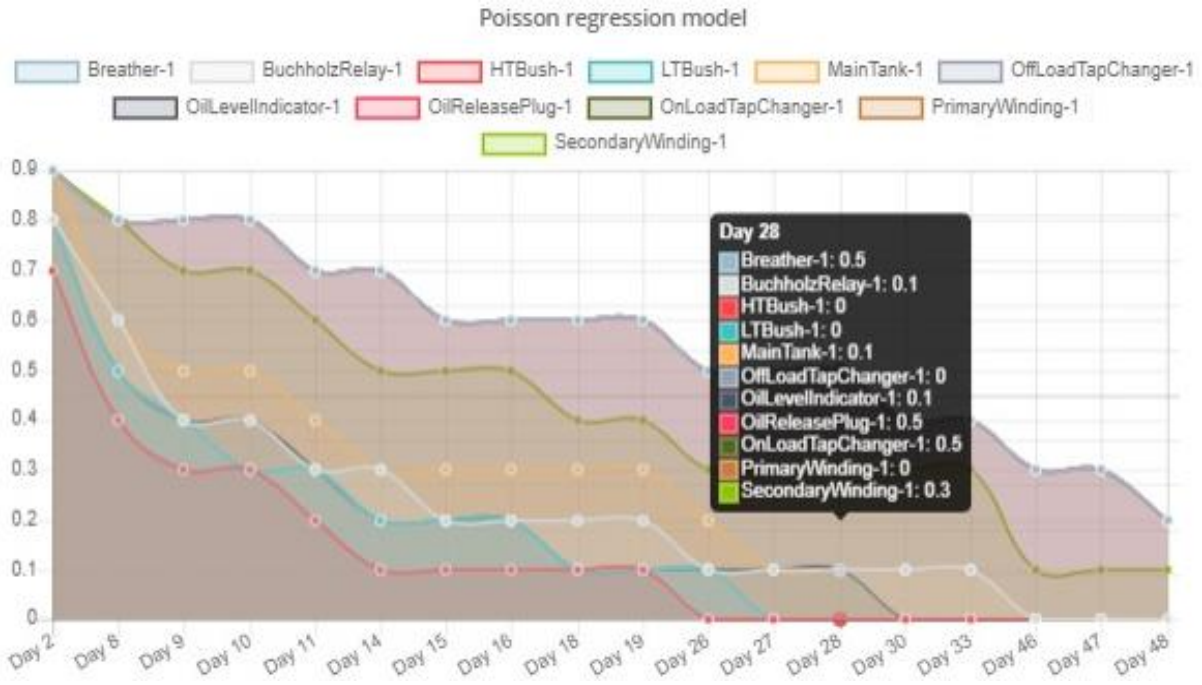


Fig. 16. Status of Substation Components.

V. Conclusion

In this paper, a simulation tool to predict failure and subsequently generate a maintenance report for a particular power substation configuration is presented. The proposed system consists of three components. The developed evaluation system consists of three main components which are Configuration Generator Tool (GCT), Failure Prediction Model (FDM), and Hybrid Simulation Platform (HSP). The GCT was used to generate alternative configurations and developed using Web Ontology Language (OWL). While the HSP is developed using several simulations and ontology-based tools such as HTML5, Javascript, Apache Jena, Unity Engine, C# Programming Language, Visual Studio 2017 IDE, and Blender 3D modeling. The FDM is developed based on the impact and relationship of weather factors (e.g. rain, wind, lightning) to power grid components and the impact of a failed component to other components. The results showed a powerful tool to predict failure with 3D simulation, however, further validation needs to be conducted to tune the prediction model using concrete instances and this is left for future work.

Acknowledgment

This project is sponsored by Universiti Tenaga Nasional (UNITEN) under the Bold Research Grant Scheme No.RJO10517844/010.

References

[1] Shafiullah, M., Abido, M. A., & Abdel-Fattah, T. (2018). Distribution Grids Fault Location employing ST based Optimized Machine Learning Approach. Energies, 11(9), 2328.

[2] Lee H, Mousa AM. GPS travelling wave fault locator systems: investigation into the anomalous measurements related to lightning strikes. Power Deliv IEEE Trans on 1996;11:1214–23.

[3] Sant MT, Paithankar YG. Online digital fault locator for overhead transmission line. Electr Eng Proc Inst of 1979;126:1181–5.

[4] Purushothama GK, et al. ANN applications in fault locators. Int J Electr Power Energy Syst 2001;23:491–506, [8//].

[5] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97, [1995/ 09/01].

[6] Pradhan AK, et al. Higher order statistics-fuzzy integrated scheme for fault classification of a series-compensated transmission line. IEEE Trans Power Deliv 2004;19:891–3.

[7] Bedekar PP, et al. Fault section estimation in power system using Hebb's rule and continuous genetic algorithm. Int J Electr Power Energy Syst 2011;33:457–65, [3//].

[8] Mokhlis H, Li HY. Fault location estimation for distribution system using simulated voltage sags data. In: Proceedings of the universities power engineering conference, 2007. UPEC 2007. 42nd International, 2007, p. 242–7.

[9] Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, methods and applications. The knowledge engineering review, 11(2), 93-136.

[10] Schachinger, D., Kastner, W., & Gaida, S. (2016, April). Ontology-based abstraction layer for smart grid interaction in building energy management systems. In 2016 IEEE International Energy Conference (ENERGYCON) (pp. 1-6). IEEE.

[11] Mahmoud, M. A., Maseleno, A., Tang, A. Y., Lim, F. C., Kasim, H. B., & Yong, C. (2019, September). Analysis of the Publications on Ontology-Based Smart Grid Applications: A Bird's Eye View. In International Conference on Applied Computing to Support Industry: Innovation and Technology (pp. 491-502). Springer, Cham.

[12] Law, N. L. L. M. F., Mahmoud, M. A., Tang, A. Y. C., Lim, F. C., Kasim, H., Othman, M., & Yong, C. (2019). A review of ontology development aspects. International Journal of Advanced Computer Science and Applications, 10(7), 290-298.

[13] Mahmoud, M. A., Tang, A. Y., Maseleno, A., Lim, F. C., Kasim, H., & Yong, C. (2019, September). Towards the Development of a Smart Energy Grid. In International Conference of Reliable Information and Communication Technology (pp. 673-682). Springer, Cham

[14] Maseleno, A., Hashim, W., Tang, A. Y., Mahmoud, M. A., & Othman, M. (2020). A Brief Understanding on Smart Grid Technology. Journal of Computational and Theoretical Nanoscience, 17(6), 2866-2868.

[15] Ullah, I., Yang, F., Khan, R., Liu, L., Yang, H., Gao, B., & Sun, K. (2017). Predictive maintenance of power substation equipment by infrared thermography using a machine-learning approach. Energies, 10(12), 1987.

[16] Bhattacharya, B., & Sinha, A. (2017, November). Intelligent fault analysis in electrical power grids. In 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 985-990). IEEE.

[17] Zhang, S., Wang, Y., Liu, M., & Bao, Z. (2017). Data-based line trip fault prediction in power systems using LSTM networks and SVM. IEEE Access, 6, 7675-7686.

[18] Novoselnik, B., Bolfek, M., Bošković, M., & Baotić, M. (2017). Electrical power distribution system reconfiguration: Case study of a real-life grid in croatia. IFAC-PapersOnLine, 50(1), 61-66.

[19] Lawrence, T. M., Watson, R. T., Boudreau, M. C., & Mohammadpour, J. (2017). Data flow requirements for integrating smart buildings and a smart grid through model predictive control. Procedia engineering, 180, 1402-1412.

[20] Tang, A. Y. C., Mahmoud, M. A., Lim, F. C., Kasim, H. (2020) Description for Smart Grid: Towards the Ontological Approach, The 8th International Conference on Information Technology and Multimedia (ICIMμ 2020).

[21] Tang, A. Y. C., Mahmoud, M. A., Lim, F. C., Kasim, H. (2020) A review of Smart Grid Technology, Components, and Implementation, The 8th International Conference on Information Technology and Multimedia (ICIMμ 2020).

[22] Yang, S., Zhou, W., Zhu, S., Wang, L., Ye, L., Xia, X., & Li, H. (2019). Failure probability estimation of overhead transmission lines considering the spatial and temporal variation in severe weather. Journal of Modern Power Systems and Clean Energy, 7(1), 131-138.

[23] Mahmoud MA., Ramli R., Azman F, and Grace J. (2018). A Development Methodology Framework of Smart Manufacturing Systems (Industry 4.0), MySEC 2018.

[24] Ahmed M, Ahmad MS, Yusoff MZ. Modeling agent-based collaborative process. InInternational Conference on Computational Collective Intelligence 2010 Nov 10 (pp. 296-305). Springer, Berlin, Heidelberg.

[25] Mahmoud, M., Ahmad, M. S., Mostafa, S., & Subramainan, L. (2020). How New Individuals Behave in a Heterogeneous Community: A Computational Approach to Norm Assimilation Using Agent-Based Systems. Journal of Systems Science and Complexity, 33(4), 849-881.

[26] Mahmoud MA, Ahmad MS, Yusoff MZ. A norm assimilation approach for multi-agent systems in heterogeneous communities. InAsian Conference on Intelligent Information and Database Systems 2016 Mar 14 (pp. 354-363). Springer, Berlin, Heidelberg.

[27] Mahmoud MA, Ahmad MS, Yusoff MZ. Development and implementation of a technique for norms-adaptable agents in open multi-agent communities. Journal of Systems Science and Complexity. 2016 Dec 1;29(6):1519-37.

Home Security System with Face Recognition based on Convolutional Neural Network

Nourman S. Irjanto[1], Nico Surantha[2]
Computer Science Department, BINUS Graduate Program - Master of Computer Science
Bina Nusantara University, Jakarta

***Abstract*—Security of house doors is very important and becomes the basis for the simplest and easiest security and sufficient to provide a sense of security to homeowners and along with technological developments, especially in the IoT field, which makes technological developments in locking house doors have developed a lot like locking house doors with faces and others. The development of facial recognition systems has also developed and has been implemented for home door locking systems and is an option that is quite simple and easy to use and is quite accurate in recognizing the face of homeowners. The development of the CNN method in facial recognition has become one of the face recognition systems that are easy to implement and have good accuracy in recognizing faces and has been used in object recognition systems and others. In this study, using the CNN Alexnet facial recognition system which is implemented in a door locking system, data collection is done by collecting 1048 facial data on the face of the homeowner using a system which is then used to train machine learning where the results are quite accurate where the accuracy is the result is 97.5% which is quite good compared to some other studies. The conclusion is the CNN Alexnet method can perform facial recognition which is quite accurate which can be implemented on the IoT device, namely, the Raspberry Pi.**

Keyword—Home door security; CNN Alexnet; facial recognition; Raspberry Pi

I. Introduction

Over the past few years, there have been quite a several choices in conventional technology and biometric technology to meet security needs for households or offices. Some conventional security systems, for example using keys, passcodes, ID cards, and/or RFID cards, can be unreliable if objects for access are stolen or lost [1]. Such security systems have disadvantages when access is stolen by people who do not have the authority to gain access and also daily activities sometimes force someone to leave the house empty, such as during work or school hours. This makes the house vulnerable to break into and theft, even when the house is locked or securely locked. The development of Information Technology and Communication currently offers convenience to users in various lines of life. One technology that is currently trending is the smart home or what is commonly known as the smart home. A smart home is a term used to define a residence that has the equipment, lighting, heating, air conditioning, TV, computer, audio system, video entertainment, security, and camera systems that can communicate with each other and can be controlled remotely with a timetable. through the internet or telephone [2]. Biometric systems are developing rapidly, especially for home security technology because they can fulfill two functions, namely identification, and verification, biometrics have characteristics that cannot be lost, cannot be forgotten, and cannot be faked where their inherent presence in humans will differ between humans and other humans so that their uniqueness is guaranteed [3]. In the journal [4] facial recognition as authentication is very good because the face is a physiological feature that is easiest to distinguish between individuals so face recognition is one of the biometrics technologies that are often studied and developed.

Convolutional Neural Networks combines three basic architectures, namely local receptive fields, shared weight in the form of filters, and spatial subsampling in the form of pooling. Convolution or what is commonly known as convolution is a matrix that functions to perform filters[5]. In the filtering process, there are two matrices, namely the input value matrix and the kernel matrix. In the Convolutional Neural Network, several layers function to carry out the filters that have been determined during the training process, namely Convolutional Layer, Pooling Layer, and Fully Connected Layer [6]. The architecture that is owned by the Convolutional Neural Network can be seen in Fig. 1.

The previous paper described a prototype of a safe room access control system based on facial recognition. This system consists of a webcam to detect faces and a solenoid door lock to access the room. Every user detected by the webcam will be checked for compatibility with the database on the system using the Haar cascade classifier method embedded in OpenCV. If the user has access rights, the solenoid door lock will open and the user can enter the room. In this paper, the Haar cascade classifier embedded in OpenCV can recognize multiple captured images [7]. Another project is designing facial recognition systems for smart home/office security applications. The design is implemented using a webcam and programmed using dlib and OpenCV. The connection between the cam and the computer can be made by cable and wireless. We'll be using a very simple approach to dealing with recognition using deep learning [8] and also in other research journals aimed at designing a door security system that uses Arduino as a microcontroller and utilizes open source OpenCV as a face reader where this research reads faces that have been entered into the database which will then match the images captured by the webcam. where the results of the accuracy measurement based on the test table carried out three times get a success rate of 71.40%, 85.71%, 71.42% [9] and In another study, a door security system was developed using facial recognition as a key to open doors. The method used in

this tool is the fisherface method. The main steps in facial recognition are face detection, PCA calculation, FLD calculation. where the measurement results of the accuracy of the system are 80% [10] and also in other research is an effort to develop assistance to maintain security in important places. We used the Viola-Jones algorithm to detect faces and the Eigenfaces algorithm to recognize people. The test results were recorded and we achieved 95% accuracy in recognition under fluorescent lighting conditions [11]. In this paper, we construct a face recognition system. In this work, we present the advantages and disadvantages of different techniques in a literature survey. It helps to choose a suitable technique among many as per our application requirements and solve current problems to some extent for real-time applications. We achieve 96.8% accuracy in real-time scenarios under many variations and seamless environments and also measure performance using the Multi-task Cascaded Convolutional Networks (MTCNN) method [12].

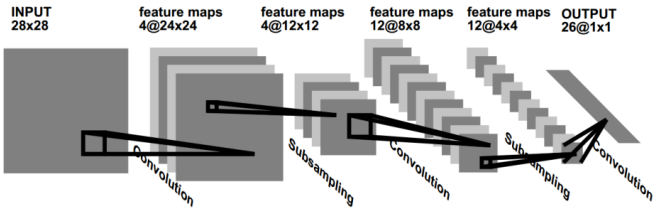


Fig. 1. Convolutional Neural Network Architecture.

Based on the existing cases, a new system must be devised to prevent house burglaries and thefts due to the weak security of the lock or padlock. So that the idea of facial recognition-based door security system innovation using the convolutional neural network (CNN)[13] method was created, of course, it has better security than locks or RFID. It can be said that this system is an automatic electronic lock. This system is expected to be able to tackle the occurrence of theft in houses that are often abandoned by the occupants.

This study expects a significant contribution to a new domain of knowledge regarding the application of accurate facial recognition technology to the home door locking systems. Therefore, this research is an attempt to build a facial recognition system that can work on house doors.

II. Method

In this study, we propose a facial recognition process for the process of opening the door of a house that can replace the process of home security using an electronic key or RFID, where the research stages are divided into 3 parts, namely the stages of collecting homeowner data, the data training process, and also the facial recognition process using Raspberry Pi. In this journal, we implement the facial recognition process with the CNN [14] method which will install it on a mini-computer, namely the raspberry pi which will serve as a microcontroller to lock and open the door automatically which is controlled by the face of the homeowner [15].

A. Homeowner Face Data Collection

The stages of data collection are carried out manually, namely by using a program designed to collect facial data from each homeowner consisting of 5 people where the total data is 1100 data which will then be divided 1040 for training data and 60 data is used for validation during training by doing the facial augmentation process starts from shifting 10-15 degrees with various expressions [16]. The results of data collection can be seen in Fig. 2 and 3.

B. Training Model

At this stage the training process is not carried out on the raspberry pi due to the small computation of the raspberry pi with that the training process is carried out on a separate computer with Intel Core i5 8500 Processor specifications and 8GB DDR4 RAM where this training process will also form a model that will be used to detect the face [17]. The stages of the training process use the CNN Alexnet method with two convolution processes and two pooling processes and softmax with several iterations of 20 times with the parameters shown in Fig. 4.

C. System Implementation

This prototype will be made by connecting the modified Pi Camera as a camera module to identify the face of the homeowner connected to the Raspberry Pi 3 Model B + where the Raspberry Pi will be connected via WLAN as a process of identifying the homeowner[18] as seen in Fig. 5.

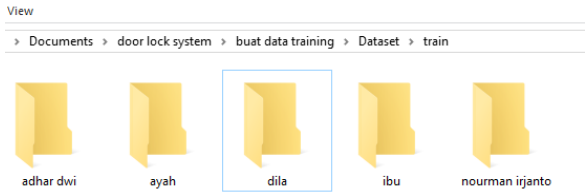


Fig. 2. Dataset of Homeowners Faces.

Fig. 3. Dataset of Face.

```
"""
Parameters
"""
img_width, img_height = 224, 224
batch_size = 32
samples_per_epoch = 1000
validation_steps = 300
nb_filters1 = 32
nb_filters2 = 64
conv1_size = 3
conv2_size = 2
pool_size = 2
classes_num = 5
lr = 0.0004
```

Fig. 4. Parameter Training Method.

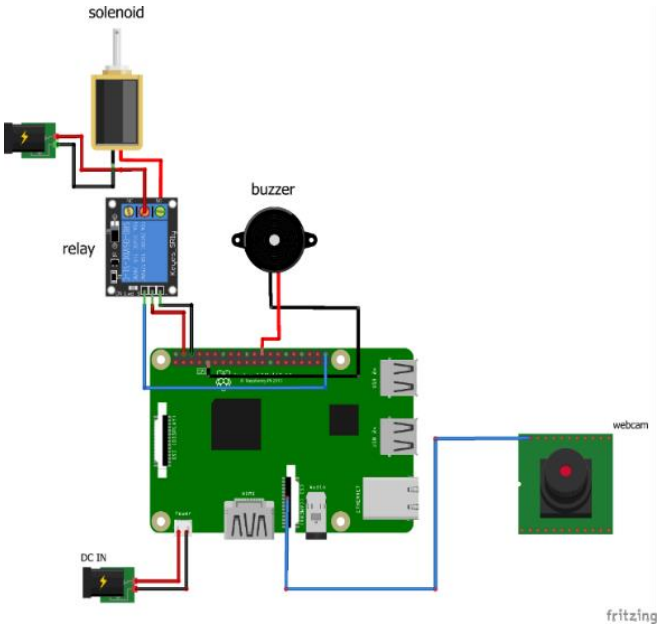


Fig. 5. System Design.

1) Flowchart system: The workflow of this system is divided into two parts, namely the registration stage where at this stage the data generated will be used as training data [19]. At this stage, there will be a registration process for the face data of the homeowner who will be trained on the computer to produce a training model which will be stored in the database on the Raspberry Pi and will be backed up and if it is already the device will be standby and ready to use as in Fig. 6.

The system installation process is carried out at the front door of the house which is the only entrance to the existing house as seen in Fig. 7 and Fig. 8.

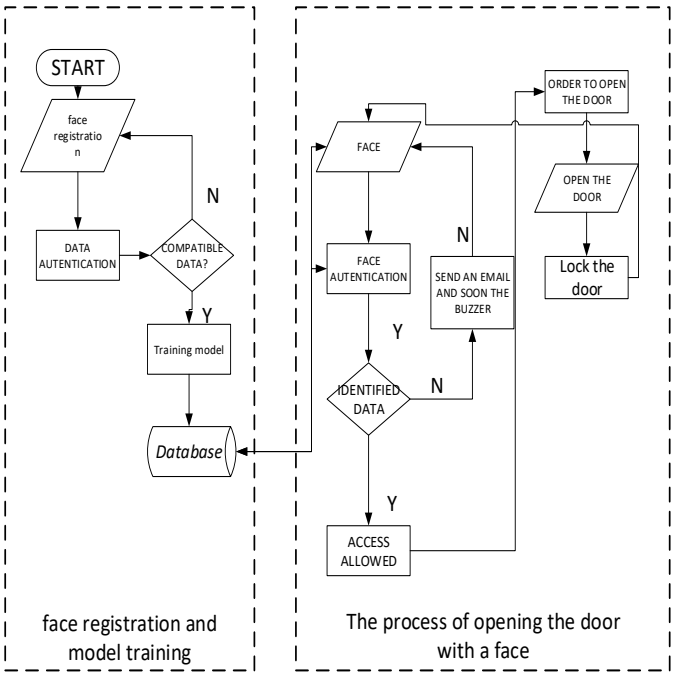


Fig. 6. Flowchart.

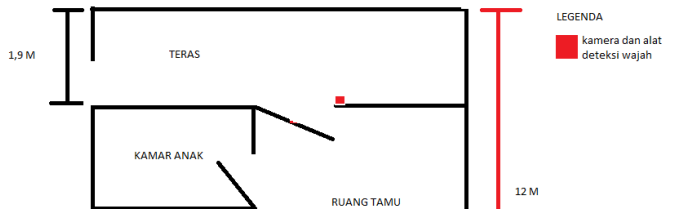


Fig. 7. Test Area.

Fig. 8. Device Placement.

III. Testing and Comparison

Testing is carried out by providing input in the form of five homeowners and also five faces of non-homeowners or neighbors who have been tested in morning conditions with a duration of 07.00 to 09.00, noon 12.00 to 14.00, noon 15.00 to 17.00, and night 19.00 to 22.00 with a distance 0.5 meters, 1 meter, and 1.5 meters, respectively and the distance from the front door [9].

A. Homeowner Testing

Testing of homeowners using the system built can be seen in Table I.

B. Testing is not a Homeowner or Neighbor

The non-homeowner test is carried out with the same conditions as the home owner's condition, namely in the morning, afternoon, evening, and night, which is shown in Table II.

C. Latency Testing

Latency testing is done by measuring the time it takes for the system to perform a face reading, the calculation process starts when the system is on standby until the solenoid functions and the door opens until the door closes again [20]. The test was conducted 20 times, 10 times for homeowners and 10 times for non-homeowners, where the time taken to take the average reading was 5.90 seconds, as shown in Fig. 9.

D. Comparison with other Studies

After being reviewed from previous journals, namely in research [11] using a dataset from AT&T Laboratories Cambridge face dataset, the training process using 400 negative images produces 95% accuracy and with the same dataset in this study, this study tests the accuracy of this method using this dataset with simulation results The same test produces better accuracy results where the accuracy value obtained is 97.83% which can be seen in Table III.

TABLE I. HOMEOWNER TESTING SAMPLES

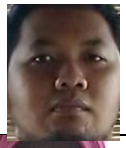
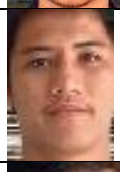
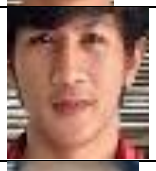
People	Face
	
	
	
	
	

TABLE II. NEIGHBOR'S FACE TESTING SAMPLE

People	Face
	
	
	
	
	

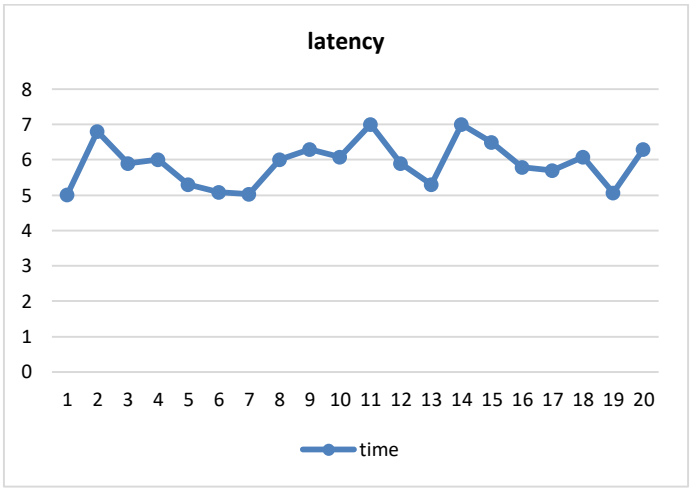


Fig. 9. Latency Testing.

TABLE III. COMPARISON WITH OTHER STUDIES

No	Paper	Method	Accuracy
1.	[11]	OpenCV	95%
2.	propose method	CNN	97.83%

E. Result

After carrying out the above tests, it resulted in a significant development both in terms of the accuracy of the image reading which has increased, and the processing is quite faster. Complete data can be seen in Table IV and Table V.

After testing four times at three different times and based on Table IV and Table V, it can be concluded that this system is running well which has test results in the morning, afternoon, evening, and night with three conditions and it can be concluded in Table VI.

TABLE IV. HOMEOWNER TEST RESULTS

Time	Face	Distance					
		1.5 m		1 m		0.5 m	
		Success	Fail	Success	Fail	Success	Fail
Morning	5	4	1	5	0	5	0
Afternoon	5	4	1	5	0	5	0
Evening	5	4	1	5	0	5	0
Night	5	5	0	5	0	5	0

TABLE V. TEST RESULTS ARE NOT HOMEOWNERS

Time	Face	Distance					
		1.5 m		1 m		0.5 m	
		Success	Fail	Success	Fail	Success	Fail
Morning	5	0	5	0	5	0	5
Afternoon	5	0	5	0	5	0	5
Evening	5	0	5	0	5	0	5
Night	5	0	5	0	5	0	5

TABLE VI. CALCULATION OF ACCURACY VALUE

Total sample: 120		Prediction	
		Negative	Positives
actual	Negative Positive	TN : 60 FN : 0	TP : 57 FP : 3
Akurasi: (TP+TN)/(TP+TN+FP+FN) = 0.975			

IV. CONCLUSION

The research was carried out in three distances, namely, 1.5 meters, 1 meter, and 0.5 meters, and carried out at four times, namely, morning, afternoon, evening, and night, where there was an error three times, namely, at a distance of 1.5 meters where there was excessive light on the background of the standing place. resulting in unclear images, and this research has used a method to increase the accuracy of facial recognition which can reach an accuracy of 97.5%. And also after comparisons with the proprietary OpenCV method [11] using the same dataset and testing stages, this research is a little better, producing an accuracy of 97.83% wherein in the previous research, 95% accuracy was obtained. Further research includes optimizing the facial data augmentation process used as a dataset, better camera resolution, and using the latest Raspberry Pi model to improve computing capabilities.

ACKNOWLEDGEMENT

The publication of this research is supported by Bina Nusantara University.

REFERENCES

[1] Y. D. S. V. D, A. Rakhmansyah, and N. A. Suwastika, "Implementasi Sistem Kunci Pintu Otomatis Untuk Smart Home Menggunakan SMS Gateway," e-Proceeding Eng., vol. 2, no. 2, pp. 6395–6407, 2015.

[2] A. Siswanto, A. Efendi, and A. Yulianti, "Alat Kontrol Akses Pintu Rumah Dengan Teknologi Sidik Jari Di Lingkungan Rumah Pintar Dengan Data Yang Di Enkripsi," J. Penelit. Pos dan Inform., vol. 8, no. 2, p. 97, 2019.

[3] A. Yudhana, "Perancangan pengaman pintu rumah berbasis sidik jari menggunakan metode uml," (Jurnal Teknol. Informasi) Sist. PENGGAJIAN KARYAWAN PADA LKP GRACE Educ. Cent., vol. Vol.1, No., no. 2, p. 12, 2018.

[4] B. Septian, A. Wijayanto, F. Utaminingrum, and I. Arwani, "Face Recognition Untuk Sistem Pengaman Rumah Menggunakan Metode HOG dan KNN Berbasis Embedded," Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 3, no. 3, pp. 2774–2781, 2019.

[5] M. Yan, M. Zhao, Z. Xu, Q. Zhang, G. Wang, and Z. Su, "VarGFaceNet: An Efficient Variable Group Convolutional Neural Network for Lightweight Face Recognition," Iccvw 2019, pp. 2647–2654, 2019.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Adv. Neural Inf. Process. Syst., vol. 2, pp. 1097–1105, 2012.

[7] A. Najmurrokhman, K. Kusnandar, A. B. Krama, E. C. Djamal, and R. Rahim, "Development of a secured room access system based on face recognition using Raspberry Pi and Android based smartphone," MATEC Web Conf., vol. 197, pp. 1–6, 2018.

[8] R. A. Isaac, A. Agarwal, and P. Singh, "Face Recognition Security Module using Deep Learning," J. Netw. Commun. Emerg. Technol., vol. 8, no. 10, pp. 10–13, 2018.

[9] J. Nasir and A. A. Ramli, "Design of Door Security System Based on Face Recognition with Arduino," vol. 3, no. 1, pp. 127–131, 2019.

[10] B. M. Susanto, F. E. Purnomo, and M. F. I. Fahmi, "Sistem Keamanan Pintu Berbasis Pengenalan Wajah Menggunakan Metode Fisherface Security System Based On Face Recognition Using Fisherface Method," J. Ilm. Inov., vol. 17, no. 1, p. 10, 2017.

[11] F. Faisal and S. A. Hossain, "Smart security system using face recognition on raspberry Pi," 2019 13th Int. Conf. Software, Knowledge, Inf. Manag. Appl. Ski. 2019, no. August, 2019.

[12] R. Singh, M. Singh, and L. Ragha, "Real-time Face Recognition Under Different Environment," SSRN Electron. J., 2019.

[13] M. F. A. Hassan, A. Hussain, M. H. Muhamad, and Y. Yusof, "Convolution neural network-based action recognition for fall event detection," Int. J. Adv. Trends Comput. Sci. Eng., vol. 8, no. 1.6 Special Issue, 2019.

[14] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An All-In-One Convolutional Neural Network for Face Analysis," Proc. - 12th IEEE Int. Conf. Autom. Face Gesture Recognition, FG 2017 - 1st Int. Work. Adapt. Shot Learn. Gesture Underst. Prod. ASL4GUP 2017, Biometrics Wild, Bwild 2017, Heteroge, pp. 17–24, 2017.

[15] Soe Sandar | Saw Aung Nyein Oo, "Development of a Secured Door Lock System Based on Face Recognition using Raspberry Pi and GSM Module," Int. J. Trend Sci. Res. Dev., vol. 3, no. 5, pp. 357–361, 2019.

[16] N. A. Al-Johania and L. A. Elrefaei, "Dorsal hand vein recognition by convolutional neural networks: Feature learning and transfer learning approaches," Int. J. Intell. Eng. Syst., vol. 12, no. 3, 2019.

[17] N. A. Muhammad, A. A. Nasir, Z. Ibrahim, and N. Sabri, "Evaluation of CNN, alexnet and GoogleNet for fruit recognition," Indones. J. Electr. Eng. Comput. Sci., vol. 12, no. 2, pp. 468–475, 2018.

[18] P. Barsocchi, A. Calabrò, E. Ferro, C. Gennaro, E. Marchetti, and C. Vairo, "Boosting a low-cost smart home environment with usage and access control rules," Sensors (Switzerland), vol. 18, no. 6, 2018.

[19] A. R. Syafeeza, M. K. Mohd Fitri Alif, Y. Nursyifaa Athirah, A. S. Jaafar, A. H. Norihan, and M. S. Saleha, "IoT based facial recognition door access control home security system using raspberry pi," Int. J. Power Electron. Drive Syst., vol. 11, no. 1, pp. 417–424, 2020.

[20] N. Surantha and W. R. Wicaksono, "An IoT based house intruder detection and alert system using histogram of oriented gradients," J. Comput. Sci., vol. 15, no. 8, pp. 1108–1122, 2019.

The Relationship of Trustworthiness and Ethical Value in the Healthcare System

Rajes Khana[1], Manmeet Mahinderjit Singh[2], Faten Damanhoori[3], Norlia Mustaffa[4]
School of Computer Science, Universiti Sains Malaysia, Penang-11800, Malaysia

***Abstract*—Females prefer discovering social media or healthcare systems to finding information and presenting their cases with any physician; however, the behavior of physicians tends to be uncontrollable on the healthcare system. Physicians have the capacity to share all of their patients' information with their colleagues without any permission or concern from the patients. For this reason, it is of utmost importance to design a breast self-examination system that can keep monthly track of self-exam data and communication between patient and physician. To develop such a system, identify the ethical values and trustworthiness as an indicator. Then, the survey will provide the details on ethical values and trustworthiness applicable in the system. Therefore, this research objective on the importance of ethical value and trustworthiness in the healthcare system. The survey on 772 respondents leading to the importance of the ethical value being used in the healthcare system is required. The ethical value of interaction, integrity, confidentiality, protection, caring, and fairness have a significant influence on the healthcare system. The path coefficients are answering Hypothesis I in presenting the positive relationship and significant effect between ethical value and BSE system (P<.001). On the other side, trustworthiness has a significant influence on the healthcare system. The path coefficients are answering Hypothesis II in presenting the positive relationship and significant effect between trustworthiness and the BSE system (P<.001). Finally, the relationship in healthcare between trustworthiness and ethical value is on integrity with honesty and belief.**

***Keywords*—Ethics; ethical value; trustworthiness; breast self-examination; healthcare system; social media**

I. Introduction

Breast cancer mortality cases are growing every year, and it becomes the number one cancer cause for females [1]. In the country of the United States, India, Malaysia, and Indonesia, practicing breast self-examination in reducing breast cancer mortality [2]–[5].

The trend of using the internet and social media is growing every year. People are connected to over 3.5 billion using the internet and social media [6]. The demand for using social media brings healthcare to become the most important area to discuss [7], [8]. People are curious to get information on a particular disease on social media before meeting the physician. They prefer to looking for information from social media [7]. Many users in social media who were suffering from illnesses such as AIDS, breast cancer, or recovered from sexual abuse used social media as a convenient venue to discuss [9]–[11]. At the same time, physicians use social media for assisting, treating, and consulting with patients who are suffering from cancer [7], [12], [13]. Social media naturally creates an opportunity for the unethical person for accessing any person's private information and confidential information related to any disease while communicating with a physician [12]. However, the violation increasing on patient confidentiality in social media [14]–[18] and patients are lack trust in the physician's conduct [12], [19]–[21]. On the other hand, physicians (65.8%, 187) are reluctant to use social media fully due to worries of protection on public access [13]. Therefore, the trust between patient and physician is needed [9] as well as ethical value should implement into the healthcare system [17].

The aims of this paper are the relationship between trustworthiness and ethical value on the interaction process between patient and physician in the use of the healthcare system. The objective is the importance of ethical value and trustworthiness in the healthcare system.

This study will describe the literature review, methodology, results, discussion, conclusion, and future challenges.

II. Literature Review

This section contains a brief discussion of the theory of ethics, the theory of trust, and the healthcare system.

A. Principles of Ethics on Healthcare

Belmont Report [22] introduces three principles of biomedical ethics; respect for persons, beneficence, and justice as mentioned in Table I. Follow by, Beauchamp & Childress[23] bring four principles of biomedical ethics (Table I). There are:

1) Respect for autonomy, every person has their right to make their own life choices and free from any influences. Every individual should be able, to tell the truth, respected others' privacy, and the ability to protect confidential information.

2) Beneficence, every person should be able to respect their own decision and protecting themselves from harm. It involves secure patient welfare and promoting good.

3) Non-maleficence, as a person we are supposed not to hurt anybody, do not cause pain or suffer to others, and do not harm.

4) Justice, a person should be able to do services to other people fairly. They are serving two different persons in equal treatment. The justification of treatment should be equally the same among persons.

B. Theory of Ethics

In the common term, ethics is part of accepting and analyzing moral life [23]. Philosophically, Ethics is "the branch of philosophy that explores the conduct of human character and human values rationally" [17]. The human character and human values were identified as what is good, bad, right, or wrong in a moral sense [25]. Every human conduct will generate an ethical value.

Ethical value is objects with value or as part of the human conduct being approval or disapproval [17], [26]. According to Khana et al. [17], ethical values consist of six values such as interaction, integrity, confidentiality, caring, and fairness. Those values carry 19 indicators such as communication, sharing information, clinician judgment, informed consent, patient interest as the priority, expert advice, honest information, respect, anonymity, responsibility, improving quality, clinical result, refrain harm, de-identification, safeguard, look up information, engagement, optimal care, and inequalities [12], [17], [27], [28]. In this study, ethical value identify as independent variable 1 (Table II).

C. Analysis Relationship between Biomedical Ethics and Ethical Value

The relationship between biomedical ethics and ethical value is presented in the Venn diagram in Fig. 1. The first principles of respect for autonomy carry, to tell the truth, respect privacy, and confidential information. Those indicators are related to the ethical value of integrity (respect), integrity (honest information), and confidentiality (clinical result). Tell the truth is the action of honesty which is related to honest information. Respect the privacy is related to respect in ethical value. Whereas confidential information is related to the confidentiality of clinical result information.

The second principle beneficence carries protect and defend the rights of others, prevent harm from occurring others, help persons with disabilities, and maximize possible benefits. Those indicators are related to the ethical value of protection and caring. Protect and defend the rights of others is related to protection (safeguard). Protection is the act of protecting somebody or something [29]. Prevent harm from occurring to others is related to protection (refrain harm). Refrain from harm is the abstain of doing harm to others. Especially physicians do not cause harm to the patients [17]. They help persons with disabilities related to caring (optimal care). The optimal care of any patient is essential for every physician [27]. Maximize possible benefits related to interaction (sharing information). A physician should share health-related information to the patient/public [12] to provide maximum benefits to them.

TABLE I. PRINCIPLES OF BIOMEDICAL ETHICS

Belmont Report (1974)	**Beauchamp & Childress (1979)**	**Moral Rules**
Respect for Persons	Respect for Autonomy	a. Tell the truth (B&C)
		b. Respect the privacy of others (B&C)
		c. Protect Confidential Information (B&C)
		d. When asked, help others make important decisions (B&C)
Beneficence	Beneficence	a. Protect and defend the rights of others (B&C)
		b. Prevent harm from occurring to others (B&C)
		c. Remove conditions that will cause harm to others (B&C)
		d. Help persons with disabilities (B&C)
		e. Rescue persons in danger (B&C)
		f. Maximize possible benefits (BR)
		g. Minimize possible harms (BR)
	Non-maleficence	a. Do not kill (Gert)
		b. Do not cause pain or suffering (Gert)
		c. Do not incapacitate (Gert)
		d. Do not cause offense (Gert)
		e. Do not deprive others of the goods of life (Gert)
		f. Do not harm (BR)
Justice	Justice	a. Everyone gets an equal share/Fair (BR)
		b. Distribution according to need (BR)
		c. According to individual effort (BR)
		d. According to societal contribution (BR)
		e. According to merit (BR)
		f. To each person according to free-market exchanges (B&C)

Noted: B&C = Beauchamp & Childress (2009)[23], BR = Belmont Report (1974)[22], Gert = Gert (2005)[24]

TABLE II. ETHICAL VALUE VARIABLE ALONG WITH THE INDICATORS

Variable	Sub-variables	Indicators	Definition
Ethical Value (EV)	Interaction (INT)	Communication (INT1)	Communication is the activity of expressing feelings and ideas or providing information to another person [29]. In healthcare, communication is the way of conveying a self-condition to the physician and vice versa[17].
		Sharing Information (INT2)	Sharing information is the process of the physician to deliver medical information to the public[12]
	Integrity (ITG)	Patient interest as the priority (ITG1)	The patient is critical in the healthcare world. They put patients as the highest priority to compare with a physician [12]. Patient interest as the priority is the professional integrity of physician that bring patient as the highest priority[23]
		Expert Advice (ITG2)	Expert advice is a physician's capability to explain healthcare matters to society. It is part of the trust given to the physician or any healthcare professional[12].
		Honest Information (ITG3)	A physician being honest to all his/her patient in relation to fees, promotion of any product, and any conflict of interest[12].
		Respect(ITG4)	Respect is part of respect for a person which is as a basis of the moral norm[23]
		Responsibility (ITG5)	Responsibility is part of the education given by the physician on a particular disease[12], [27]
		Improving Quality (ITG6)	To maintain the quality of clinical care and healthcare system[27]
		Clinician Judgment (ITG7)	Clinician judgment is the best of clinician outcome on patient health interest[27]
	Confidentiality (CFT)	Informed consent (CFT1)	Informed consent is the consent of a person to undertake a medical procedure or any other information[30].
		Anonymity (CFT2)	Anonymity is an unknown person accessing a system without any identification[31], [32]
		Clinical Result (CFT3)	The clinical result is the diagnosis data or historical data of the patient based on the examination process[27]
		De-identification (CFT4)	De-identification is a process of removing identifiers from health information and mitigating privacy risks to individuals[33]
		Look up information (CFT5)	To look up information is to find information on a particular patient's treatment on the internet. It is part of the patient's privacy that is openly accessed in the public environment. It will create a violation and a compromise of trust[12].
	Protection (PRT)	Refrain harm (PRT1)	Harmful is the condition of causing harm to other people, such as posting unprofessional content[17].
		Safeguard (PRT2)	Safeguard is the system that used a secure closed system with data encryption[12]
	Caring (CRG)	Engagement (CRG1)	Engagement is the approach of the physician to convince the patient[12].
		Optimal care (CRG2)	Optimal care is the maximum effort of the physician to take care of the patient until the clinical result/outcome appears[27].
	Fairness (FRS)	Inequalities (FRS)	Inequalities are incomplete of medical evidence for physician decisions on patient treatment[27].

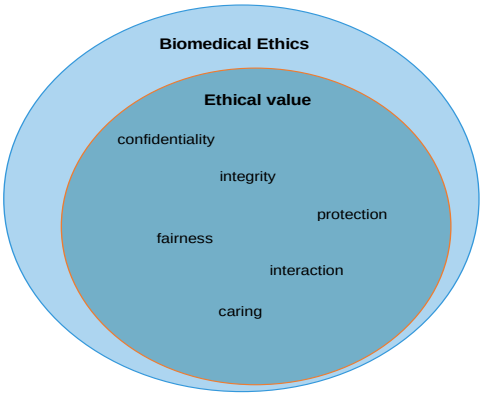


Fig. 1. Venn Diagram of The relationship between Biomedical Ethics with Ethical Value.

The third principle non-maleficence carries do not harm and do not cause pain. Those indicators are related to the ethical value of protection (refrain harm). Do not harm and do not cause pain is related to refrain harm of ethical value. Refrain harm is the action of abstaining from harm to others [17].

The fourth principle justice carries everyone gets an equal share/fair and distribution according to need. Those indicators are related to the ethical value of fairness. Every patient essentially must get equal treatment and avoid inequalities treatment [27].

D. Theory of Trust

Trust is "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" [34]. The trust definition applied to the relationship with another identifiable party that perceived to act and react based on willingness toward the trustor [34], [35]. Whereas, trustworthiness is the ability to be relied on as honest or truthful [29]. Trustworthiness also is considered as a virtual synonym for character or virtue on honesty and integrity in the context of health care [23], and it is context-dependent and personalized [36].

Mayer and Davis had introduced the trust principle. There are factors relating to the trustor and trustee that lead to trust. The trustor characteristic is based on the propensity to trust. The propensity to trust is the general willingness to trust

others. At the same time, the trustee characteristic is based on trustworthiness. Trustworthiness is measured as the motivation to lie. For example, if a trustee will get something dishonest, he will be seen as less trustworthy [34].

Trust theory has moved to several areas of concern. This study brings trust concern on social media, especially to the healthcare area. Quinn et al. [37] introduce a personalized model of trustworthiness to cater to the internet environment and social media [36].

In this study, the enhancement of trustworthiness relationship in the healthcare system and its relationship with ethical value. The trustworthiness identifies as independent variable two which carries eight indicators such as honesty, reputation, competency, reliability, credibility, belief, confidence, and faith [36], [37], as mentioned in Table III.

E. Healthcare System

In this section, the healthcare system will define based on the breast self-examination system (BSE) due to the number of females who are suffering from breast cancer is growing. The estimated number on new cases 2,088,849 and mortality cases 626,679 [1], and around 87% of posts on Facebook consist of cancer cases [8].

BSE is a method of doing breast examination on early detection from any abnormal fear of cancer [41]. BSE is an independent regular self-diagnostic technique done by a woman to observe any suspicious and changes on her breast through the use of a mirror [42].

Patients prefer to access social media with the facility that able to make appointments, receive reminders, diagnostic test results, provide information about their health, and as a forum for asking general questions [15]. There are some features requested by patients as a reference to develop the BSE system such as user account, calendar, self-exam wizard, history, chat room, location, knowledge, video tutorial, and forum.

In this study, the BSE system [43] identifies as a dependent variable that carries nine indicators, as mentioned in Table IV.

F. Formulation of Research Hypothesis

After reviewed ethical value, trustworthiness, and the BSE system, the researcher is formulating the research hypothesis, as mentioned below.

H1: Ethical value will positively affect the BSE system.

H2: Trustworthiness will positively affect the BSE system.

TABLE III. TRUSTWORTHINESS VARIABLE ALONG WITH THE INDICATORS

Variable	Indicators	Definition
Trustworthiness (TW)	Honesty (TW1)	Honesty is one that makes good faith agreements, tells the truth, and fulfills any promises made[37].
	Reputation (TW2)	Reputation is part of the social notion of trust[38] or “an expectation about agent’s behavior based on information about the observation of its past behavior”[39].
	Competency (TW3)	Competency is the ability of one person to fulfill another person's needs[37].
	Reliability (TW4)	The quality of being trustworthy or of performing consistently well[29]
	Credibility (TW5)	The quality of being trusted and believed in[29]
	Belief (TW6)	Belief is justified and should be accepted(acceptable without argumentative support)[23]
	Confidence (TW7)	Confidence is ”a feeling of certainty or easiness regarding a belief one holds”[40].
	Faith(TW8)	Faith is the complete trust or confidence in someone or something[29].

TABLE IV. BSE SYSTEM VARIABLE ALONG WITH THE INDICATOR

Variable	Indicators	Definition
BSE System (BSE)	User Account (BSE1)	A user account is privileged access by a user for keeping personal information safe.
	Calendar (BSE2)	Calendar (Reminder system) for setting menstrual schedule as an alarm system.
	Self-Exam Wizard (BSE3)	Users are able to tap/sign/mark on the breast picture to plot the lesion area and share it with the physician. And It has the capability to take a photo when the lump appears on the breast surface.
	History (BSE4)	The function of this feature is to record all activity on breast self-examination.
	Chat room (BSE5)	The interaction or dialog privately between public/user and physician. User able to share her history data on self-exam
	Forum (BSE6)	The interaction or dialog publicly between public and physician
	Knowledge (BSE7)	The knowledge will provide information such as history, breast anatomy, breast cancer, diagnosis, breast self-exam, and treatment
	Location for Treatment(BSE8)	The user has the capability to get a selection of the nearest doctor for consultation or treatment. Physicians being informed by the patient for an appointment.
	Video Tutorial (BSE9)	This tutorial video will be presented visually shown on how to do the correct practice of BSE.

III. METHODOLOGY

The methods section will describe the research flow and research method and data collection.

A. Research Flow

The research flow in Fig. 2 describes the ethical phenomena in social media, supporting theory, identify the variable, the hypothesis of theoretical ethical framework, survey, validation proses, and finally ethical BSE system as the final outcome.

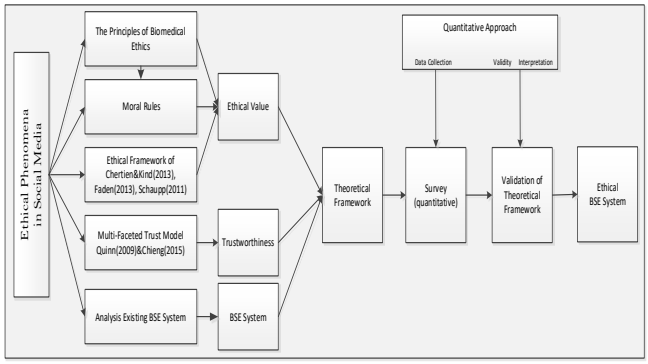


Fig. 2. Research Flow.

B. Research Method

The research method in this study is a quantitative approach. The quantitative design will do several steps on data gathering and analysis. There are identify participants, sampling size, reliability and validity, and the survey finding.

1) Participants: The questionnaire is distributed to potential participants such as doctors[44], and the public, which is taken to hospitals (doctors) and universities (students and lecturers). The study consent participants must be 18 years old and above, which is staying in the city of Jakarta, Indonesia[45]. The participants came from different backgrounds of study, and all were speaking and writing in the Indonesian language. The questioner has been designed into one form of instrument, whereas the instrument is designed into two languages, Indonesian and English (Appendices A.1).

2) Sampling size: The sampling size does base on target respondents. The target respondents are doctors and patients (outpatient females such as university students, housewives, or others). The sample size of doctors is based on the total population of doctors in Jakarta (the capital city of Indonesia). The total population of the doctor is 19536 [46], [47]. The sample size calculation for doctors based on the Slovin formula as mention below:

$\mathbf{n} = \frac{N}{1+Ne^2}$, *where n= number of samples, N = total population and e =error tolerance (e = 0.05)* [48].

So, sample size for doctor is $n = \frac{19536}{1+19536(0.05)^2} = 392$ samples of the respondent.

Furthermore, the sample size for the patient (university student, housewife, or others) is an unknown population. So, the sample size is based on Partial Least Squares (PLS) guidance. PLS is important for structural equation modeling tools in developing theories for exploratory research, such as describing the independent variance variable on investigative the model. The sample size required by PLS is First, ten times the largest number of formative indicators used to measure a single construct (independent variable); Second, ten times the largest number of structural paths directed at a particular construct in the structural model[49]. The researcher has identified four variables, six sub-variables, and each sub-variables includes 1-6 measuring indicators with a total of 38 indicators. The researcher chooses PLS as a suitable analytical tool to fit the small sample size requirement. In this case, the sample size was calculated based on a number of indicators, so ten times 38 equivalent to 380. Therefore, the minimum sample size for patients is 380.

3) Data analysis: A series of data analyses will be conducted on this study to test the research hypotheses. The data analysis through a quantitative approach is applied to structural equation modeling (SEM) in this study for identifying the relationship among observed variables. The correlation and multiple regression analyses will be executed from the collected quantitative[50], [51]. SEM consists of partial least squares SEM (PLS-SEM) and covariance-based SEM(CB-SEM). PLS-SEM is used to build theories by emphasizing on explaining the variance in the dependent variables while examining the model. Whereas, CB-SEM is used to accept or reject theories on a proposed theoretical model to estimate the covariance matrix as a sample data set[49]. This study prefers to used PLS-SEM due to less sample size required and simple on the multivariance normality[49]. Therefore, this method is appropriate for being used in the proposed model. PLS-SEM provides two main components: 1. Structural model, and 2. Measurement model. The structural model (represents the construct in circles) is the relationship between the variables(constructs). The measurement model is the relationships between the variables(constructs) and the indicator(represented in rectangles) [49].

4) Reliability and validity: The reliability is "a matter of whether a particular technique, applied repeatedly to the same object, yields the same result each time." Whereas validity is "the extent to which an empirical measure adequately reflects the real meaning of the concept under consideration" [52]. In other words, validity is "the extent to which differences found with a measuring instrument reflect true differences among those being tested" [53]. Reliability and validity will be addresses in quantitative.

In this study, the validity separated into internal validity and external validity [51], [54].

a) Internal validity is a causal claim by the researcher in a single experiment on the relationship between two variables. The result of correlation and multiple regression analysis will not guarantee the relationship of the variables between the independent variable and the dependent variable [55].

b) External validity is an inference validity on the effect of variables relationship results that being generalize to the population [55]. In phase 1: the significance value of the quantitative analysis indicates that the accessible population can be inferred from the sample in the study [51].

5) Survey findings: The instrument's strategies of printing and distribution were divided into several rounds until achieving the number of the valid questionnaire. In the first round, instruments were printed and distributed 1000 questionnaires. Those questionnaires were distributed to six hospitals and one university. Out of 1000 questionnaires, 772 valid questionnaires, 123 invalid questionnaires, and 105 questionnaires were not returned. The instrument has been designed for the ordinal scale data type, and the statistical analysis has been done through smartPLS 3.2.8.

IV. Results

This section will describe demographics, descriptive statistics, analysis of the formative measurement model, analysis of the reflective measurement model, and the relationship analysis between trustworthiness and ethical value.

A. Demographics

This section is explaining the data related to the respondent's profile, such as profession, age, gender, education, and healthcare system user. Based on Table V has shown the total respondents are divided into two categories, doctors 51%, and outpatient 41%, whereas the outpatient dominated by female students. The target respondents of this study are females 75% and males 25% on using the BSE system. The respondents are dominated by experience users on using the healthcare system (72%). In age, the respondents majority 18-29 was 57%, 30-39 (24%), and 40-49 (12%). At the same time, education dominant by bachelor's degrees 51%, followed by high school 24%, master's degree 16%, and Ph.D. 4%. From the educational background, the researcher could categories most of the respondents are educated and aware of the healthcare system.

B. Descriptive Statistic

The descriptive statistic shows the indicators with mean, standard deviation, min, max, kurtosis, and skewness. The mean values are in the range of 3.968 to 4.367 for BSE1 (user account) and ITG7 (clinical judgment), which means the ITG7 (clinical judgment) as the highest implies by the users and BSE1 (user account) as the lowest implies by the users. The user accepts the integrity of physicians on their clinical judgment. In parallel, the mean values for interaction (INT1-INT2), integrity (ITG1-ITG7), confidentiality (CFT1-CFT5), protection (PRT1-PRT2), caring (CRG1-CRG2), and fairness (FRS) is above 4.0, which means respondents agree that the use of BSE system must have ethical value on it. Moreover, the trustworthiness (TW1-TW8) shows a mean value above 4.0 with the lowest TW8 (faith) 4.100 and the highest TW6 (belief) 4.350, which means the respondents trust the BSE system. The indicator's value of kurtosis and skewness is less than the -1 and +1 range, meaning that the data is accepted as a normal distribution.

TABLE V. Respondents Demographic

Demographic	Category	Count (N=772)	Percentage
Profession	Doctor Outpatient: • Student • Housewife • Others	392 202 65 113	51% 26% 8% 15%
Gender	Male Female	195 577	25% 75%
Age	18-29 30-39 40-49 50-59 Above 60	442 189 94 41 6	57% 24% 12% 5% 1%
Education	High School Diploma Bachelor Degree Master Degree Ph.D. or equivalence	183 40 393 122 34	24% 5% 51% 16% 4%
Healthcare system User's	Yes No	553 219	72% 28%

C. Analysis of Formative Measurement Model

The formative measurement model is the relationship between latent variables and the indicators. The development of the construct must consider the reflective and formative measurement model [49]. In this study, Ethical value has six sub-constructs, namely, interaction, integrity, confidentiality, protection, caring, and fairness are identified as formative measures. Those six sub-constructs are not correlated with each other. The valid measurements for ethical value are based on convergent validity, collinearity, and weight significance assessments [49]. Fig. 3 shows the ethical value measurement model based on smartPLS analysis, and each indicator carries a loading value.

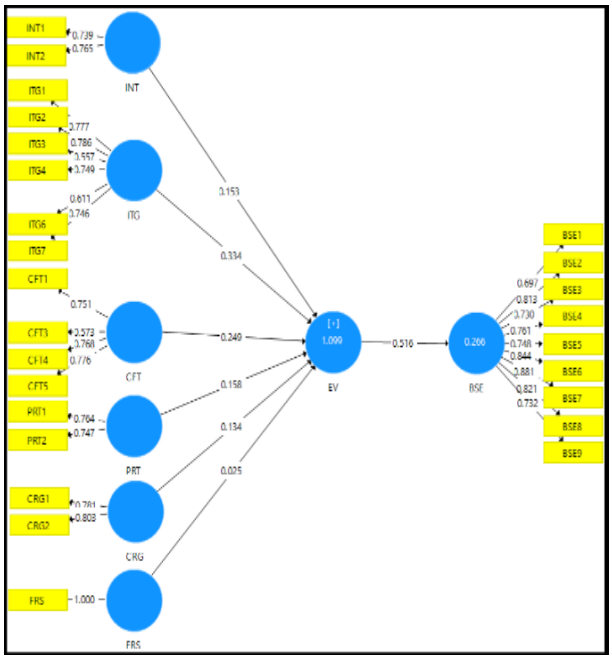


Fig. 3. The Formative Measurement Model of Ethical Value.

Convergent validity is a positive measure that correlates with alternative measures of the same construct [49]. This validity will assess the degree construct indicators which are related to each other by getting their loadings, composite reliability (CR), and average variance extracted (AVE) in the outer model [56]. The observed of indicator loadings of interaction(INT), integrity (ITG), confidentiality (CFT), protection (PRT), caring (CRG), and fairness (FRS) are above 0.70 except for ITG3 (honest information), ITG5 (responsibility), ITG6 (improving quality), CFT2 (anonymity), and CFT3 (clinical result) indicators (Appendixes A). ITG5 (responsibility) and CFT2 (anonymity) indicators with a poor loading of 0.569 and 0.573 is then deleted to improve AVE scores of construct. The improvement of AVE score happened to confidentiality (CFT) from AVE=0.482 to 0.521, integrity (ITG) from AVE=0.472 to 0.504 after ITG5 and CFT2 deletion. Those improvements have been satisfying the requirement of CR and AVE scores above the thresholds [49]. Based on the results, convergent validity has shown the latent constructs measurements achieved their loading, CR, and AVE values have exceeded the criteria.

The collinearity assessment of ethical value makes sure the indicators represent the latent construct and not high level correlated between the indicator variables. At the same time, the measurement of collinearity is based on the variance inflation factor (VIF). The VIF values are five and higher that represents the collinearity problem. Therefore, the VIF value threshold should be below 5(Hair et al., 2017). The highest value of VIF, among other indicators, occurs to PRT1=4.205, which means that not every ethical value indicator is part of the collinearity problem due to those indicators below the VIF threshold (Appendixes B).

To assess the formative measurement must look at each indicator weight of the t-value, and it significant to the indicator validity [57]. The significance level to measure the formative on ethical value is evaluated using a bootstrapping feature with 5000 subsamples [49]. The indicators of ethical value are significant at p<0.001 for outer weights and outer loading. As a result, that component (interaction, integrity, confidentiality, protection, caring, and fairness) have causal relationships with the construct (ethical value).

D. Analysis of Reflective Measurement Model

The reflective measurement in the study is based on the construct (independent variables) of trustworthiness (IV2), BSE system (DV), and moderator of trust propensity. The analysis is similar to the formative measurement, which is evaluated based on convergent validity and discriminant validity. Fig. 4 represents the reflective measurement of trustworthiness (IV2), each indicator reflecting trustworthiness (IV2) based on an outer loading value above 0.5.

The convergent validity is on reflective measurement based on trustworthiness (IV2), BSE system (DV), and moderator of trust propensity. Similar to formative convergent validity, this validity will evaluate the construct outer loadings, composite reliability (CR), and average variance extracted (AVE). The convergent validity of reflective measures has shown the latent constructs measurements achieved their loading (>0.50), CR (0.60-0.95), and AVE (0.50) values have exceeded the criteria (Appendixes C).

The significance level to measure the reflective measures on trustworthiness and BSE system are evaluated using a bootstrapping feature with 5000 subsamples [49]. The indicators of trustworthiness and the BSE system are significant at p<0.001 for outer weights and outer loading.

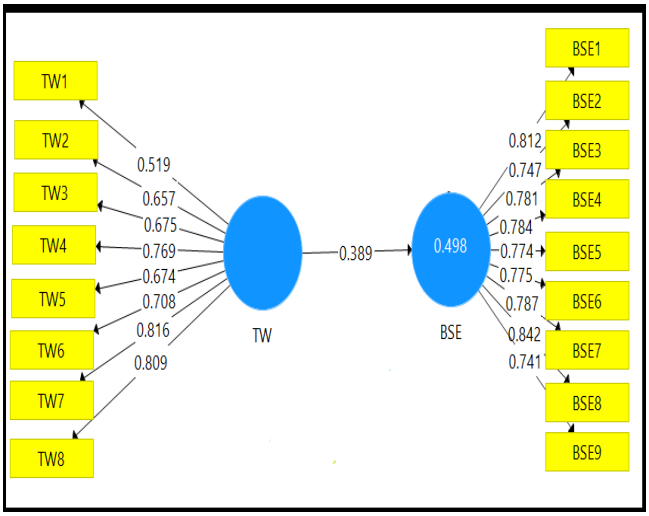


Fig. 4. Reflective Measurement of Trustworthiness (IV2).

E. Analysis of Structural Model

After the analysis of formative and reflective measurement is completed, the researcher proceeds with the structural model evaluation and hypothesis testing, whereas the evaluation is based on collinearity assessment, coefficient of determination and predictive accuracy, and path coefficients.

1) Collinearity assessment, the structural model evaluation should be first tested the collinearity assessment among the reflective measures. This evaluation to make sure during regression analysis not affected by collinearity problems [58]. The assessment for formative and reflective is based on inner VIF values. Table VI shows the inner VIF values for the independent variable (exogenous construct) in contradiction of the dependent variable (endogenous construct), which means that the model is far from collinearity problems (VIF value lower than five).

2) Coefficient of Determination (R^2) is to evaluate the structural model by calculating the squared correlation between the actual dependent variable and predicted values. To evaluate the value of R^2 using smartPLS through consistent PLS algorithm and look at the quality criteria→R square as a result. The coefficient represents the combined effect of independent variables and dependent variables [49]. The value of R^2 is 0.509, meaning that 51% of the variance (BSE system) is explained by the independent variables (ethical value and trustworthiness). The R^2 = 0.509 are identified as substantial since the basis of variance explanations above 50% [59].

3) The path coefficients are the way to evaluate the significance and relevance of constructs path relationship.

There are three criteria values accepted to conduct the two-tailed test in PLS-SEM: 1.65 on significance level 10%, 1.96 on significance level 5%, and 2.57 on significance level 1% that depends on the study field [49]. The bootstrapping test with 5000 subsamples to execute path coefficients, standard errors, t-Value, and confidence intervals. The significant results 1% level of H1 and H2, which is answering first and second research questions. Notably, the effect of trustworthiness on the BSE system with the highest path coefficient of 0.465 and significant p <.001.

The researcher provides the PLS path model illustrated in Fig. 5, all relationship and effects between the variables (constructs) are positive and significant (p values = 0.000).

Finally, the ethical framework on the BSE system has been valid on the bootstrapping test with 5000 subsamples shown in Fig. 5. The relationship between IV1-DV and IV2-DV.

F. Analysis of the Relationship between Ethics and Trustworthiness in the Healthcare Field

Based on the analysis of Venn diagram (Fig. 1), the ethical value indicators are part of the principles of biomedical ethics.

TABLE VI. COLLINEARITY STATISTICS FOR FORMATIVE AND REFLECTIVE MEASURES

Construct	BSE System(DV)
Ethical Value(IV1)	VIF: 2.257
Trustworthiness(IV2)	VIF: 3.287
Trust Propensity	VIF: 1.923

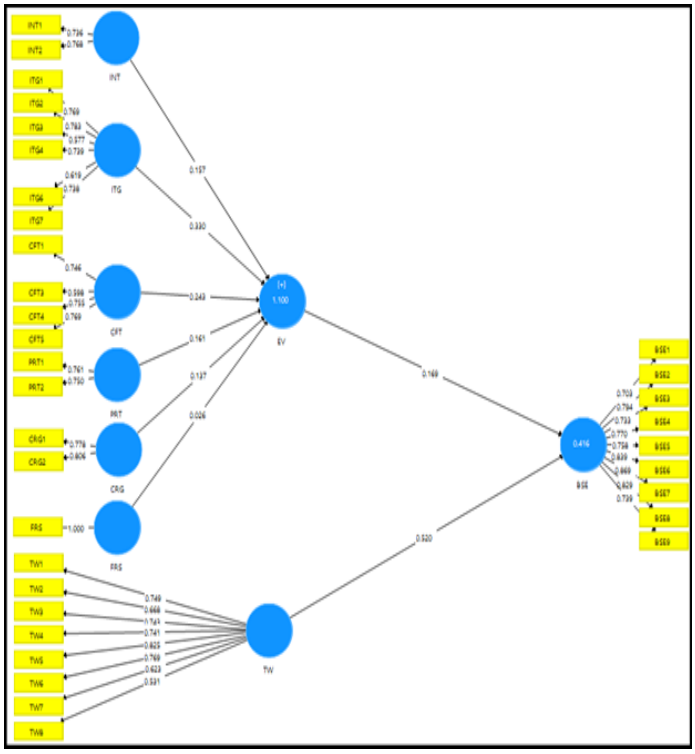


Fig. 5. Structural Model of Ethical Framework.

The relationship between ethics and trustworthiness in healthcare is presented in the Venn diagram in Fig. 6. Ethical value carries interaction, integrity, confidentiality, protection, caring, and fairness. Whereas trustworthiness carries honesty, reputation, competency, reliability, credibility, belief, confidence, and faith. Integrity brings several related indicators with trustworthiness. There is honest information (Table II) related to honesty (Table III). Honesty is telling the truth and fulfills any promises made [29], [37]. Respect is related to respect for autonomy (Table I). Whereas respect for autonomy is the right to hold views, to make choices, and take action based on personal values and beliefs [23]. The belief in ethics is related to the belief of trustworthiness. Therefore, respect is related to the belief of trustworthiness.

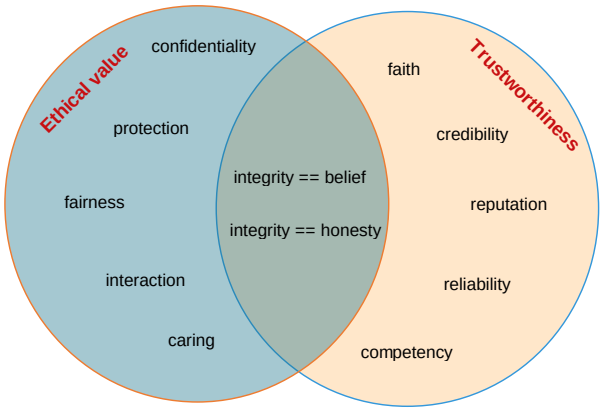


Fig. 6. Venn Diagram of the Relationship between Ethics and Trustworthiness.

G. Relationship Analysis between Ethical Value and Trustworthiness on the Healthcare System

1) The Relationship between Ethical value and trustworthiness:

- Based on the theory, ethics is an individual character of person and persons, which carries several aspects such as interaction, integrity (honest information), confidentiality, protection, caring, and fairness.
- Trust is the ability to be honest or truth to another person. The capability of being honest is part of human integrity. Once a person could tell the truth at any moment and situation, he/she will identify as a person with good integrity. Trust also carries several values such as honesty, reputation, competency, reliability, credibility, belief, confidence, and faith.
- Since honesty is one of the trust values, it will connect to the ethical value of integrity that carries "honest information" as the value. The relationship between ethics and trust is on the value of honesty. However, the honesty of ethical value emphasizes honest information, which identifies as an adjective. At the same time, honesty on trust emphasizes the action of the patient's statement saying the truth in the area of pain to the doctor. Doctors are telling the truth to the patient on the treatment given.

2) The Differences between Ethical value and trustworthiness:

- Ethics emphasizes the overall individual character of a person or persons, whereas trust emphasizes on the capabilities of a person being honest or being truthful.
- The difference is in the scope of ethics and trust functions related to the healthcare system. The scope of ethics is to the physician, patient, and BSE system. At the same time, the scope of trust focuses on the trust of patients to the physician.
- Ethical value carries 19 indicators, whereas trustworthiness carries eight indicators. Moreover, the nineteen indicators of ethics and eight indicators of trust are not related to each other.

The 19 indicators of ethical value significantly influence the BSE system as well as trustworthiness, with eight indicators significantly influences the BSE system. The trustworthiness is emphasized in measuring physician direct performance, whereas ethical value carries 19 indicators to measure patient, physician, and healthcare systems.

V. Discussion

The main finding in this study is; first, the ethical value positively influences the BSE system. The correlation effect between the ethical value and the BSE system is significant (P<.001). Second, trustworthiness positively influences the BSE system. The correlation effect between trustworthiness and the BSE system is significant (P<.001).

There are 772 participants who are filled with a valid questionnaire consist of 51% doctors and 49% outpatient female. The demographic of education background shown most of the respondents are educated and aware of the healthcare system.

Based on Fig. 3, the formative measurement model analysis has bring ethical value with a loading such as interaction (0.153), integrity (0.334), confidentiality (0.249), protection (0.158), caring (0.134), fairness (0.025). Integrity becomes the most preferences of ethical value. Therefore, the physician should have good integrity when accessing the BSE system. The indicators of integrity with the loading are patient interest as the priority (ITG1) 0.777, expert advice (ITG2) 0.786, honest information (ITG3) 0.557, respect (ITG4) 0.749, improving quality (ITG6) 0.611, and clinician judgment (ITG7) 0.735. Out of the seven indicators, the higher outer loading is ITG2 0.786 and ITG1 0.777, which means that the "expert advice" and "patient as the priority" as the essential aspects of doctors and patients to be concerned. A doctor, as an expert able to provide a clear treatment solution on a particular disease [12]. In this case, the doctor as an expert to advise a patient based on their best treatment solution. Doctor integrity must put patients as their highest priority to compare with other things [12], [60]. Respect (ITG4) with loading 0.749 that concern between physician and patient must respect each other before interaction. A doctor must write a full respect narrative on social media [12] and respect the right of the patient [27]. Clinical judgment (ITG7) with loading 0.735, the capabilities of doctors to make a clinical judgment to a particular patient on her breast cancer disease. The clinical judgment must base on the scientific evidence and physician experience in treating the patient [27]. A patient expected an excellent clinical judgment from the doctor on their disease. Honest information (ITG3) with 0.557 that the honest information given by the patient will help a doctor to provide a solution, as well as the doctor, must be honest to inform to the patient related to the treatment and fees related [12].

The indicators of interaction are communication (INT1) and sharing information (INT2) with outer loading 0.882 and 0.887, which means that communication and sharing information as an essential aspect that influences the healthcare system. Physicians and patients must be able to identify communication in the healthcare system are credible [12]. The sharing of information within the community is essential to avoid strangers [28]. The indicators of confidentiality and the loading are informed consent (CFT1) 0.751, the clinical result (CFT3) 0.573, de-identification (CFT4) 0.768, and lookup information (CFT5) 0.776. The higher outer loading on lookup information 0.776 and informed consent 0.751. The way a doctor to look up patient information as an essential part of the BSE system, and patient consent is necessary before accessing the data. In the open internet space, to look up patient information is not wrong [12] as patient consent exists. De-identification (CFT4) with loading 0.768 will help the patient to de-identify their self before meeting with the doctor. The capabilities of the system to be able to de-identify patient identity [12]. The feel of shameful while consultation with male doctors was solved. Clinical result (CFT3) with loading 0.573, the medical history of self-exam recorded along with the communication history with the doctor [27]. The indicators of protection are refrain harm (PRT1) and safeguard (PRT2) with the higher outer loading on refrain harm 0.764 and safeguard 0.747. A patient's medical history posted by a doctor with patient consent will refrain harm to the patient [12]. The healthcare system protection must base on the encryption mode [12]. With the encryption on the user account that the BSE system will secure the interaction between patient and physician. The indicators of caring are engagement (CRG1) and optimal care (CRG2) with the higher outer loading on engagement 0.803 and follow by optimal care 0.781. The healthcare system is able to engage the patient's own care and perform quality assessment [12]. The capabilities of a system to make comfort to the patient and physician on accessing the system. Optimal care also helps the attention of a physician to provide the best treatment as the final outcome [27]. The indicator of fairness is inequality with loading 1. Inequalities of scientific health evidence will create wrong clinical decisions [27]. The equalities of the clinical result will bring better treatment evidence on clinical decisions.

Fig. 4 shown reflective measurement model analysis on trustworthiness. The trustworthiness with eight indicators, such as honesty, reputation, competency, reliability, credibility, belief, confidence, and faith, carries a variety of loading values. The highest loading of the trustworthiness indicator is confidence (TW7) 0.816, followed by faith (TW8) 0.809, reliability (TW4) 0.769, belief (TW6) 0.708. Others are competency (TW3) 0.675, credibility (TW5) 0.674, reputation

(TW2) 0.657, and the lowest loading is honesty (TW1) 0.519. When looking at the path relationship, trustworthiness has the strongest path coefficient relationship with the BSE system. The path coefficient 0.465, t-value 11.667 and p-value 0.000.

Confidence with loading 0.816 has a feeling of easiness related to the belief hold [40], which means a doctor and patient have good confidence while accessing the BSE system. Faith with loading 0.809 has the capability to fully trust something [61], which means the full trust of patients and physicians on the interaction process in the healthcare system. Reliability with loading 0.769 has the quality of being trustworthy based on the system [62], which means a doctor and patient expected a reliable medical history. Belief with loading 0.708, belief is accepted without any doubt [23], which means the patient and physician accepted using the healthcare system without any doubt. Competency with loading 0.675 has the capability to fulfilled other person needs [37], which means the capabilities of a doctor to do treatment to every patient through the system. Credibility with loading 0.674 has the quality of belief in particular things [62], which means the capabilities of the healthcare system to support the need of patients and doctors. Reputation with loading 0.657, the expectation of an agent behavior from the information[39], which means doctors' reputation depends on the trust value given by the patients. Honesty with loading 0.519, the promises made by a particular person as a good faith [37], which means doctors must keep their statement on explaining the case to a particular patient. As well as patients must saying the truth of their sickness to the doctor for correct treatment.

In the general healthcare field, the relationship between ethics and trustworthiness are related to each other (Fig. 5). Two indicators are related to each other (Fig. 6). The integrity of ethics and honesty of trustworthiness are related. As well as the integrity of ethics and honesty of trustworthiness are related.

The relationship between ethical value and trustworthiness in the healthcare system has been identified on the honesty indicator. However, Honesty is the indicator of trustworthiness as a noun, whereas honest information exists on an ethical value indicator as an adjective. Honesty emphasizes the action of the patient's statement to the physician in the area of sickness. Doctors are telling the truth to the patient on medical results. Therefore, there is no relationship between trustworthiness and ethical value in the healthcare system.

VI. Conclusion

The study has confirmed an ethical value formative measurement model to predict the relationship with the BSE system. The relationship between ethical value and the BSE system has proved significant findings. On the other hand, the trustworthiness reflective measurement model to predict the relationship with the BSE system has proved significant findings. The relationship between ethical value and trustworthiness happens in honesty and belief indicators.

The implementation of ethical value and trustworthiness enables patients to be confident in using the BSE system. Patients and physicians will be secure in using the healthcare system, which is protected by the ethical value.

VII. Future Challenges

The challenge of target participants happening due to a specialist of an oncologist is limited time for the survey. The survey is not specific to any particular race, culture, or religion and the study also focuses on the use of the BSE system.

Acknowledgment

The ethical code of this study was approved by the research ethics board of Esa Unggul University committee (No. 0155-20.133/DPKE-KEP/FINAL-EA/UEU/V/2020).

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” CA. Cancer J. Clin., vol. 68, no. 6, pp. 394–424, 2018.

[2] L. M. Hassan et al., “Evaluation of effect of self-examination and physical examination on breast cancer,” The Breast, vol. 24, no. 4, pp. 487–490, 2015.

[3] M. Y. Roth, J. G. Elmore, J. P. Yi-Frazier, L. M. Reisch, N. V. Oster, and D. L. Miglioretti, “Self-Detection Remains a Key Method of Breast Cancer Detection for U.S. Women,” J. Women’s Heal., vol. 20, no. 8, pp. 1135–1139, 2011.

[4] S. Solikhah, S. Promthet, N. Rakkapao, and C. P. Hurst, “Validation of an Indonesian Version of the Breast Cancer Awareness Scale (BCAS-I),” Asian Pac. J. Cancer Prev., vol. 18, no. 2, pp. 515–522, 2017.

[5] E. Sujindra and T. Elamurugan, “Knowledge, attitude, and practice of breast self-examination in female nursing students,” Int. J. Educ. Psychol. Res., vol. 1, no. 2, p. 71, 2015.

[6] S. Kemp, “Digital in 2020,” We Are Social Inc., 2020. [Online]. Available: https://wearesocial.com/blog/2020/07/digital-use-around-the-world-in-july-2020. [Accessed: 10-Aug-2020].

[7] M. Modahl, L. Tompsett, and T. Moorhead, “Doctors , Patients & Social Media,” QuantiaMD, 2011. [Online]. Available: whttp://www.quantiamd.com/q-qcp/social_media.pdf. [Accessed: 10-Aug-2020].

[8] T. Hale, A. Pathipati, S. Zan, and K. Jethwani, “Representation of health conditions on Facebook: content analysis and evaluation of user engagement,” J Med Internet Res, 2014.

[9] K. Denecke et al., “Ethical Issues of Social Media Usage in Healthcare,” IMIA Yearb. Med. Informatics, pp. 137–147, 2015.

[10] S. Fox, “Peer-to-peer Health Care,” PewReseachCenter, 2011. [Online]. Available: www.pewinternet.org. [Accessed: 12-Dec-2016].

[11] M. White and S. M. Dorman, “Receiving social support online: Implications for health education,” Health Educ. Res., vol. 16, no. 6, pp. 693–707, 2001.

[12] K. C. Chretien and T. Kind, “Social media and clinical care: Ethical, professional, and social implications,” Circulation, vol. 127, no. 13, pp. 1413–1421, 2013.

[13] J. Brown, “How Doctors View and Use Social Media: A National Survey,” JMIR mHealth uHealth, vol. 16, 2014.

[14] K. C. Chretien, S. R. Greysen, J.-P. Chretien, and T. Kind, “Online Posting of Unprofessional Conduct by Medical Students,” Jama, vol. 301, no. 12, pp. 1309–1315, 2009.

[15] J. Fisher and M. Clayton, “Who Gives a Tweet: Assession Patients’ Interest in the Use of Social Media for HealthCare,” Worldviews Evidence-Based Nurs., vol. 9, 2012.

[16] S. R. Greysen, D. Johnson, K. C. Chretien, C. Gross, A. Young, and H. J. Chaudhry, “online professionalism investigations by State Medical Boards-First, DO NOT HARM,” Ann. Intern. Med., 2013.

[17] R. Khana, M. Mahinderjit Singh, F. Damanhoori, and N. Mustaffa, “Investigating the Importance of Implementing Ethical Value on a Healthcare System within a Social Media context,” Int. J. Innov. Creat. Chang., vol. 12, no. 5, pp. 352–369, 2020.

[18] C. L. Ventola, "Social Media and Health Care Professionals : Benefits , Risks , and Best Practices," P&T, vol. 39, no. 7, pp. 491–500, 2014.

[19] T. Kind, S. R. Greysen, and K. C. Chretien, "Pediatric clerkship directors' social networking use and perceptions of online professionalism," Acad. Pediatr., vol. 12, no. 2, pp. 142–148, 2012.

[20] S. R. Greysen, T. Kind, and K. C. Chretien, "Online professionalism and the mirror of social media," J. Gen. Intern. Med., vol. 25, no. 11, pp. 1227–1229, 2010.

[21] K. Krot and I. Rudawska, "The Role of Trust in Doctor- Patient Relationship : Qualitative Evaluation of Online Feedback," vol. 9, no. 3, pp. 76–88, 2016.

[22] Belmont Report, "Ethical Principles and Guidelines for the Protection of Human Subjects of Research," Washington D.C., 1974.

[23] T. L. Beauchamp and J. F. Childress, Principles of Biomedical Ethics, 6th ed. Oxford University Press, 2009.

[24] B. Gert, Its Nature Justification of Morality, Revised. Oxford University Press, 2005.

[25] J. P. Thiroux and K. W. Krasemann, Ethical Theory And Practice, 11th ed. Pearson, 2012.

[26] J. H. Tufts, "Ethical Value," J. Philos. Inc., vol. 75, no. 11, pp. 664–677, 1908.

[27] R. R. Faden, N. E. Kass, S. N. Goodman, P. Pronovost, S. Tunis, and T. L. Beauchamp, "An Ethics Framework for a Learning Health Care System: A Departure from Traditional Research Ethics and Clinical Ethics," Hastings Cent. Rep., vol. 43, no. SUPPL. 1, 2013.

[28] L. C. Schaupp, L. D. Carter, D. L. Schaupp, and N. C. a, "Ethics in Social Networking: A Framework for Evaluating Online Information Disclosure," Proc. 44th Hawaii Int. Conf. Syst. Sci. HICSS, pp. 1–7, 2011.

[29] A. Hornby, Oxford Advanced Learner's Dictionary of Current English. Oxford University Press, 2005.

[30] "Medical Dictionary," Farlex, Inc, 2003. [Online]. Available: https://medical-dictionary.thefreedictionary.com. [Accessed: 09-Aug-2019].

[31] C. Rizza, P. Curvelo, I. Crespo, M. Chiaramello, A. Ghezzi, and Â. G. Pereira, "Interrogating Privacy in the digital society: media narratives after 2 cases 6 Interrogating Privacy in the digital society: media narratives after 2 cases," IRIE Int. Rev. Inf. Ethics, vol. 16, no. 12, pp. 6–17, 2011.

[32] M. Mahinderjit Singh, P. J. Ng, K. M. Yap, M. H. Husin, and N. H. A. H. Malim, "Cyberbullying and a mobile game app? An initial perspective on an alternative solution," J. Inf. Process. Syst., vol. 13, no. 3, pp. 559–572, 2017.

[33] US-HHS, "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule," Health and Human Services(HHS.gov), 2015. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#rationale. [Accessed: 22-Oct-2019].

[34] R. C. Mayer and J. H. Davis, "An Integrative Model of Organizational Trust," Acad. Manag., vol. 20, no. 3, pp. 709–734, 1995.

[35] J. B. Rotter, "A new scale for measurement of interpersonal Trust," J. Pers., vol. 35, no. 4, 1967.

[36] L. B. Chieng, M. Manhinderjit Singh, Z. F. Zaaba, and R. Hassan, "Multi-Facet Trust Model for Online Social Network Environment," Int. J. Netw. Secur. Its Appl., vol. 7, no. January 2015, 2015.

[37] K. Quinn, D. Lewis, D. O. Sullivan, and V. P. Wade, "An Analysis of Accuracy Experiments Carried out over of a Multi-Faceted Model of Trust," Int. J. Inf. Secur., vol. 8, no. 2, pp. 103–119, 2009.

[38] J. Golbeck and J. Hendler, "Accuracy of Metrics for Inferring Trust and Reputation in Semantic Web- Based Social Networks," no. April 2004, 2004.

[39] J. A. N. M. Leimeister and W. Ebner, "Design , Implementation , and Evaluation of Trust-Supporting Components in Virtual Communities for Patients," J. Manag. Inf. Syst., vol. 21, pp. 101–135, 2005.

[40] D. H. McKnight and N. L. Chervany, "The Meanings of Trust," no. May, 1996.

[41] H. Thornton and R. R. Pillarisetti, "'Breast awareness' and 'breast self-examination' are not the same. What do these terms mean? Why are they confused? What can we do?," Eur. J. Cancer, vol. 44, no. 15, pp. 2118–2121, 2008.

[42] J. L. Longe, The Gale Encyclopedia of Cancer "A guide to Cancer and its treatments," 2nd ed. ThomasGale, 2005.

[43] R. Khana, M. Mahinderjit Singh, F. Damanhoori, and N. Mustaffa, "Breast Self-Examination System Using Multi-Faceted Trustworthiness: Observational Study," JMIR Med. Informatics, 2020.

[44] P. White, "Getting health professionals to work together," BMJ, vol. 320, pp. 1021–1022, 2000.

[45] S. Kemp, "Digital in 2018: World's internet users pass the 4 billion mark," wearesocial.com, 2018. [Online]. Available: https://wearesocial.com/blog/2018/01/global-digital-report-2018. [Accessed: 13-Feb-2018].

[46] D. Budijanto, F. Sibuea, W. Widiantini, and Y. Aryantin, "Data dan Informasi Profil Kesehatan Indonesia 2018," Jakarta, 2019.

[47] IDI, Profil Dan Direktori Ikatan Dokter Indonesia. Jakarta: Ikatan Dokter Indonesia(IDI), 2018.

[48] T. P. Ryan, Sample Size Determination and Power. Canada: John Wiley & Sons,Inc, 2013.

[49] J. F. Hair, G. T. M. Hult, C. M. Ringle, and M. Sarstedt, A Prime on Partial Lease Squares Sructural Equation Modelling (PLS-SEM). Sage Publications Asia-Pacific Pte. Ltd, 2017.

[50] J. W. Creswell, Research design, 4th ed. Sage Publications, 2014.

[51] S. Wong and P. Cooper, "Reliability and validity of the explanatory sequential design of mixed methods adopted to explore the influences on online learning in Hong Kong bilingual cyber higher education," Int. J. Cyber Soc. Educ. Pages, vol. 9, no. 2, pp. 45–64, 2016.

[52] E. Babbie, The Basics of Social Research, Fifth Edit. United States of America, 2011.

[53] C. Kothari, Research Methodology "Methods and Technique," 2nd ed. New Age Int Publishers, 2004.

[54] M. Zohrabi, "Mixed Method Research : Instruments , Validity , Reliability and Reporting Findings," Theory Pract. Lang. Stud., vol. 3, no. 2, pp. 254–262, 2013.

[55] W. R. Shadish, T. D. Cook, and D. T. Campbell, Experimental and Quasi-Experimental Designs For Gneralized Causal Inference. Boston, New York: Houghton Mifflin Company, 2002.

[56] L. Tomei and R. Morris, ICTs for Modern Educational and Instructional Advancement. Information Science Reference, 2009.

[57] P. Andreev and H. Maoz, "Validating Formative Partial Least Squares (PLS) Models : Methodological Review and Empirical Illustration," AIS Electron. Libr., 2009.

[58] M. Sarstedt, C. M. Ringle, D. Smith, R. Reams, and J. F. Hair, "Journal of Family Business Strategy Partial least squares structural equation modeling (PLS-SEM): A useful tool for family business researchers," J. Fam. Bus. Strateg., vol. 5, no. 1, pp. 105–115, 2014.

[59] W. W. Chin, "The partial least squares approach for structural equation modeling.," Mod. methods Bus. Res., no. April, pp. 295–336, 1998.

[60] H. C. Sox, "Medical Professionalism in the New Millenium: A Physician Charter," Ann. Intern. Med., vol. 136, no. 3, pp. 243–246, 2002.

[61] L. A. Anderson and R. F. Dedrick, "Development of the Trust in Physician Scale : A Measure to Assess Interpersonal Trust in Patient-Physician Relationships," Phsycological Reports, no. November, 1990.

[62] K. Quinn, "A Multi-Faceted Model of Trust that is Personalisable and Specialisable," University of Dublin, Trinity College, 2006.

APPENDICES

A. Convergent Validity of Formative Measure

Construct		Indicators	Convergent Validity		Internal Consistency Reliability	
			Loading >0.50	AVE >0.50	Composite Reliability 0.60-0.90	Cronbach' Alpha 0.60-0.90
EV	INT	INT1	0.739	0.566	0.722	0.722
		INT2	0.765			
	ITG	ITG1	0.777	0.504	0.857	0.852
		ITG2	0.786			
		ITG3	0.557			
		ITG4	0.749			
		ITG6	0.611			
		ITG7	0.735			
	CFT	CFT1	0.751	0.521	0.811	0.805
		CFT3	0.573			
		CFT4	0.768			
		CFT5	0.776			
	PRT	PRT1	0.764	0.570	0.726	0.726
		PRT2	0.747			
	CRG	CRG1	0.781	0.627	0.771	0.771
		CRG2	0.803			
	FRS	FRS	1.000	1.000	1.000	1.000

B. Collinearity Statistics of Ethical Value

Interaction		Integrity		Confidentiality		Protection		Caring		Fairness	
INT1	3.941	ITG1	3.054	CFT1	2.643	PRT1	4.205	CRG1	1.647	FRS	1.000
INT2	1.469	ITG2	1.639	CFT3	1.646	PRT2	1.482	CRG2	1.647		
		ITG3	1.259	CFT4	1.534						
		ITG4	2.699	CFT5	2.873						
		ITG6	1.420								
		ITG7	3.209								

C. Convergent Validity for Reflective Measures

Construct	Indicators	Convergent Validity		Internal Consistency Reliability	
		Loading >0.50	AVE >0.50	Composite Reliability 0.60-0.95	Cronbach' Alpha 0.60-0.95
TW	TW1	0.519	0.503	0.888	0.891
	TW2	0.657			
	TW3	0.675			
	TW4	0.769			
	TW5	0.674			
	TW6	0.708			
	TW7	0.816			
	TW8	0.809			
TP	TP1	0.872	0.744	0.853	0.853
	TP2	0.853			
BSE	BSE1	0.812	0.613	0.934	0.935
	BSE2	0.747			

Examining the Effect of Online Gaming Addiction on Adolescent Behavior

Maha Abdullah Al-Dwehy[1], Dr. Hedia Zardi[2]
Department of Computer Science
College of Computer, Qassim University
Buraydah, Saudi Arabia

***Abstract*—It exceeds daily rates of Internet use among adolescents compared to adults' use of the Internet, as it was monitored that the number of adolescents on the Internet is increasing all over the world. Today, as a result of the ease of access to the Internet in the world, most adolescents' access to the internet world is easier and more common. In this paper, we review some studies that explain the behavior of adolescents while gaming online and its effects. There are some statistics to determine the impact of the Internet on teenagers. The study reviews past studies on adolescent behavior and privacy with a potential impact on adolescent behavior, which has become one of the most important problems. We focused on exploring online game addiction concerns and their effects on teens' behavior. The purpose of this type of study is to determine the objective and examine this study within the backdrop of social reality. This study employed a quantitative methodology. We have selected this methodology because it has been proven to be reliable and has sound construct validity. The data was analyzed using the SPL smart tool and the main objective of this study was to investigate adolescent's behavior in terms of their addiction to online games, and to study parents' awareness of the dangers of online games for their children. The study explored various factors that can influence addiction fears and examines their effects on adolescent behavior and contributed to the literature by identifying correlation factors and addressing this gap by applying through SEM application specifically the Smart PLS tool.**

Keywords*—Online gaming addiction; adolescent behavior on internet; privacy

I. Introduction

A. Adolescent Behavior on the Internet

The rapid progress in technology today has brought about a major social transformation in all areas of society, and as a result of the use of the Internet for adults as well as adolescents at the same level of facilities and interaction, whether through a communication or online games or other [1]. The results of the analysis showed that the number of adolescents exceeds the number of adults, because there is a rapid increase in the number of teenagers around the world, and the results indicate that they are more vulnerable, as they are constantly exposed to inappropriate con- tent for them [1]. Today this generation is called a technical generation, teens spend a lot of time gaming online and interacting on social networking. As a result of the increased accessibility of the Internet in developed countries, the numbers of Internet callers increased, in turn, to a number that cannot be registered today [2]. A complete technological revolution has occurred during the past two decades and a new era for childhood has emerged, the digital age. Which made many researchers focus on their studies on the impact of the use of the Internet in general, and for adolescents in particular, where their research touched on emotional and psychosocial aspects [2]. As they focus on the extent of the negative effects of adolescent addiction to use the Internet [3]. The results of the paper showed that, in the year 2009-2010, approximately 93% of Internet users in the United States are adolescents [3] and the number of Internet users from adolescents in Europe constitutes 60% as the study results showed that they spend their time using the Internet in games via the Internet [2]. And still, to this day, the number of teenagers who use the Internet are increasing, and as a result, the fears are increasing.

Where the authors of the study found that more than half of adolescents aged 9-19 have a local internet connection worse, whether through schools or regular Internet users [1]. This number is still increasing, and even children between the ages of 5 and 8 have a local internet connection. Society has given more attention to protecting adolescents and children from external physical hazards, and they have not paid attention to technological interaction [2], [4] and [5]. Computer and artificial intelligence experts say there is an emotional connection between teens and electronic games around the world, as they seem to know in their depths that they are the computer generation [15]. As a result, today the methods of collecting information have become easier through the use of data acquisition and extraction systems. Numerous studies have shown that adolescents are easy targets for information col- lectors because they are often under-conscious. Data collection for adolescents occurs when there is an interaction between them and the fictional character. Among the factors that affect adolescents, social impact, more than half of the parents believe that the child tends to isolate as a result of using the Internet, especially in online games, so parents have a responsibility to think about whether there is harm or benefit when their children participate in society from During the internet and interpersonal relationships, in addition to focusing on social activities, previous studies indicated that there is a decrease in social activities for individuals who spend more time on the Internet [15]. There is also the problem of the health factor, where attention must be paid to the impact of adolescents' mental and physical health as a result of their use of the Internet, where, as it is clear, the obesity factor has increased with the invasion of the Internet in Western society [15]. The results of the studies showed that the rate of obesity has multi

plied the number of times over the rate found in the sixties and the beginning of the seventies, in addition to inactivity and lack of movement, the child spends long hours up to more than five hours per day either on the computer or mobile [16]. Also, with regard to their health and safety, great attention must be paid to knowing the people they meet through the network, in terms of misfortune in online games or social media. It is important to educate teenagers and draw their attention that there is a hidden danger and must be careful throughout and take all precautions to avoid the danger [16].

B. Online Games

Online games are an important factor in affecting adolescent behavior, as it greatly affects their behavior and may lead to excessive addiction [3]. Concerns have been expressed about online gaming, there is a growing concern that information about them can be collected through online games, which is becoming an increasingly important issue today [3]. In addition, there is a risk of internet content that may be harmful to teenagers as well as children [17]. Teens may place their faith in these technologies without being fully aware of the risks and implications of this, the content may carry violent material that supports hate, or the teenager may be exposed to an illegal content, wrong information, difficult and dangerous games that affect their thinking and behavior [3]. In fact, 85% of parents think that Internet content (photos, games, videos, etc.) pose a greater risk to teenagers and more dangerous than TV [3].

C. Information Privacy

This is personal privacy and there is also information privacy [6]. Many authors have provided different meanings and definitions of the word privacy in all its forms, but there is no standardized and specific description covering all aspects of this term [6]. However, it was agreed that privacy is one of the most important ethical issues for the information age [7]. As a result of adolescents entering the online world, there are increasing concerns about disclosing their privacy [7]. So privacy is a critical issue, and it exceeds adolescents because they are the easiest prey for their privacy to be violated by unauthorized persons or people with malicious intent [8]. They are vulnerable to sharing their personal data easily compared to more conscious adults and are more difficult to provoke compared to adolescents [9]. If privacy refers to the collection of information and unauthorized use, as well as errors of improper access to the control of the individual when issuing personal information [11]. Privacy in terms of disclosure of personal information is another aspect to consider. It is self-disclosure of personal information [13]. When people communicate with each other, this process is called self-disclosure [13]. Intimacy and sensitivity are two advantages of information that seem important to disclosing information, and disclosure of privacy relates to the level and type of information that individual wishes to dis- close to another [13]. The paper [10] specifically found that users disclosed their birth date of 87.8%, and the profiles they examined contained an image of about 90.8% and 50.8% included their current residence. In addition, 39.9% included a phone number, and most users revealed their full names [10]. In socially based Internet domains, privacy disclosure can also be seen online. One such medium is games [10]. Because such games were controlled by youth and adolescents between the ages of 13 and 19 [10]. In online social networks, when an individual's private information is disclosed about a person, the P of his / her boyfriend's FP is disclosed to another FFP (FFP is not a friend of P), then it is called a privacy leak [11]. Privacy leaks may remain all the time, while personal information is shared by interacting with friends [14]. Some of those who want to obtain user information try to use methods to obtain private information from any possible way (sometimes legal and sometimes illegal) to obtain this desired information do not give up and this is a big problem [12] [14].

II. Methods

A. Experiment

1) Data collection: In this study a procedure was used for the data collection design process, which includes sampling technique, target population and questionnaires.

2) Target population: We chose to be a sample of parents, because they are more likely than others to be concerned about the impact of online games on their children's behavior, especially in adolescence, in addition to that they are knowledgeable about adolescent behavior online and have easy access to dis- tribute the questionnaire. When we talk about their children, they pay special attention to this topic. *Sampling techniques:* This study focuses on the method of the online survey. Whereas, as we dis- cussed above, we will collect data from parents as they are concerned about adolescent behavior. The questionnaire will be available through a URL: Put the link on Twitter (Retweet). Send the link via groups on WhatsApp.

B. Questionnaire

The questionnaire for this study was created by Google Docs. Various types of questions were asked which are multiple-choice, select multiple answers, yes or no questions, agree or not, and scale questions). The questions of this study were written in English and Arabic to obtain the largest possible participation from parents. At the beginning of the questionnaire we added a pre-test questionnaire in order to understand the project and its questions (Table I), and then we focused on our hypothesis (Addiction concerns, Psychological effect, Risk, parental control, subjective norm, content) when we developed the questionnaire. This study formulated questions for the questionnaire (as shown in Table II) and asked questions to know the background of the participants.

C. Pre-Test Questionnaire

The questionnaire should not take more than 10 minutes to complete.

TABLE I. SAMPLE OF QUESTIONNAIRE

Name | الاسم	Short answer text.....
1. Are you parent? | هل أنت أحد الوالدين *	Yes / No
2. Do you have child? | هل لديك أطفال؟ *	Yes / No
3. Gender | الجنس *	Female / انثى Male / ذكر
4. Age | العمر *	
20 - 30 / 30 - 40 / 40 - 50 / 50 +	
4. Education level | مستوى التعليم * Please check only one. | يرجى تحديد خيار واحد فقط.	Diploma / دبلوم Bachelor / بكالوريوس Master / ماجستير PhD / دكتوراه Uneducated / غير متعلم Other / ذلك غير
5. How many children/child do you have? | كم طفل لديك؟ * Please check only one. | يرجى تحديد خيار واحد فقط.	1 2 3 4 5+
6. The Average of their age | متوسط أعمار أطفالك *	0-5 Years / سنة 5-10 Years / سنة 10-12 Years / سنة 12-15 Years / سنة 15-18 Years / سنة 18-21 Years / سنة

D. Post Questionnaire

Please rate your agreement with the following statements on a scale of 5.

TABLE II. SAMPLE OF QUESTIONNAIRE 2

	Statement	Scale (Strongly Disagree / Disagree / Neither / Agree / Agree Strongly)
Adolescents's behavior on the Internet (AHB)	I will let my child use the Internet | سأسمح لطفلي بإستخدام الإنترنت *	
	Even if there is a risk, I will let my child use the Internet | حتى لو كان هناك خطر سأسمح لطفلي بإستخدام الإنترنت *	
	My child can use the Internet all the time | يمكن لطفلي إستخدام الإنترنت في كل وقت *	
	In the future, I will not monitor my children when they use the Internet | في المستقبل لن أراقب أطفالي عند إستخدامهم للإنترنت *	
Psychological Effect (PE)	I feel my child is interacting and being greatly affected while gaming online | أشعر أن طفلي يتفاعل ويتأثر بشدة أثناء مشاركته اللعب على شبكة الإنترنت *	
	I am concerned that bad people may negatively affect my child while interacting with them | أنا أقلق من أن الأشخاص السيئين قد يُأثرون بشكل سلبي على طفلي أثناء تفاعله معهم *	
	I am concerned that harmful games have negatively affected my child | أنا أقلق من أن الألعاب السيئة قد تأثر بشكل سلبي على طفلي *	
Risk (R)	My child is affected quickly | طفلي قد يتأثر بشكل سريع *	
	I don't care what kind of games my child interacts with on the Internet | أنا لا أهتم بطبيعة الألعاب التي يتفاعل معها طفلي على الإنترنت *	
	Spending long time playing online is dangerous | قضاء الطفل وقت طويل في اللعب عبر الإنترنت يشكل خطر *	
Subjective norm (SN)	My child tends to share his / her real personal information with others while playing online | يميل طفلي إلى مشاركة معلوماته الشخصية الحقيقية مع الآخرين أثناء اللعب عبر الإنترنت *	
	People who interact with my child gaming online do not share their personal information | لا يشارك الأشخاص الذين يتفاعلون مع طفلي في اللعب عبر الإنترنت معلوماتهم الشخصية الخاصة *	
	My child interacts with online games because she/he is influenced by the opinions of her/his friends | يتفاعل طفلي مع الألعاب عبر الإنترنت لأنه يتأثر بآراء أصدقائه *	
Content (C)	Online gaming content affects my child's behavior | محتوى الألعاب عبر الإنترنت يؤثر على سلوك طفلي *	
	I control all kinds of games that my child can participate in | أنا أتحكم في كل أنواع الألعاب التي يمكن أن يشارك بها طفلي *	
	I use programs to block harmful games | استخدم برامج لحجب الألعاب السيئة *	
parental control (PC)	I have a major responsibility for controlling my children's access to the Internet | لدي مسؤولية كبيرة في التحكم في إستخدام طفلي للإنترنت *	
	I'm aware of the risks faced by my children when they interact in online games | أنا على علم بالمخاطر التي يوجهها أطفالي عند تفاعلهم في الألعاب عبر الإنترنت *	
	I allow my child access to the Internet with limits and with my supervision | أسمح لطفلي باللعب عبر الإنترنت بحدود وبإشرافي *	
	What are your notes about your children behavior while gaming online: Select all the correct answers | ما هي ملاحظاتك حول سلوك ابنك المراهق أثناء اللعب عبر الإنترنت: حدد جميع الإجابات الصحيحة *	Angry / Exited / worry / Others .....
	Please add any additional comments and notes about how you feel as a parent when you see your child playing online games | يرجى كتابة أي تعليقات إضافية وملاحظاتك حول شعورك كأحد الوالدين حين ترى طفلك يشارك الألعاب عبر الإنترنت	Long answer text.....
Restrictions and Rules (RR)	There is no worry if a child has restrictions and rules when playing online games | لا يوجد قلق إذا تم وضع قيوداً وقواعداً على الطفل عند ممارسة الألعاب عبر الإنترنت *	Agree / Disagree
	What kind of restrictions and rules do you suggest? Is it about the time she/he spends in playing? Or protection from psychological and physical harm? Or both? | ما هو نوع القيود والقواعد التي تقترحها؟ هل فيما يتعلق بالوقت الذي يقضيه في اللعب؟ أم بالحماية من الضرر النفسي والجسدي؟ أو كلاهما؟ *	Short answer text.....

E. Analysis

1) Structural equation modeling: This study used the modeling of the structural equation, which is a statistical method for testing and estimating causal relationships using a set of statistical data and qualitative causal assumptions. Addictive causal models and linear support theoretically can be tested using a tool in research called SEM like PLS, it was done by measured the items and then statistically tested. Usually one or more hypotheses are required that are represented as a model (see Fig. 1 and Table III).

2) Proposed model: The following model was developed.

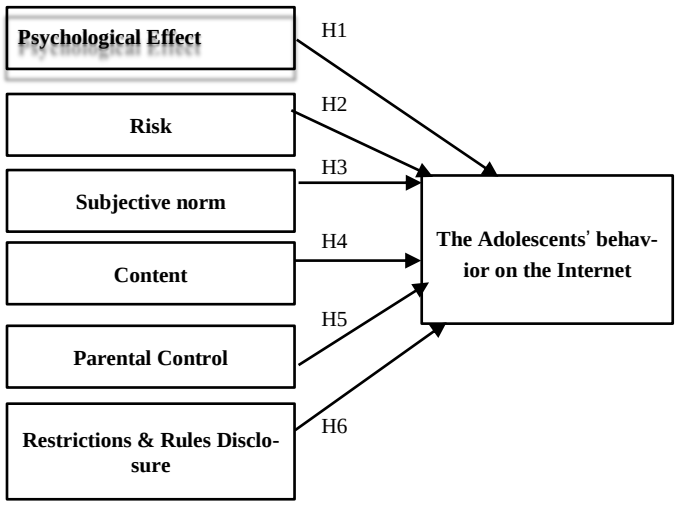


Fig. 1. Conceptual Framework and Hypotheses (Model A).

3) Proposed Hypothesis

TABLE III. DEFINITIONS OF T HYPOTHESIS

H1	*Internet gaming addiction fears will affect children's online behavior.*
H2	*Information disclosure will have a positive effect on the level of privacy concerns.*
H3	*Privacy risk will have a positive effect on online gaming addiction*
H4	*parental control will have a negative effect on the level of addiction to online games*
H5	*The subjective norm will have positive effect on teen addiction*
H6	*Internet content will have a negative impact on teenagers*

4) Data preparation for smartPLS: In this study, the participants' results were manually entered in Microsoft Excel and saved as xlsx format as shown in Fig. 2. This data set contains a sample size of 1500 without any missing values, invalid observations, or outliers. The first row of the Excel spreadsheet has the names of these indicators, such as (AHB, PE, R), to ensure that the software can correctly import the file data. Because SmartPLS can- not take the Excel file format directly, the file has been converted to the appropriate extension for the data set, which is .csv format.

5) Building the inner model: A proposed model for this study was designed (see Fig. 1) and it is a basic model based on which an internal model will be built. An internal model can be created easily in the SmartPLS program by representing latent variables by red circles, after drawing circles it is possible to change the default name by right- clicking on each latent variable. In order to link these variables together, arrows are drawn by clicking on the arrow symbol in the menu (see Fig. 3).

6) Building the outer model: The indicators are pulled from the "indicators" tab to the corresponding red circle, in order to link the latent variable and thus build an outer model. When the link is established, the color of the latent variable will change from red to blue. By right-clicking on the blue variable, indicators can be easily moved on the screen using the "Align Top / Bottom / Left / Right" function (see Fig. 4).

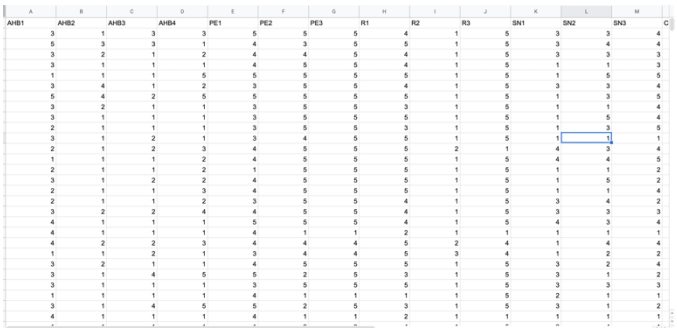


Fig. 2. Dataset from our Survey.

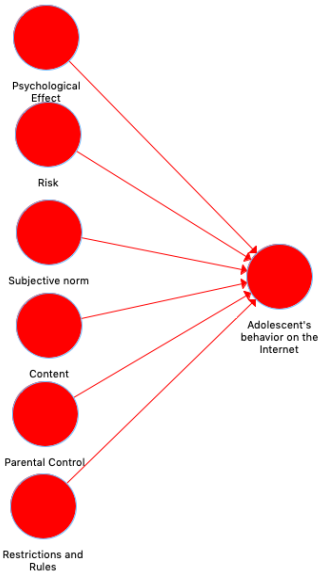


Fig. 3. Building the Inner Model.

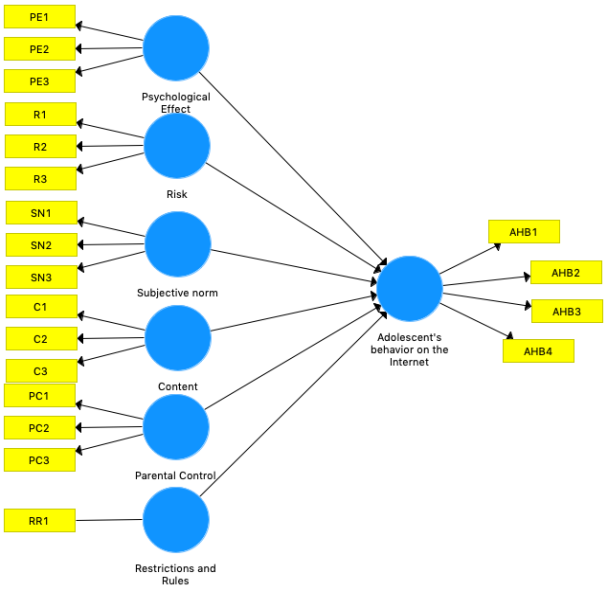


Fig. 4. Building the Outer Model.

7) Running the Factor-Modeling Estimation: The factor modeling procedure can be implemented, by going to the "Calculate" menu and choosing "PLS Algorithm" and then successfully linking the indicators and latent variables together in the SmartPLS with no red circles and arrows. As a result that there is no missing value for the data set in this study, we proceed directly to configure the settings of the PLS algorithm with the following parameters:

a) Weighting Scheme will be: Factor Weighting Scheme.

b) Data Metric will be: Mean 0, Variance 1.

c) Maximum Iterations will be: 300.

d) Abort Criterion will be: 1.0E-5.

e) Initial Weights will be: 1.0 (see Fig. 5).

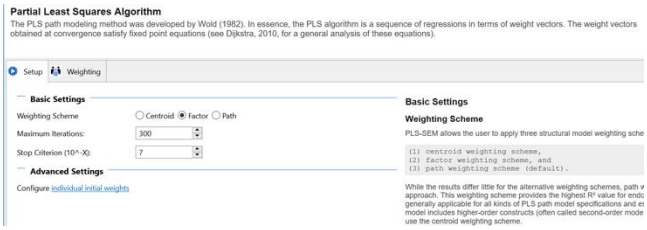


Fig. 5. Configuring the PLS Algorithm.

8) Assessing the PLS Output: SmartPLS text-based report estimates provide path modeling that can be accessed via the Report menu not only in the Modeling window (see Fig. 6).

9) Structural model evaluation: Then, we validate the suggested model for modeling the process specified in the hypotheses.

10)Inner model path coefficient sizes and significance: Results are shown based on the suggested inner model that the Restrictions and Rules has the strongest effect on Adolescents' behavior on the internet (0.261), followed by Parental Control (0.119), subjective norms (0.046), Psychological Effect (-0.016), Content (-0.245) and the weakest effect is Risk on (-0.505). Thus we can conclude that Restrictions and Rules, Parental Control and subjective norms are moderately strong predictors of Adolescents' behavior on the internet, but the opposite of the factors Psycho- logical Effect, Content and Risk.

11)Outer model loadings: The following table shows the correlations between the latent variable and indicators in its outer model (Table IV).

12)Indicator reliability: All other research considers that the reliability and validity of the latent variables are essential to complete the examination of the structural model. The above table outlined the various items of reliability and validity that must be examined and reported when performing the PLS-SEM.

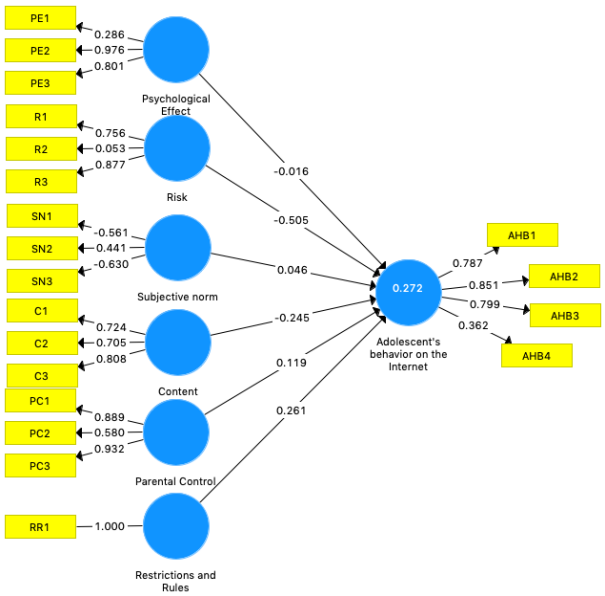


Fig. 6. PLS Results.

TABLE IV. RESULT SUMMARY FOR REFLECTIVE OUTER MODELS

	AVE	CR	Cranach's Alpha	Item Loading	Mean	Load- ing
PE	0.668	0.732	0.509	PE1 PE2 PE3	3.01 3.64	0.875 0.896
R	0.841	0.941	0.905	R1 R2 R3	3.25 3.35 3.56	0.930 0.921 0.900
SN	0.780	0.914	0.859	SN1 SN2 SN3	3.99 3.85 3.93	0.883 0.884 0.882
C	0.574	0.799	0.641	C1 C2 C3	2.55 3.18 3.21	0.607 0.810 0.835
PC	0.698	0.874	0.788	PC1 PC2 PC3	3.80 3.37 3.55	0.822 0.836 0.847
RR	0.590	0.732	0.354	RR1	2.88	0.925

Composite Reliability (CR): composite ratability effect between 0.7 and 0.8 is considered normal, if either the value does not exceed -0.5 or 0.06 then there is a problem. 0.9 or above is great.

Cronbach's Alpha: It is always considered valid except if the value in the study is less than 0.5.

13)Checking structural path significance in bootstrapping: To test the importance of both inner and outer models, we create T-statistics by SmartPLS and using a procedure called bootstrapping. Let's do this by selecting "Bootstrapping" from the "Calculate" menu (see Fig. 7 and Table V).

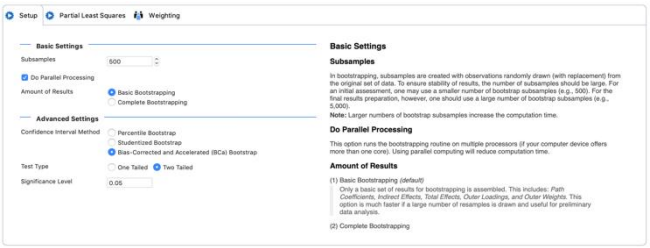


Fig. 7. Bootstrapping Algorithm.

TABLE V. RESULTS OF THE STRUCTURAL MODEL AND HYPOTHESES TESTING

Hs	**Associations**	**Coefficients**	**Supported**
H1	PE > AHB	0.199	Yes
H2	R > AHB	0.546	Yes
H3	SN > AHB	0.681	Yes
H4	C > AHB	0.254	Yes
H5	PC > AHB	0.099	Yes
H6	RR > AHB	0.014	Yes

If the effect between two constructs in the model is less than 0.05, there are not support, Otherwise sup- port. The findings of the study indicate that:

H1:PE > AHB is support.

H2: R > AHB is support.

H3: R > AHB is support.

H4: C > AHB is support.

H5: PC > AHB is support.

H6: RR > AHB is support.

14)Study demography: Below are the results of the pre-test questionnaire (see Table VI):

TABLE VI. STUDY DEMOGRAPHY

Demography Variable	Demography Classification	Frequency
Gender	Female Male	990 510
Age	18 - 24 25 - 34 35 - 44 45+	24 600 665 211
Education	High school Bachelor No degree Others	412 876 186 26
Parent?	Yes No	1500 0
Children	1 2 3 4 5+	321 550 405 190 34

III. RESULT

The Adolescents' behavior on the Internet, Psychological Effect, Risk, Subjective Norm, Content, Parental Control, and Restrictions and Rules, Scale was used in our evaluation as an overall score. The questions from this survey were always completely relevant, making this test rather practical on these points. We have used it in its broadest sense as we try to find the factors that lead to addictive behavior in adolescents in online games on a more formal and uniform level. In the previous part, we saw the analysis of the questionnaire using smart SPL tool and in the next part we will dis- cuss the hypotheses through the view of the end-users who participated in the performance of the experiment:

A. Psychological Effect

The study found that Information disclosure had a positive effect on the Privacy concerns (β = 0.199) (H1).

The result partly showed the extent to which parents are concerned about the negative effects of either the bad games or the bad people they interact with on the Internet. With average scores, we found that overall scores were varied, although they were closely related to each other, which can be seen through an illustration in the previous section. The difference in scores is relative where the participants' total support was 80% concerned that their children were gaming online.

B. Risk

The study found that Information disclosure had a positive effect on the Privacy concerns (β = 0.546) (H2).

The result shows that participants' awareness of risks poses a positive impact on their perception of the risk of their adolescent's addiction to game online. This result confirms the hypothesis of this study. This indicates that the greater the awareness of parents in monitoring their children, the fewer risks involved. It was recorded that there was a similarity in the score by all the participants, who recorded that they were interested in the nature of the games their children interacted with online.

C. Subjective Norm

The study found that Information disclosure had a positive effect on the Privacy concerns (β = 0.681) (H3).

The result of the factor shows that subjective standards will have a negative impact in relation to privacy concerns. As this result appears in a varying percentage between that the children of the participants tend to share their information, while the people who interact with them do not show all of their information, this may be because users who voluntarily dis- close personal information are associated with their behavior as a result of the confidence they have gained with other users of the network. This finding also con- firms the results of previous studies that found subjective criteria positively associated with privacy concerns.

D. Content

The study found that Information disclosure had a positive effect on the Privacy concerns (β = 0.199) (H4).

The result in the content factor showed a slight difference between the opinions of parents on the extent of their control over the content, but the highest percent- age of 85% stressed the importance of controlling the content of the games, as this confirmed that they are aware of the risks and implications of that, so they emphasized their use of blocking programs.

E. Parental Control

The study found that Information disclosure had a positive effect on the Privacy concerns (β = 0.254) (H5).

The result here confirms that the participants have strict control over their children and the study con- ducted in different societies may show close results. With a difference of 90% of parents are aware of the risks that their children can face and how important it is to monitor their children and allow them to game online under their supervision.

F. Restrictions and Rules

The study found that Information disclosure had a positive effect on the Privacy concerns (β = 0.099) (H6).

The last factor is the Restrictions and Rules, which gave a difference in the result by the participants, where 80% of the participants recorded their support for establishing rules and restrictions when children game online, and some participants commented by 7.4% the importance of establishing restrictions and rules to protect against psychological harm, in When 3.6% of the respondents recorded the importance of establishing restrictions and rules to protect against physical harm, while the highest vote was 73%, the importance of establishing restrictions and rules to protect against psychological and physical harm together.

1) The Adolescents' behavior on the Internet: The result recorded by the participants in this factor confirms the importance of the role of parents in reducing the risks of their children using the Internet, as the biggest difference in the result was given by the participants in the first point of the adolescent behavior factor on the Internet, who recorded that

they have no objection to the use of their children The Internet in the event that there is no danger to them, and within certain time limits. The result of this factor shows that there is nothing wrong with using the Internet under certain conditions that parents place on their children such as time as well as the element of safety, this result is consistent with what we have reached, and the reason may be that our participants may have strict control over their children.

IV. Conclusion

The main objective of this study was to reduce the risk of adolescent addiction in online games by under- standing and analyzing factors that influence their behavior in order to achieve the goal, and this study was developed as a model and measured adolescent perceptions from a parent's perspective. The study uses smartPLS to analyze results. This study discusses the hypothesis and the results of the research. Finally, it ends with contributions and future work.

The research question is: What factors have influenced teen addiction to online games? To answer this question, some objectives were raised as follows:

- Study the primary role of parents in allowing their children to use the Internet and monitor their behavior while playing online, and what procedures are followed accordingly.
- To study parents' awareness of their children's behavior while playing online.
- Focus on adolescents when they use the Internet for games by reducing the risks of influencing their behavior.
- Provide a comprehensive explanation of how to assess factors that influence addiction behavior, especially adolescents, in order to help protect them from the psychological harm that many games may cause.
- To describe a method for measuring parenting concerns about their children's addiction to Internet games through an effective quantitative research application.

This study was designed to serve the community in protecting adolescents from the harm of internet gaming addiction. In addition, the study contributes to educating and awareness parents about the importance of protecting their children from the dangers of the Internet by contacting them and obtaining their opinions on the behavior of teenagers on the Internet. This study provides support for parents to make the Internet safer and more secure for their children. Since this study provided observations related to protecting the privacy of information for adolescents and examining some of the factors that can influence their behavior, these topics will be of interest to researchers and practitioners in the public domain. This study inevitably faced many limitations. We will then summarize the limitations of this study and identify proposed further improvements for future research. This study needs to identify more factors. Some items did not achieve the desired value resulting in deletion. Depending on the limitations we mention, future work can be: The study may involve and test more factors like trust, trust in the Internet, and the risk of making friends while gaming online.

Acknowledgment

I would like to thank Allah for all the blessings that I have. A special thanks to supervisor Dr. Hedia Zardi for her help and for her directions for this paper.

References

[1] S. Livingstone and D. R. Brake, "On the rapid rise of social networking sites: New findings and policy impli- cations," Adolescents & society, vol. 24, no. 1, pp. 75-83, 2010.

[2] F. Annansingh and T. Veli, "An investigation into risks awareness and e-safety needs of adolescents on the in- ternet: a study of Devon, UK," Interactive Technology and Smart Education, vol. 13, no. 2, pp. 147-165, 2016.

[3] M. O. Lwin, A. D. Miyazaki, A. J. Stanaland, and E. Lee, "Online usage motive and information disclosure for preteen adolescents," Young Consumers, vol. 13, no. 4, pp. 345-356, 2012.

[4] S. Bannon, T. McGlynn, K. McKenzie, and E. Quayle, "The Internet and young people with additional support needs (ASN): Risk and safety," Computers in Human Behavior, vol. 53, pp. 495-5032015.

[5] A. Alkhalifah and J. D'Ambra, "Identity Management Systems Research: Frameworks, Emergemce, and Fu- ture Opportunities," in ECIS, 2015.

[6] J. Bryce and J. Fraser, "The role of disclosure of person- al information in the evaluation of risk and trust in young peoples' online interactions," Computers in Hu- man Behavior, vol. 30, pp. 299-306, 2014.

[7] E. Aghasian, S. Garg, L. Gao, S. Yu, and J. Mont- gomery, "Scoring users' privacy disclosure across multi- ple online social networks," IEEE access, vol. 5, pp. 13118-13130, 2017.

[8] K. Li, Z. Lin, and X. Wang, "An empirical analysis of users' privacy disclosure behaviors on social network sites," Information & management, vol. 52, no. 7, pp. 882-891, 2015.

[9] W. Shin and N. Ismail, "Exploring the role of parents and peers in young adolescents' risk taking on social networking sites," Cyberpsychology, Behavior, and So- cial Networking, vol. 17, no. 9, pp. 578-583, 2014.

[10] T. Buchanan, C. Paine, A. N. Joinson, and U. D. Reips, "Development of measures of online privacy concern and protection for use on the Internet," Journal of the American society for information science and technolo- gy, vol. 58, no. 2, pp. 157-165, 2007.

[11] Z. De Souza and G. N. Dick, "Information disclosure on MySpace–the what, the why and the implications," Pas- toral Care in Education, vol. 26, no. 3, pp. 143-157, 2008.

[12] N. K. Lankton, D. H. McKnight, and J. F. Tripp, "Face- book privacy management strategies: A cluster analysis of user privacy behaviors," Computers in Human Behav- ior, vol. 76, pp. 149-163, 2017.

[13] C. S. Silva, G. A. Barbosa, I. S. Silva, T. S. Silva, F. Mourão, and F. Coutinho, "Privacy for Adolescents and Teenagers on Social Networks from a Usability Perspec- tive: A Case Study on Facebook," in Proceedings of the 2017 ACM on Web Science Conference, 2017, pp. 63-71: ACM.

[14] K. Li, X. Wang, K. Li, and J. Che, "Information privacy disclosure on social network sites: An empirical investi- gation from social exchange perspective," Nankai Busi- ness Review International, vol. 7, no. 3, pp. 282-300, 2016.

[15] X. Zhang, L. Zhang, and C. Gu, "Security Risk Estima- tion of Social Network Privacy Issue," in Proceedings of the 2017 the 7th International Conference on Communi- cation and Network Security, 2017, pp. 81-85: ACM.

[16] J. Bremer, "The internet and adolescents: advantages and disadvantages," Child and Adolescent Psychiatric Clinics, vol. 14, no. 3, pp. 405-428, 2005.

[17] J. Wojniak and M. Majorek, "Adolescents in internet space–the European Union policies on adolescents's safety online," in SHS Web of Conferences, 2016, vol. 26, p. 01048: EDP Sciences.

BOTNETs: A Network Security Issue

From Definition to Detection and Prevention

Engr. Umar Iftikhar[1], Engr. Kashif Asrar[2], Dr. Maria Waqas[3], Dr' Syed Abbas Ali[4]
Computer and Information Systems Engineering Department
NED University of Engineering and Technology
Karachi, Pakistan

***Abstract*—With the technological advancements in the field of networking and information technology in general, organizations are enjoying the technological blessings and simultaneously under perpetual threats that are present in the form of attacks, designed especially to disable organizations and their infrastructure, as the gravest cyber threats in recent times. Compromised computers or BOTNETs are unarguably the most severe threat to the security of internet community. Organizations are doing their best to curb BOTNETs in every possible way, spending huge amount of their budget every year for available hardware and software solutions. This paper presents a survey on the security issues raised by the BOTNETs, their future; how they are evolving and how they could be circumvent to secure the most valuable resource of the organizations which is data. The compromised systems may be treated like viruses in the network which are capable of performing substantial loss to the organization including theft of confidential information. This paper highlights the parameters that should be considered by the organizations or Network administrators to find out the anomalies that may point to the presence of BOTNET in the network. The early detection may reduce the impact of damage by taking timely actions against compromised systems.**

***Keywords*—BOTNET; malware; drones; zombies; threats**

I. INTRODUCTION

The emerging and rapidly growing internet era has led the mankind to an exceptional world of facilitation where one can find endless social and economic benefits. On the other hand, this technology has introduced numerous challenges. Despite of various advanced security methodologies, the network security threats are continuing to evolve day by day.

Network security can be described as the actions taken for the protection of the network. Usually, these actions safeguard the usability, reliability, integrity and safety of the data and network. Operative security in networks is capable of addressing various types of threats as well as prevents them from entering or spreading into the network.

There are numerous types of threats that are being faced by network security. Some of them are Trojan horses, viruses and worms, spyware, malware, BOTNETS, zero- hour attacks, hacker attacks, DoS (Denial of Service) attacks, data interception and theft, identity theft, etc. [1].

This paper deals with a review of a very important network security issue that is BOTNET. The paper begins with the demonstration of network security issues in section [I] and explaining some of the important threats in network security due to the BOTNETS. Botnet administrators can moreover run them as a commercial operation for creating a distress for the organization especially those who rely more on the IT infrastructure for their business continuity in Section II. Threats immerged due to the existence of BOTNETS are mitigated by major operations that require significant worldwide participation in Section III.

A. What is BOTNET?

The terminology BOTNET is extracted from the term bot that is the short form for the robot. Intruders use different tricky techniques to distribute malicious software that is capable of converting a computer into bot or zombie. When such a situation arises in which a computer is being controlled not by user but by a hacker, it performs several suspicious tasks on internet without the knowledge of the user.

In other words, the collections of several computers that are associated to perform suspicious tasks using malicious software are termed as BOTNETs.

Attackers usually utilize the bots to infect huge number of computers. These computers form a group known as BOTNET. These zombies can be utilized to spread out spam emails, distribute viruses, attack the servers, and commit various kinds of fraud and cybercrimes [2].

The size of BOTNET is variable that is it can be small or large. The size of BOTNET depends upon the sophistication and complexity of the bots that are used. A large BOTNET consist of tens and hundred thousand zombies. While on the other hand, a smaller BOTNET comprised of a few thousand of zombies.

The owner whose computer has become the zombie, do not know that the affected computer and all of its resources are being remotely controlled, subjugated and misused by an single or a group of malware runners that uses Internet Relay Chat (IRC) as a substantial tool for these malicious attacks. There are several kinds of malwares and malicious software and applications that have already trapped and are continuing to trap the internet. Large bots use their own spreaders to spread the viruses while smaller kinds of bots do not possess such capabilities. The whole scenario of BOTNETs is illustrated in Fig. 1 [3].

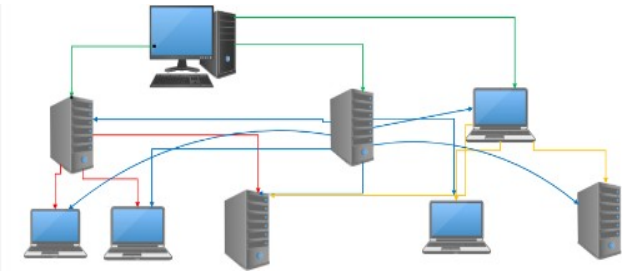


Fig. 1. BOTNET Central Command.

B. Formation of BOTNET

BOTNETs are formed with a use of an IT tool known as "drones". These drones penetrate into an open computer through internet that has low security measures. When a drone penetrates into a computer that computer becomes a "Zombie Computer". Now that zombie computer acts like a BOTNET that is being controlled centrally by a BOTNET owner. The zombie computer runs malicious software itself and the owner of infected computer doesn't know that the computer is being trapped and became a BOTNET [4].

The whole group of BOTNETs is being controlled centrally by the herder that is actually a BOTNET owner. The more increment in the size of the BOTNETs, the larger is the impact of BOTNETs on internet. In other words, herder gets stronger and stronger as more and more computers are affected by this malware 5].

C. Propagation of BOTNET

The BOTNETs propagate through various bot software that contains spreaders. This spreader in the bot software automates the job of IP addresses scanning for various susceptible software holes. Once the holes are found by software, the unsecured and defenseless machines are attacked and infected by this bot software, and this pattern continues thus resulting in increasing the number of infected computers. With each new machine infected by the drones, the BOTNET becomes more and more powerful to infect more machines. The one and only difference between a bot and worm is the presence of a unifying control mechanism [6].

Command and Control Tools for BOTNET:

A large number of infected machines are useless without some controlling mechanism. The command and control (C&C) mechanism provides the interface between the BOTNET and the herder. The C&C get commands from herder and control the bots.

The BOTNETs are traditionally been controlled using Internet Relay Chat (IRC). This framework is the most popular one because of its easiness, tractability and ease of administration. IRC is a global and commonly used communication standard over the internet and can be easily modified for any specific purpose. Bot software has the tendency to connect the infected computer to IRC server and accept instructions from centrally controlled channel (herder). The herders have rights to use current chat service and network or they can implement their own separate server for control by using the IRC daemon [7].

D. Mitigation of BOTNET

In today's high tech era, where internet has penetrated into the lives of humans and made the world a global village, BOTNETs are of major concerns and can be very dangerous if they are in a very large number. With drone population counting as 60,000 – 80,000, the access and the control that herders can have over the largest network giants is incredible and gigantic.

Therefore, the best possible way to diminish BOTNETS is to prevent and block them from establishing at initial stages. If malware is controlled from propagation and infection into the system, BOTNETs would no longer remain the serious threat to the network security. The owner of the computer should take care of the system itself by properly patching and licensing the software and systems, otherwise their computers can be easily transformed into BOTNETs. The mitigation of BOTNETs is further illustrated in Fig. 4 [8].

II. Application of BOTNETs

BOTNET damages magnificently, the security of businesses and individuals the data and resources of an infected computer losses its legitimate user's control. Most of the users store their sensitive information on their personal machines. If the security of this machine is compromised, the attacker can easily harvest that sensitive and confidential data. Bot herders used to sell or rent their BOTNETs to those who want to perform hacker activities.

The strong penetrating capability and strength of BOTNETs, give attacker more and more power on the internet. With the increase in number of BOTNETs, the control over compromised systems of the herder becomes stronger thus performing more complicated, advance and typical activities that internet has never seen before.

Some of the severe applications of BOTNETs are discussed below:

A. Click Fraud

BOTNETs can be utilized to engage in Click Fraud. In this type of scam, the bot software used to navigate different websites on browser and automatically click on advertisements. Now consider about a herder having a bot network of several thousand computers and stealing a large amount of money from online advertisement organizations that pay small amount on each click. With a large network, each click for few times, returns heavy amount of money. As the clicks are coming from each separate entity distributed across the globe, so investigators can't find out that this is a scam [9].

B. Distributed Denial of Service (DDoS)

BOTNETs are used to remunerate confrontation on various computers over the network accessing the internet by completely trapping and saturating its bandwidth and various other resources. Such DDoS attacks can disable the access the web pages for a long span of time. While considering the financial organizing, this delay of accessibility places a marvelous and enormous burden on financial operators that are unable to service their customers.

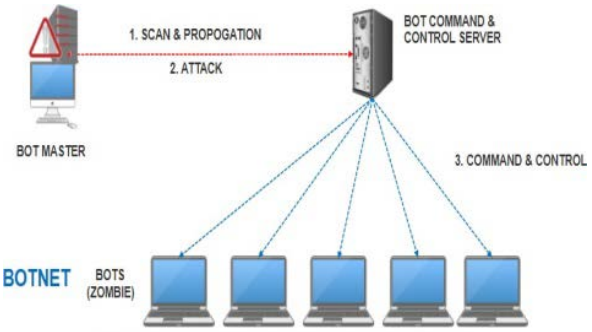


Fig. 2. Distributed Denial of Service (DDoS).

Another type of attack comes under the umbrella of DDoS where the attackers demand for the payment to free the attacked resources and allow the traffic to flow again. This type of attack is known as Extortion attacks. The complete illustration can be referred from Fig. 2 [10].

C. Key logging and Mass Identity Theft

Key logging is the technique that is used to record the sequence of the key pressed. This can be done by installing the key logger software. This is encryption software that is used to gather the sequence of keys pressed by the user. This includes the personal information of the user and passwords. This is one of the important reasons behind the massive PayPal account theft for past several years.

Bots can also be utilized by the attacker as an agent for mass identity theft. This involves methods of phishing or pretending to be the agent of a company and enforcing the Client to give their personal information like password and credit card numbers. The phishing technique is implemented by spam emails in which a fake link is given for the renowned financial or online transaction website that traps the client to submit the personal information.

Despite of key logging, many bots allow the herder to completely access the file system. This enable the herder to modify and transfer any file, can read any personal document stored in the user's computer and can upload the malicious files [11].

D. Traffic Monitoring and Spamming

BOTNETs are utilized by using the TCP/IP proxy protocol for several applications of network. After the IP of a computer is compromised, bot commander can use this IP to propagate the massive spams, malware, phishing and fraud email to various email address. This is achieved by stealing an IP address of any bot and in conjunction with other bots, the bot commander send these massive spam emails.

Also, a zombie can act as a packet sniffer to monitor the traffic and ongoing activities over the network with the help of infected machines. Typically these sniffers look for the username and passwords for different accounts which a bot commander can use later for its personal interests [12].

E. Warez

Another application of the BOTNETs is Warez. Warez is technique in the world of hacking that is being used for stealing the licenses of the software or applications. BOTNETs possess the tendency to steal, store or propagate Warez. They can do this by scanning the hard drives of the infected machines looking for the software and applications that are licensed. After successful searching, the herder can easily transfer or duplicate that license and can distribute over the internet thus violating the copyrights of the software. The illustration of Warez in Fig. 3 [13].

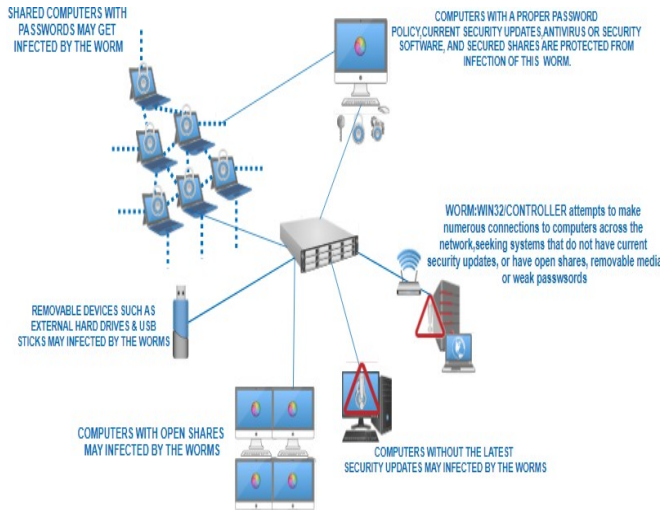


Fig. 3. Illustration of Warez.

III. BOTNET Detection and Prevention

The detection and prevention of BOTNETs is of immense importance as it is one of the major issues of the network security.

A. Detection of BOTNETs

The detection of BOTNETs is typical and very much difficult. The reason is that bots used to operate within a network or infected machine without prior information to the owner [15]. However there are some common indications while the routine functioning of the machine from which the owner can identify whether the machine is under attack or not. These indications are listed below:

1) Usage of Internet Relay Chat (IRC) by monitoring the traffic.

2) Time to time multiple connection attempts with known C&C servers.

3) Multiple machines generating identical DNS requests.

4) Outgoing SMTP traffic becomes very high.

5) Unexpected popups.

6) Slow processing speed, processor utilization is at extreme while no certain heavy application is running.

7) High and repetitive spikes in the data traffic over the network. Particularly Port 6667 that is used for IRC. Port 25 usually used for spam emails. Port 1080 used by proxy servers.

8) Outbound messages send by user without any knowledge like on email, instant messengers, social media, etc. with the name of owner.

9) Internet access becomes unavailable for no reason. Web pages couldn't be accessed [5, 7, 14, 16].

B. Prevention from BOTNETs

There are various method, techniques and measures that the owner of the infected machine can take to avoid the BOTNET attacks. The major concerned in transforming a normal machine to a BOTNET is malware. The measures described below mainly focuses on how to avoid malware. In other words, if malware are being stopped from penetration into a machine the chances of transforming a machine into BOTNET gets lower extensively.

Recommended practices from different network security providers are summarized below:

1) Installation and enabling of Windows Firewall: The users should always install a recommended firewall and must keep that firewall enabled especially when accessing the internet. As firewall block many network based threats.

2) Disabling the Auto-Run option: The auto-run option in the windows must be disabled or it must be enabled with the permission of user. The user must know that which software is installing on the machine. If auto-run option is enabled, it will automatically install the software without the permission of the user.

3) Breaking password Trusts: While taking in consideration the local accounts, especially the account with a local network administrator, it is really important to isolate and eliminate the threats by making a judicious policy for the implementation of the local network. By disabling the computer's capability to automatically connect to the other networks that are closer in the path, the property of BOTNET to make itself multiple will be eliminated.

4) Network Compartmentalization: In various computing environments, the workstations don't communicate with each other within the same vicinity or the departments. Disabling this feature will help in prevention of BOTNETs spreading feature up to a great extent. The network administrators should establish VLANs and ACLs between several sub networks to minimize the exposure. Although this approach is not much appreciable, but it fits in the environment where there is a mix voice and data communication.

5) Providing Least Privilege: A central control mechanism must be implemented in a network where every user must not be given the administrative rights. This approach can minimize the propagation of the malware to infect the individual machines as the user have very limited and specific privileges.

6) Installation of Host-based Intrusion Prevention Application: IT managers should focus on taking additional measures for the security and protection by adding vulnerability to the specified network layers for example at points of contact between specific hardware and the software. Although this approach cannot fix the technicalities but still it prevents the system not to be exploited easily. Also these types of security applications are very expensive and very much typical to deploy.

7) Enhancement in the Monitoring of Traffic over Network: The monitoring of the network traffic can play a very essential part in preventing the BOTNET attacks. The enhancement and routine scheduling of the network traffic monitoring is essential and Network administrator must concentrate of that very seriously.

8) Filtering of data that is outgoing from Network: BOTNETs use to communicate with bot commander through remote severs. The agencies must stop these communications by prohibiting the unwanted traffic leaving from the network. A very essential tool for this purpose is Egress Filtering. Agencies should deploy Data Loss Prevention (DLP) solution.

9) Usage of Proxy Servers: Although it is impossible to block all the outbound traffic, but forcing the outbound traffic using a proxy server provides agencies a secondary choke point to monitor and control the out-bound data that is accessing the web.

10)Monitoring of queries generated by DNS: The workstations responds to DNS queries in a way is a pre sign of warning that workstation can get infected by a drone. Particularly, the responses with very low time-to live values should be seriously monitored by the network administrator. Monitoring helps the network administrator to act early as the attack gets stronger infecting a large are or might be whole network [17 – 21].

IV. Conclusion

BOTNETs are one of the most severe threats in the domain of network security. This paper addressed some of the attention-grabbing aspects of BOTNETs and provided a viewpoint as to why BOTNETs are so much dangerous and harmful in the field of network security. Hence it becomes very essential to aware the users regarding the threats that can be caused from this type of malware.

After the study and analysis, it was concluded that creation and prevention of BOTNETs can be considered as the cold war between the intruders that creates the BOTNETs and preventers that counters the attacks. The security experts are focusing on the prevention of new attacks making use of regression techniques in unsupervised learning algorithms to identify the malicious traffic pattern.

References

[1] Geer, "Malicious bots threaten network security," Computer, vol. 38, no. 1, pp. 18–20, Jan. 2005.

[2] Z. Zhu, G. Lu, Y. Chen, Z. Fu, P. Roberts, and K. Han, "Botnet Research Survey," in Computer Software and Applications, 2008. COMPSAC '08. 32nd Annual IEEE International, 2008, pp. 967–972.

[3] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna, "Your botnet is my botnet: analysis of a botnet takeover," presented at the Proceedings of the 16th ACM conference on Computer and communications security, 2009, pp. 635–647.

[4] "R. Vogt, J. Aycock, and M. Jacobson, Jr. Army of Botnets, 14th Annual Network and Distributed System Security Symposium, 2007, pp. 111-123. Reprinted in Chapter 10 of Botnets: A Cyber Threat, S. Puneet, ed., Icfai University Press, 2008, pp. 171-199." .

[5] Y. H. Moon, E. Kim, S. M. Hur, and H. K. Kim, "Detection of botnets before activation: an enhanced honeypot system for intentional infection and behavioral observation of malware," Secur. Commun. Netw., vol. 5, no. 10, pp. 1094– 1101, Oct. 2012.

[6] D. Dagon, C. C. Zou, and W. Lee, "Modeling Botnet Propagation Using

Time Zones.," in ResearchGate, 2006.

[7] W. T. Strayer, R. Walsh, C. Livadas, and D. Lapsley, "Detecting Botnets with Tight Command and Control," in Proceedings 2006 31st IEEE Conference on Local Computer Networks, 2006, pp. 195–202.

[8] "Measurements and Mitigation of Peer-to-Peer-based Botnets: A Case Study on Storm Worm." [Online]. Available: https://www.usenix.org/legacy/event/leet08/tech/full_papers/holz/holz_html/. [Accessed: 29-Sep-2015].

[9] B. J. Jansen, "Click Fraud," Computer, vol. 40, no. 7, pp. 85–86, Jul. 2007.

[10] F. Lau, S. H. Rubin, M. H. Smith, and L. Trajkovic, "Distributed denial of service attacks," in 2000 IEEE International Conference on Systems, Man, and Cybernetics, 2000, vol. 3, pp. 2275–2280 vol.3.

[11] M. Bailey, E. Cooke, F. Jahanian, Y. Xu, and M. Karir, "A Survey of Botnet Technology and Defenses," in Conference For Homeland Security, 2009. CATCH '09. Cybersecurity Applications Technology, 2009, pp. 299– 304.

[12] H. R. Zeidanloo, A. Bt Manaf, P. Vahdani, F. Tabatabaei, and M. Zamani, "Botnet detection based on traffic monitoring," in 2010 International Conference on Networking and Information Technology (ICNIT), 2010, pp. 97–101.

[13] D. Dagon, G. Gu, C. P. Lee, and W. Lee, "A Taxonomy of Botnet Structures," in Computer Security Applications Conference, 2007. ACSAC 2007. Twenty- Third Annual, 2007, pp. 325–339.

[14] K. Anestis, R. Brian, and H. David, "Wide-scale Botnet Detection and Characterization."

[15] P. Barford and V. Yegneswaran, "An Inside Look at Botnets," in Malware Detection, M. Christodorescu, S. Jha, D. Maughan, D. Song, and C. Wang, Eds. Springer US, 2007, pp. 171–191.

[16] W. Lee, C. Wang, and D. Dagon, Botnet Detection: Countering the Largest Security Threat. Springer Science & Business Media, 2007.

[17] K. A. Cole, R. L. Silva, and R. P. Mislan, "All Bot Net: A Need for Smartphone P2P Awareness," presented at the International Conference on Digital Forensics and Cyber Crime, 2011, pp. 36–46.

[18] "Indian Journals." [Online]. Available: http://www.indianjournals.com/ijor.aspx?target=ijor:ij mt&volume=1&issue=3&article=004. [Accessed: 29- Sep- 2015].

[19] M. R. Thakur, D. R. Khilnani, K. Gupta, S. Jain, V. Agarwal, S. Sane, S. Sanyal, and P. S. Dhekne, "Detection and prevention of botnets and malware in an enterprise network," Int. J. Wirel. Mob. Comput., May 2012.

[20] N.-Y. Lee and H.-J. Chiang, "The research of botnet detection and prevention," presented at the Computer Symposium (ICS), 2010 International, 2010, pp. 119– 124.

[21] "The Analysis of Botnet Transmission Model and the Prevention & Cure Methods--《 Journal of Changzhou Institute of Technology 》 2008 年 06 期 ." [Online]. Available: http://en.cnki.com.cn/Article_en/CJFDTOTAL- CZGB200806009.htm. [Accessed: 29-Sep-2015].

Assessment of Surface Water Quality on the Upper Watershed of Huallaga River, in Peru, using Grey Systems and Shannon Entropy

Alexi Delgado[1], Jharison Vidal[2], Jhon Castro[3], Jhonel Felix[4], Jorge Sáenz[5]
Faculty of Systems Engineering, Universidad de Ciencias y Humanidades, Lima, Perú[1]
Environmental Engineering Program, Universidad Nacional de Ingeniería, Lima, Perú[2, 3, 4, 5]

***Abstract*—The assessment of the quality of surface water is a complex issue that entails the comprehensive analysis of several parameters that are altered by natural or man-made causes. In this sense, the Grey Clustering method, which is based on Grey Systems theory, and Shannon Entropy, based on the artificial intelligence approach, provide an alternative to evaluate water quality in an integral way considering the uncertainty within the analysis. In the present study, the water quality on the upper watershed of Huallaga river was evaluated taking into account the monitoring results of twenty-one points carried out by the National Water Authority (ANA) analyzing nine parameters of the Prati index. The results showed that all the monitoring points of the Huallaga river were classified as not contaminated, which means that the discharges, generated by economic activities, are carried out through of treatment plants meeting the quality parameters. Finally, the results obtained can be of great help to the ANA and the regional and local authorities of Peru in making decisions to improve the management of the Huallaga river watershed.**

***Keywords*—Grey clustering; Huallaga river; Prati index; Shannon entropy; water quality**

I. Introduction

The Huallaga river watershed is one of the main watersheds of Peru and with a great potential of water resources, this due to the existence of a large number of lagoons, rivers, streams and springs, an important source of resources natural resources, food and work for the native communities and populated centers of the area being the main economic activities: agricultural, industrial, energy, mining and fishing for direct human consumption [1]. In addition, the benefits provided by the watershed have been diminished, due to its waters have been polluted by domestic wastewater, wastewater municipal, solid waste, as well as mining environmental liabilities, being a risk to public health [2].

This watershed, located on the Atlantic watershed, is one of the largest tributaries of the Marañon River and It is made up of the Lower Huallaga Inter-watershed, Paranapura, Middle Lower Huallaga Inter-watershed, Mayo Watershed, Middle Huallaga Inter-watershed, Biabo Watershed, Middle Upper Huallaga Inter-watershed, Huyabamba Watershed and Upper Inter-watershed Huallaga [1]. The assessment of the surface water quality will be carried out in the upper part of the watershed due to its environmental importance.

For the development of the assessment we will use the Grey Clustering method, as well as the Shannon Entropy. Flock Clustering is a method that is based on the theory of Grey systems, an approach within what is called Intelligence Artificial, so it has a great variety of applications [3]. For the case to be studied, we will use the center-point triangular whitenization weight functions (CTWF) method, which is applied in studies on water [4] or in the assessment of urban transport [5]. On the other hand, the weight method Entropy, based on Entropy of Shannon, is also an approach within artificial intelligence developed initially by Claude E. Shannon [6], this method was used to calculate the weights objectives of the assessment criteria within the CTWF method.

Therefore, our specific objective in this study is the classification of 21 monitoring points on the upper watershed of Huallaga river according to the water quality criteria, using the Grey Clustering method and the Entropy of Shannon.

In the present study, Section II details the CTWF method. Section III describes the case study, followed by the results and discussions of Section IV. The conclusions will be presented in Section V.

II. Methodology

In this section, we describe the CTWF method, which can be described as follows: first, suppose the area is set of m objects, a set of n criteria, and a set of s Grey classes, according to the sample value (i = 1, 2,..., m; j = 1, 2,…, n). Then, the steps of the CTWF method can be developed with the following points according to different research [3], [7] and [8].

A. Step 1: Determination of Center Points

The ranges of the criteria are divided into 5 Grey classes, and then their central points are $\lambda_1, \lambda_2, \dots, y\ \lambda_s$, this is determined by the Prati index.

B. Step 2: Dimension Removal

At this point it is assumed that there are objects for assessment and n criteria or Grey classes, which forms the following Matrix $Z = \{Z_{ij}; i = 1,2, \dots, m\ ; j = 1,2, \dots, n\}$. In that sense, it is normalized for each criterion $C_j (j = 1, 2, \dots, n)$. The value is normalized P_{ij}, which is calculated by Equation 1.

$$P_{ij} = \frac{z_{ij}}{\frac{\sum_{j=1}^{n} z_{ij}}{n}} \quad (1)$$

C. Step 3: Determination of the Triangular Functions and their Values

The Grey classes are expanded in the addresses of each parameter used and for this the index will be used as a reference Prati, who provided the data to measure quality, in this research Prati provides us with five 5 levels of quality for each parameter, so there will be five (5) functions for each parameter. The new sequence of points central is λ_1, λ_2,…y λ_5. For the class k = 1,2,3,4 and 5, j = 1, 2, …, n, for an observed value xij. The calculation is displayed of the CTWF by means of Equations 2, 3 and 4; and Fig. 1 shows the graph of the triangular functions.

$$f_j^1(x_{ij}) = \begin{cases} 1, \ x \ni [0, \lambda_j^1] \\ \frac{\lambda_j^2 - x}{\lambda_j^2 - \lambda_j^1}, \ x \in \langle \lambda_j^1, \lambda_j^2 \rangle \\ 0, x \in [\lambda_j^2, +\infty\rangle \end{cases} \quad (2)$$

$$f_j^k(x_{ij}) = \begin{cases} \frac{x - \lambda_j^{k-1}}{\lambda_j^k - \lambda_j^{k-1}}, \ x \ni \langle \lambda_j^{k-1}, \lambda_j^k] \\ \frac{\lambda_j^{k+1} - x}{\lambda_j^{k+1} - \lambda_j^k}, \ x \in \langle \lambda_j^k, \lambda_j^{k+1} \rangle \\ 0, x \in [0, \lambda_j^{k-1}] U [\lambda_j^2, +\infty\rangle \end{cases} \quad (3)$$

$$f_j^5(x_{ij}) = \begin{cases} \frac{x - \lambda_j^4}{\lambda_j^5 - \lambda_j^4}, \ x \in \langle \lambda_j^4, \lambda_j^5 \rangle \\ 1, \ x \in [\lambda_j^5, +\infty\rangle \\ 0, x \in [0, \lambda_j^4] \end{cases} \quad (4)$$

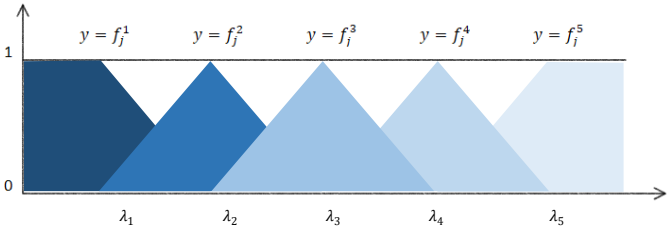


Fig. 1. CTWF According to Prati Scale.

D. Step 4: Determination of the Weight for each Criterion

In this step, the Shannon Entropy weight method is used. For everything P_i it is considered within a distribution of probability, Shannon developed the measure H, which satisfies the following properties [3], [7] and [9]:

- H is a positive continuous function
- If all p_i are equivalent and p_i=1/n, in this sense, H should be a monotonous increasing function of n.
- For all n ≥ 2, $H(P_1, P_2, \dots, P_n) = h(p_1 + p_2, p_3, \dots, p_n) + (p_1 + p_2) H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$

Shannon shows that only the functions that satisfy this condition are calculated by Equation 5.

$$H_{Shannon} = -\sum_{i}^{n} p_i \log(p_i)$$

$$Where: 0 \le p_i \le 1; \ \sum_{f=1}^{n} p_i = 1 \quad (5)$$

Around the entropy weight methodology, it can be demonstrated according to the following definition [3], [7] and [9]. As shown above, m objects are displayed for assessment, and n assessment criteria, which form the following matrix $x = \{x_{ij}; i = 1,2, \dots, m; j = 1,2, \dots n\}$. After that, the following stages continue.

1) Then the matrix $x = \{x_{ij}; i = 1,2, \dots, m; j = 1,2, \dots n\}$ is normalized by each criterion Cj. The normalization evaluates Pij are calculated by Equation 6.

$$f_j^1(x_{ij}) P_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}} \quad (6)$$

2) The entropy of each criterion is calculated by Equation 7, which was constructed based on Equation 6.

$$H_j = -k \sum_{i=1}^{m} P_{ij} ln(P_{ij}) \quad (7)$$

Where, K is a constant, k= $(\ln(m))^{-1}$

3) The degree of divergence of the intrinsic information in each criterion C_j is calculated by Equation 8.

$$div_j = 1 - H_j \quad (8)$$

4) In the weight entropy w_j of each criterion C_j is calculated by Equation 9.

$$w_j = \frac{div_j}{\sum_{j=1}^{n} div_j} \quad (9)$$

Where, w_j is equal to n_j

E. Step 5: Determination of the Clustering Coefficient

The clustering coefficient σ_i^k by objeto i, i = 1, 2, …, m, respect to the Grey classes k, k=1, 2, …, s, is calculated by Equation 10.

$$\sigma_i^k = \sum_{j=1}^{n} f_j^k(x_{ij}) n_j \quad (10)$$

Where $f_{j=1}^k(x_{ij})$ is the CTWF of the k-th grey class of j-th criterion, y n_j is the weight of criterion J, establish said weights the Shannon Entropy method will be used.

F. Step 6: Results using the Maximum Clustering Coefficient

Finally, we have the calculation of $max_{1 \le k \le s}\{\sigma_i^k\} = \sigma_i^k$, We decide which object belongs to Grey class k*. When there are several objects in Grey class k*, these objects can be ordered according to the magnitudes of their grouping coefficients integral.

III. Case Study

The analysis of the surface water quality was carried out in the upper part of the Huallaga river watershed, which is located in the central zone of Peru has an area of 89,416 km2 and a length of 1,168 km in a direction from south to north [1], which is represented in Fig. 2.

Fig. 2. Huallaga River Watershed, Peru.

A. Definition of Study Objects

For the assessment of the quality of the surface water of the upper watershed of Huallaga river , information was collected of 21 monitoring points obtained from the seventh monitoring of surface water quality carried out on 19November to December 20, 2019 by the Huallaga Water Administrative Authority and the Local Authorities of the Alto Huallaga, Tingo María, Huallaga Central, Alto Mayo and Tarapoto [1]. Which will be detailed in Table I and represented in Fig. 3.

TABLE I. MONITORING POINTS IN THE UPPER WATERSHED OF HUALLAGA RIVER

Point	Code	Coordinates (WGS 84)	
		East	*North*
1	RHual1	370546	8828639
2	RHual2	369690	8830701
3	RHual3	370966	8836230
4	RHual4	370755	8836925
5	RHual5	370959	8843646
6	RHual6	369913	8876992
7	RHual7	367820	8882470
8	RHual8	366866	8886415
9	RHual9	363585	8896395
10	RHual10	362973	8900714
11	RHual11	364061	8901392
12	RHual12	367093	8907020
13	RHual13	379125	8912553
14	RHual14	384109	8914086
15	RHual15	395255	8950625
16	RHual16	393159	8959917
17	RHual17	389447	8971679
18	RHual18	390103	8974251
19	RHual19	380015	9002651
20	RHual42	369474	8832777
21	RHual43	369809	8834456

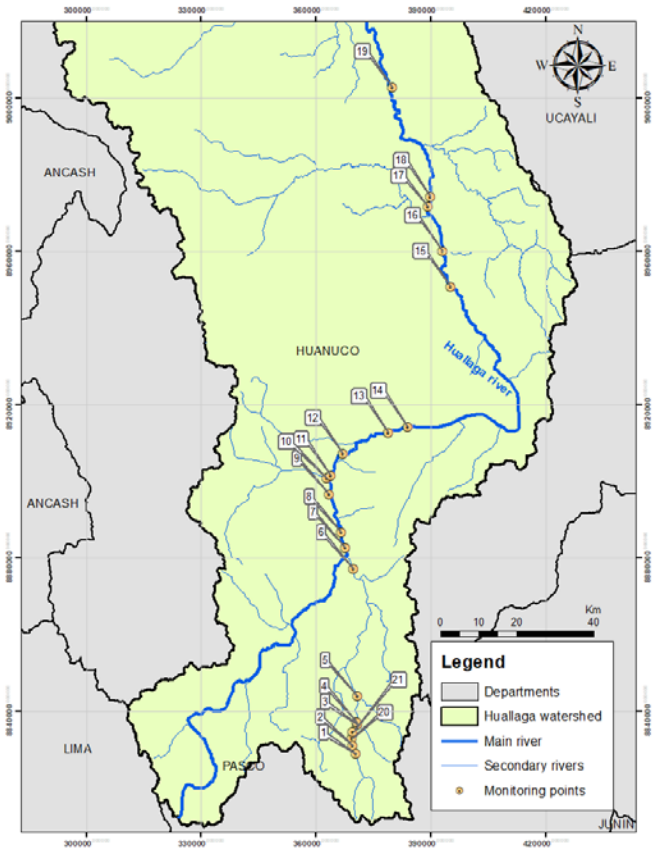


Fig. 3. Monitoring Points of Surface Water Quality in the Upper Watershed of Huallaga River.

B. Definition of Assessment Criteria

The assessment criteria for the present study are determined by the water quality parameters which are presented in Table II.

C. Definition of the Grey Classes

The classes for assessment are five and are based on the levels of water quality according to the Prati index, which are presented in Table III.

TABLE II. PRATI INDEX STANDARD DATA FOR THE ASSESSMENT OF WATER QUALITY

Criterion	Units	Notation
pH	0 -14	C_1
BOD	ppm	C_2
COD	ppm	C_3
Total Suspended Solids	mg/L	C_4
NH_3	ppm	C_5
NO_3	ppm	C_6
Cl	ppm	C_7
Manganese	ppm	C_8
Iron	ppm	C_9

TABLE III. PRATI INDEX STANDARD DATA FOR THE ASSESSMENT OF WATER QUALITY

Parameters	Quality Index Condition				
	Uncontaminated	*Acceptable*	*Moderately contaminated*	*Contaminated*	*Highly Contaminated*
pH	6.5-8.0	8.0-8.4	8.4-9.0	9.0-10.1	>10.1
BOD (ppm)	0.0-1.5	1.5-3.0	3.0-6.0	6.0-12.0	>12.0
COD (ppm)	0-10	10-20	20-40	40-80	>80
Total Suspended Solids (mg/L)	0-20	20-40	40-100	100-278	>278
NH_3 (ppm)	0-0.1	0.1-0.3	0.3-0.9	0.9-2.7	>2.7
NO_3 (ppm)	0-4	4-12	12-36	36-108	>108
Cl (ppm)	0-50	50-150	150-300	300-620	>620
Manganese (ppm)	0.00-0.05	0.05-0.17	0.17-0.50	0.50-1.00	>1.00
Iron (ppm)	0.0-0.1	0.1-0.3	0.3-0.9	0.9-2.7	>2.7

D. Calculations using the CTWF Method

The calculations based on the gray clustering method are presented below:

1) Step 1: Based on the Prati quality index, the central values of the parameters to be analyzed are obtained. These values are shown in Table IV.

2) Step 2: The non-dimensioned standard values for each parameter, according to the Prati index, were determined through the (1). These values are presented in Table V.

TABLE IV. CENTRAL VALUES OF THE PRATI INDEX PARAMETERS

Parameters	Uncontaminated (λ_1)	Acceptable (λ_2)	Moderately contaminated (λ_3)	Contaminated (λ_4)	Highly Contaminated (λ_5)
pH	7.25	8.20	8.70	9.55	10.40
BOD (ppm)	0.75	2.25	4.50	9.00	13.50
COD (ppm)	5.00	15.00	30.00	60.00	90.00
Total suspended Solids (mg/L)	10.00	30.00	70.00	189.00	308.00
NH_3 (ppm)	0.05	0.20	0.60	1.80	3.00
NO_3 (ppm)	2.00	8.00	24.00	72.00	120.00
Cl (ppm)	25.00	100.00	225.00	460.00	695.00
Manganese (ppm)	0.03	0.11	0.34	0.75	1.16
Iron (ppm)	0.05	0.20	0.60	1.80	3.00

TABLE V. NON-DIMENSIONAL STANDARD VALUES FROM PRATI INDEX

Parameters	Notation	Uncontaminated (λ_1)	Acceptable (λ_2)	Moderately contaminated (λ_3)	Contaminated (λ_4)	Highly Contaminated (λ_5)
pH	C_1	0.822	0.930	0.986	1.083	1.179
BOD	C_2	0.125	0.375	0.750	1.500	2.250
COD	C_3	0.125	0.375	0.750	1.500	2.250
TSS	C_4	0.082	0.247	0.577	1.557	2.537
NH_3	C_5	0.044	0.177	0.531	1.593	2.655
NO_3	C_6	0.044	0.177	0.531	1.593	2.655
Cl	C_7	0.083	0.332	0.748	1.528	2.309
Mn	C_8	0.053	0.231	0.704	1.576	2.437
Fe	C_9	0.044	0.177	0.531	1.593	2.655

Similarly, based on the results of the participatory monitoring report of surface water quality in the Huallaga river watershed, developed by the National Water Authority (ANA), the values without dimension were obtained for each parameter of the 21 selected monitoring points. These values are presented in Table VI.

TABLE VI. NON-DIMENSION MONITORING DATA IN THE CASE STUDY

Point	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
1	0.985	0.167	0.150	0.239	0.060	0.045	0.007	0.112	0.454
2	1.002	0.167	0.100	0.387	0.020	0.056	0.010	0.166	0.531
3	0.999	0.167	0.100	0.568	0.088	0.058	0.009	0.302	0.672
4	0.999	0.167	0.175	0.428	0.034	0.044	0.007	0.219	0.729
5	0.989	0.167	0.225	0.659	0.028	0.047	0.008	0.239	1.010
6	0.994	0.167	0.500	12.545	0.003	0.032	0.004	2.292	22.673
7	0.998	0.167	0.050	10.988	0.005	0.024	0.010	1.295	21.407
8	0.993	0.833	0.350	0.428	0.081	0.019	0.009	2.225	33.487
9	1.001	0.167	0.025	9.292	0.032	0.024	0.010	1.397	19.823
10	0.955	0.833	0.250	12.315	0.090	0.019	0.009	2.006	32.195
11	0.992	0.167	0.050	4.481	0.040	0.026	0.009	0.992	14.513
12	0.989	0.500	0.250	13.213	0.100	0.020	0.008	2.160	0.333
13	0.982	0.167	0.025	12.842	0.100	0.020	0.007	2.134	0.326
14	0.971	0.167	0.375	4.992	0.055	0.033	0.010	0.780	9.735
15	0.940	0.167	0.800	0.593	0.005	0.012	0.004	0.131	2.270
16	0.981	0.167	0.475	4.893	0.005	0.018	0.008	1.054	14.646
17	0.863	0.167	0.300	0.832	0.081	0.018	0.014	0.218	3.965
18	0.935	0.167	0.025	1.425	0.005	0.015	0.008	0.282	4.985
19	0.918	0.167	0.025	1.928	0.091	0.017	0.016	0.327	4.554
20	1.001	0.167	0.225	0.470	0.075	0.055	0.008	0.233	0.448
21	0.935	0.167	0.300	0.387	0.072	0.053	0.008	0.302	0.668

3) Step 3: Replacing the values in Table III in (2)-(4), the triangular whitening functions of the five Gray classes were obtained for each parameter. As an example, the functions corresponding to the second parameter (BDO) are shown in (11)-(15) and Fig. 4. Then, the values in Table VI were evaluated in the triangular whitening functions of the five classes Grey for each parameter. The results obtained for the first five monitoring points are shown in Table VII.

$$f_2^1(x_{ij}) = \begin{cases} 1, x \in [0, 0.125] \\ \left(\frac{0.375-x}{0.375-0.125}\right), x \in <0.125, 0.375> \\ 0, x \in [0.375, +\infty> \end{cases} \quad (11)$$

$$f_2^2(x_{ij}) = \begin{cases} \left(\frac{x-0.125}{0.375-0.125}\right), x \in <0.125, 0.375] \\ \left(\frac{0.750-x}{0.750-0.375}\right), x \in <0.375, 0.750> \\ 0, x \in [0, 0.125] \cup [0.750, +\infty> \end{cases} \quad (12)$$

$$f_2^3(x_{ij}) = \begin{cases} \left(\frac{x-0.375}{0.750-0.375}\right), x \in <0.375, 0.750] \\ \left(\frac{1.500-x}{1.500-0.750}\right), x \in <0.750, 1.500> \\ 0, x \in [0, 0.375] \cup [1.500, +\infty> \end{cases} \quad (13)$$

$$f_2^4(x_{ij}) = \begin{cases} \left(\frac{x-0.750}{1.500-0.750}\right), x \in <0.750, 1.500] \\ \left(\frac{2.250-x}{2.250-1.500}\right), x \in <1.500, 2.250> \\ 0, x \in [0, 0.750] \cup [2.250, +\infty> \end{cases} \quad (14)$$

$$f_2^5(x_{ij}) = \begin{cases} \left(\frac{x-1.500}{2.250-1.500}\right), x \in <1.500, 2.250> \\ 1, x \in [2.250, +\infty> \\ 0, x \in [0, 1.500] \end{cases} \quad (15)$$

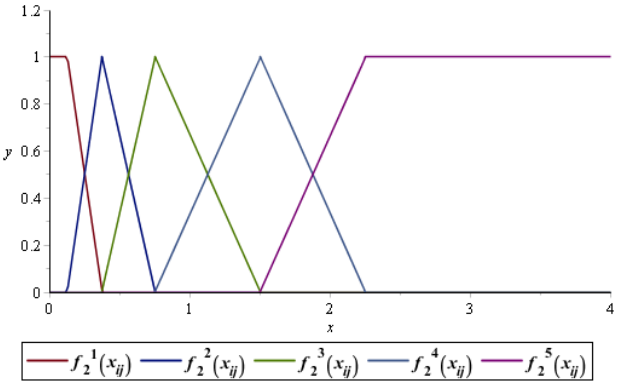


Fig. 4. CTWF for the Second Parameter.

TABLE VII. VALUES OF CTWF OF FIRST FIVE MONITORING POINT

P1	$\mathbf{f_j^1(X)}$	$\mathbf{f_j^2(X)}$	$\mathbf{f_j^3(X)}$	$\mathbf{f_j^4(X)}$	$\mathbf{f_j^5(X)}$
C_1	0	0.02	0.98	0	0
C_2	0.833	0.167	0	0	0
C_3	0.9	0.1	0	0	0
C_4	0.05	0.95	0	0	0
C_5	0.88	0.12	0	0	0
C_6	0.991	0.009	0	0	0
C_7	1	0	0	0	0
C_8	0.668	0.332	0	0	0
C_9	0	0.218	0.782	0	0
P2	$\mathbf{f_j^1(X)}$	$\mathbf{f_j^2(X)}$	$\mathbf{f_j^3(X)}$	$\mathbf{f_j^4(X)}$	$\mathbf{f_j^5(X)}$
C_1	0	0	0.835	0.165	0
C_2	0.833	0.167	0	0	0
C_3	1	0	0	0	0
C_4	0	0.575	0.425	0	0
C_5	1	0	0	0	0
C_6	0.911	0.09	0	0	0
C_7	1	0	0	0	0
C_8	0.367	0.633	0	0	0
C_9	0	0	1	0	0
P3	$\mathbf{f_j^1(X)}$	$\mathbf{f_j^2(X)}$	$\mathbf{f_j^3(X)}$	$\mathbf{f_j^4(X)}$	$\mathbf{f_j^5(X)}$
C_1	0	0	0.871	0.129	0
C_2	0.833	0.167	0	0	0
C_3	1	0	0	0	0
C_4	0	0.025	0.975	0	0
C_5	0.673	0.327	0	0	0
C_6	0.897	0.103	0	0	0
C_7	1	0	0	0	0
C_8	0	0.85	0.15	0	0
C_9	0	0	0.867	0.133	0
P4	$\mathbf{f_j^1(X)}$	$\mathbf{f_j^2(X)}$	$\mathbf{f_j^3(X)}$	$\mathbf{f_j^4(X)}$	$\mathbf{f_j^5(X)}$
C_1	0	0	0.871	0.129	0
C_2	0.833	0.167	0	0	0
C_3	0.8	0.2	0	0	0
C_4	0	0.45	0.55	0	0
C_5	1	0	0	0	0
C_6	1	0	0	0	0
C_7	1	0	0	0	0
C_8	0.068	0.932	0	0	0
C_9	0	0	0.813	0.187	0
P5	$\mathbf{f_j^1(X)}$	$\mathbf{f_j^2(X)}$	$\mathbf{f_j^3(X)}$	$\mathbf{f_j^4(X)}$	$\mathbf{f_j^5(X)}$
C_1	0	0	0.976	0.024	0
C_2	0.833	0.167	0	0	0
C_3	0.6	0.4	0	0	0
C_4	0	0	0.916	0.084	0
C_5	1	0	0	0	0
C_6	0.98	0.021	0	0	0
C_7	1	0	0	0	0
C_8	0	0.983	0.017	0	0
C_9	0	0	0.549	0.451	0

4) Step 4: The clustering weight (ηi) of each parameter was determined using Shannon entropy. For this, the following procedure:

a) Substep 4.1: The values of the parameters of the Prati index were normalized. These values are presented in Table VIII.

b) Substep 4.2: The entropy Hj of each criterion Cj was calculated through (7). The results are presented in Table IX.

c) Substep 4.3: The degree of divergence of each criterion Cj was calculated through (8). The results are shown in Table X.

d) Substep 4.4: Finally, the entropy weights wj according to (9) and were equated to the grouping weights nj of each parameter. The values are presented in Table XI.

5) Step 5: The values of the clustering coefficients(σ_i^k) were calculated using (10). The results of the the first five monitoring points are shown in Table XII.

6) Step 6: Finally, the condition was applied: if $max\{\sigma_i^k\} = \sigma_i^{k^*}$, it is decided that the object i belongs to the Grey class k*; for each monitoring point.

TABLE VIII. NORMALIZED VALUES OF EACH CRITERIA

Class / Parameter	λ_1	λ_2	λ_3	λ_4	λ_5	
C_1	0.164	0.186	0.197	0.217	0.236	1
C_2	0.025	0.075	0.15	0.3	0.45	1
C_3	0.025	0.075	0.15	0.3	0.45	1
C_4	0.016	0.049	0.115	0.311	0.507	1
C_5	0.009	0.035	0.106	0.319	0.531	1
C_6	0.009	0.035	0.106	0.319	0.531	1
C_7	0.017	0.066	0.15	0.306	0.462	1
C_8	0.011	0.046	0.141	0.315	0.487	1
C_9	0.009	0.035	0.106	0.319	0.531	1

TABLE IX. ENTROPY VALUES IN THE CASE STUDY

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
H_j	0.995	0.803	0.803	0.729	0.683	0.683	0.778	0.733	0.683

TABLE X. DEGREE OF DIVERGENCE

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
div_j	0.005	0.197	0.197	0.271	0.317	0.317	0.222	0.267	0.317

TABLE XI. CLUSTERING WEIGHT OF EACH PARAMETER

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
$w_j = n_j$	0.002	0.094	0.094	0.128	0.15	0.15	0.105	0.126	0.150

TABLE XII. VALUES OF CTWF AND σ_i^k FOR THE FIRST FIVE MONITORING POINTS

P1	$f_j^1(X)$	$f_j^2(X)$	$f_j^3(X)$	$f_j^4(X)$	$f_j^5(X)$
σ_i^k	0.639	0.241	0.12	0	0
P2	$f_j^1(X)$	$f_j^2(X)$	$f_j^3(X)$	$f_j^4(X)$	$f_j^5(X)$
σ_i^k	0.61	0.183	0.207	0	0
P3	$f_j^1(X)$	$f_j^2(X)$	$f_j^3(X)$	$f_j^4(X)$	$f_j^5(X)$
σ_i^k	0.513	0.191	0.276	0.02	0
P4	$f_j^1(X)$	$f_j^2(X)$	$f_j^3(X)$	$f_j^4(X)$	$f_j^5(X)$
σ_i^k	0.567	0.21	0.195	0.028	0
P5	$f_j^1(X)$	$f_j^2(X)$	$f_j^3(X)$	$f_j^4(X)$	$f_j^5(X)$
σ_i^k	0.537	0.18	0.204	0.079	0

TABLE XIII. VALUES OF σ_i^k FOR EACH MONITORING POINT

Point	$max\ \sigma_i^k$	Prati index
P1	0.639	Uncontaminated
P2	0.610	Uncontaminated
P3	0.513	Uncontaminated
P4	0.567	Uncontaminated
P5	0.537	Uncontaminated
P6	0.484	Uncontaminated
P7	0.577	Uncontaminated
P8	0.373	Uncontaminated
P9	0.577	Uncontaminated
P10	0.400	Uncontaminated
P11	0.577	Uncontaminated
P12	0.389	Uncontaminated
P13	0.514	Uncontaminated
P14	0.472	Uncontaminated
P15	0.554	Uncontaminated
P16	0.484	Uncontaminated
P17	0.482	Uncontaminated

IV. RESULTS AND DISCUSSION

A. Results on the Case Study

It is showed, in Table XIII, that the 21 monitoring points resulted in an uncontaminated water quality, however, a quality level comparison can be made according to the maximum clustering coefficient (max. σ_i^k), like to shown in Table XIV and Fig. 5.

It is observed that the monitoring point P1 presents the best water quality and, the point P8, the lowest water quality. This happens because point P1 is on the beginning of the Huallaga river and point P8 is in a lower zone, it means that the quality of Water decreases along the river depending on the activities that take place, such as mining and hydroelectric plants [1].

TABLE XIV. COMPARISON OF THE WATER QUALITY OF THE MONITORING POINTS ACCORDING TO THEIR MAX σ_i^k

Point	Max σ_i^k	Prati index	Color scale
1	0.639	Uncontaminated	Better water quality
2	0.610	Uncontaminated	
7	0.577	Uncontaminated	
9	0.577	Uncontaminated	
11	0.577	Uncontaminated	
18	0.577	Uncontaminated	
4	0.567	Uncontaminated	
15	0.554	Uncontaminated	
5	0.537	Uncontaminated	
19	0.524	Uncontaminated	
13	0.514	Uncontaminated	
3	0.513	Uncontaminated	
20	0.493	Uncontaminated	
6	0.484	Uncontaminated	
16	0.484	Uncontaminated	
17	0.482	Uncontaminated	
14	0.472	Uncontaminated	
21	0.470	Uncontaminated	
10	0.400	Uncontaminated	
12	0.389	Uncontaminated	
8	0.373	Uncontaminated	Lower water quality

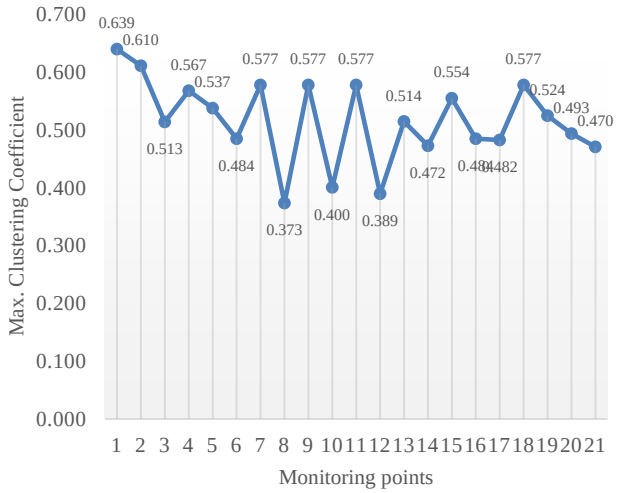


Fig. 5. Max. σ_i^k of Monitoring Points.

The reason why the water is uncontaminated may be because mining and hydroelectric companies, which operate in the upper part of the Huallaga river, they treat their industrial effluents adequately in accordance with the regulations national. This water body is classified as suitable for the irrigation of vegetables and animals [1], but according to the results of the water quality assessment, it can also be considered as water that can be made drinkable with conventional treatment.

In relation to other studies, Fu and Zou [10], applied the Grey Clustering method to evaluate the water quality of the Yellow River, the results also showed good river quality. In the assessment of the water quality carried out by Liping et al. [11], applied the Grey Clustering method for the assessment of the quality of the Fenchuan river of the Yan'an Baota area in China, however, considered the arithmetic mean for the determination of the weights of clustering. In this case, Shannon Entropy could have been applied alternatively, as was done in the present study, to calculate these weights in an objective and precise way. Similarly, in the study carried out by Wang et al. [12] the clustering weights could be obtained through the Shannon Entropy method and be complemented by the Single Factor method used in this study.

B. Discussion on the Methodology

The Grey Clustering method is the most appropriate in high uncertainty issues [3] such as assessment of surface water quality where each parameter varies depending on environmental conditions, in comparison of classic multi-criteria assessment methods such as Delphi [13] or the Analytical Hierarchy Process (AHP) [14] which do not consider uncertainty in their analysis. In Peru, the Grey Clustering method is not very widespread compared to other logic methods Aristotelian or statistical models [7] which means a limitation for its application with the national water quality standards, which are not determined based on any quality index.

Finally, the Shannon entropy method is well suited for evaluating water quality because allowed to determine the grouping weights (η_j) for each parameter in an objective way, without the need to ask to an expert and this reduces assessment costs. In addition, this method has multiple applications as in studies of social conflicts or assessments of social impact [15], due to its great capacity to process information and reduce subjectivity in assessments.

V. CONCLUSIONS

The surface water quality of the upper watershed of Huallaga river could be evaluated using Grey Clustering method and Shannon Entropy, so it was possible to classify the 21 monitoring points in this area. The results obtained in this study can be useful to the regional and local authorities of Peru, as well as to the National Water Authority to make better decisions regarding the management of this important watershed.

According to the methodology, the Grey Clustering method can be more effective than other classical methods due to that considers the uncertainty within the analysis, regarding the Shannon Entropy it allows to calculate the weights objectively of the criteria without resorting to expert judgments. Another important point is that when using the Prati index, there is an advantage when we need to compare if the water quality is affected by the activities carried out in the watershed, due to the parameters used.

Finally, in future research, the efficacy of the Grey Clustering method should be compared with the methodology established by the National Water Authority (ANA) for the determination of the Quality Index Environment of Surface Water Resources (ICARHS). In case the results are indistinct or very similar, the use of the Grey Clustering method could be extended to those rivers where the data is insufficient to apply the ANA methodology.

REFERENCES

[1] Autoridad Nacional del Agua, "Informe del Monitoreo Participativo de la Calidad del Agua Superficial en la cuenca del río Huallaga – Febrero y Marzo del 2019," Tarapoto, 2019.

[2] Congreso de la República del Perú, "Pre Dictamen del Texto Sustitutorio de la Ley que declara de interés nacional y necesidad publica la recuperación, conservación y protección de las aguas de la cuenca del rio Huallaga," Lima, 2017. [Online]. Available: http://www.congreso.gob.pe/Docs/comisiones2016/PueblosAndinosEcologia/files/ppt_predictamen_huallaga_al_10.04.17.pdf.

[3] S. Liu and Y. Lin, Grey Systems: Theory and Applications, vol. 68. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[4] C. Zhu and Q. Liu, "Evaluation of Water Quality Using Grey Clustering," in 2009 Second International Workshop on Knowledge Discovery and Data Mining, Jan. 2009, pp. 803–805, doi: 10.1109/WKDD.2009.28.

[5] Y. Leng et al., "Evaluation on Transfer Efficiency at Integrated Transport Terminals through Multilevel Grey Evaluation," Procedia - Soc. Behav. Sci., vol. 43, pp. 587–594, 2012, doi: 10.1016/j.sbspro.2012.04.132.

[6] C. E. Shannon and W. Weaver, The Mathematical Theory of Communication. 1949.

[7] A. Delgado and I. Romero, "Environmental conflict analysis using an integrated grey clustering and entropy-weight method: A case study of a mining project in Peru," Environ. Model. Softw., vol. 77, pp. 108–121, Mar. 2016, doi: 10.1016/j.envsoft.2015.12.011.

[8] Y. Zhang, J. Ni, J. Liu, and L. Jian, "Grey evaluation empirical study based on center-point triangular whitenization weight function of Jiangsu Province industrial technology innovation strategy alliance," Grey Syst. Theory Appl., vol. 4, no. 1, pp. 124–136, Jan. 2014, doi: 10.1108/GS-11-2013-0027.

[9] A. Shemshadi, H. Shirazi, M. Toreihi, and M. J. Tarokh, "A fuzzy VIKOR method for supplier selection based on entropy measure for objective weighting," Expert Syst. Appl., vol. 38, no. 10, pp. 12160–12167, Sep. 2011, doi: 10.1016/j.eswa.2011.03.027.

[10] X. Q. Fu and Z. H. Zou, "Water Quality Evaluation of the Yellow River Basin Based on Gray Clustering Method," IOP Conf. Ser. Earth Environ. Sci., vol. 128, p. 012139, Mar. 2018, doi: 10.1088/1755-1315/128/1/012139.

[11] Wang Liping, Lai Kunrong, and Zhou Weibo, "Application of Grey Clustering method for water quality evaluation in fenchuan River Yan'an Baota Area," in 2011 International Symposium on Water Resource and Environmental Protection, May 2011, pp. 838–841, doi: 10.1109/ISWREP.2011.5893142.

[12] J. Wang et al., "Application of Grey Clustering Method Based on Improved Analytic Hierarchy Process in Water Quality Evaluation," MATEC Web Conf., vol. 246, p. 02004, Dec. 2018, doi: 10.1051/matecconf/201824602004.

[13] V. Mahajan, "The Delphi method: Techniques and applications," JMR, J. Mark. Res., vol. 13, no. 000003, p. 317, 1976.

[14] T. L. Saaty, The Analytic Hierarchy Process. New York, 1980.

[15] A. Delgado, "Social conflict analysis on a mining project using shannon entropy," in 2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON), Aug. 2017, pp. 1–4, doi: 10.1109/INTERCON.2017.8079661.

Supplier Qualification Model (SQM): A Quantitative Model for Supplier Agreements Evaluation

Mohammed Omar[1], Yehia Helmy[2], Ahmed Bahaa Farid[3]
Information Systems Department, Faculty of Computers and Information, Helwan University[1]
Microsoft Company, Technology Solution Professional (TSP), 32511 Helwan, Egypt[2]
Management Information System Department, Faculty of Commerce and Business Administration[2]
Helwan University, 32511 Helwan, Egypt[2]
Information Systems Department, Faculty of Computers and Artificial Intelligence, Helwan University, 32511 Helwan, Egypt[3]
Information Systems Department, Faculty of Computers and Artificial Intelligence[3]
Beni-Suef University, 62511 Beni-Suef, Egypt[3]

***Abstract*—Recently software outsourcing has increasingly widespread due to the valuable economical and technical benefits it introduced to the software development industry. Where the software development organizations adopt a third party to acquire a software project component (product, service). In the acquisition, process companies rely on the CMMI supplier agreement management (SAM) process area to select the potential supplier. Potential suppliers (vendors) are carefully selected through a dedicated process to ensure the delivery of high-quality and reliable services. Most of the published work in the context of how to evaluate and select the right supplier is based on a normal process with plain steps, nevertheless, no literature was reported to evaluate suppliers in a measurable way and select the potentials depending on a quantitative model. The purpose of this paper is to propose a practical quantitative model called the Supplier Qualification Model that enables the organizations to easily evaluate and select the potential suppliers through a measurable approach depends on monitoring and executing the SLAs of the SAM. The proposed model has been verified by implementing it through building an extension for one of the worldwide leading Agile management platforms according to Gartner (Microsoft Team Foundation Server). Multiple versions of the extension were implemented to target the major versions of Microsoft Team Foundation Server and validated by using them in 426 worldwide companies. This proves the suitability of the model to be used.**

Keywords*—Agile practices; vendor selection; CMMI; outsourcing; software acquisition; supplier agreement management; supplier selection; supplier evaluation

I. Introduction

In the last decade, the software development markets became more competitive and demanding, where it came to be crucial for organizations to invest more effort to improve their software processes to meet specific requirements. As a result, they had to follow a quality model to improve their software development process, increase their capability and maturity level, and become a benchmark for their competitors [1].

The Capability Maturity Model Integration (CMMI) Product Suite that was resealed in January 2002 paved the way for the organizations to improve their processes for acquisition, development, and sustainability [2]. Therefore, a large number of software development organizations (according to the CMMI Institute, 5000 businesses in 49 countries all over the world, and 1900 appraisal were conducted [3]) embark on it to improve their performance, develop higher-quality software, meet stakeholders satisfaction and achieve external validation [4].

As the software market mandate not only a product of high quality but also time and cost control [1], the organizations found that working Agile based on the CMMI model would allow continuous improvement and help them to reach the required maturity level [5].

Agile is a traditional software development approach that was created in 2001[6], and its methodologies have gained widespread acceptance among a large number of organizations as a quality-focused and highly collaborative mechanism to manage software development and improve the delivery process [7]. Agility is based on the idea that high-quality software can be developed by following a set of rules that allows continuous product improvement and testing depending on rapid feedback and testing [8].

Williams and Cockburn [9] defined the Agile team at the beginning to be as small as 50 people or fewer. However, Duka [10] showed that the larger teams had adopted Agile methodologies over the waterfall methods due to the great benefits of Agile practices. As well, in 2010 studies in the Agile journal [10] showed that 88% of companies with some of them have more than 10.000 employees are adopting Agile.

Different studies [11–13] were conducted to display the great benefits of working agile from the perspective of productivity and time management which leads at the end to customer satisfaction. They showed that agility can increase productivity to about 88%, improve cost efficiency by about 26%, and 41% time to market. Also, the 12th annual state of the Agile report [14] revealed that working Agile, 71 % manage change priorities, 66% improve the project visibility, gives 65% better business and IT alignment, speed the project delivery 62%, and 61% increase the team productivity. Besides, the Standish Group Chaos Report 2018 [15] results exhibited that working agile gives higher success and lower failure rates when compared to projects adopting the waterfall method (see Fig. 1 and Fig. 2).

Another reason that many organizations adopt Agile practices is that it permits the so-called distributed development (software outsourcing) [16], where an organization selects a third-party service provider (supplier) to execute a part of a software development project. The strategy of software outsourcing brought enormous benefits over the in-house development, where it reduced the operating costs, saved time, and gave immediate access to talented and higher-level IT professionals [17,18]. Watts Humphrey [19] implied that the quality of the software development process determines the quality of the produced software, and in the case of outsourced products the quality of software acquired from suppliers. This means that the selection process of the suppliers should be done according to specific and carefully calculated criteria.

Since the process to manage and control outsourced development projects is complicated and faces obstacles as in-house projects. Agile organizations had to adopt a specific process to manage the relationship between them and the service provider to be able to identify and select the potential suppliers. For an Agile organization that needs to be compliant with the CMMI supplier agreement management (SAM) process area, it must demonstrate an explicit commitment to establish a process-based confirmation through people and should verify the execution of the process and validate the outcome of process execution through a measured result [2]. EM Soares et al. [20] mapped the relationship between the supplier agreement management process area and agile practices, where they developed a catalog of the best practices using the concepts of the agile methodologies to manage the software acquisition process in the Supplier Agreement Management (SAM) process area of CMMI-DEV. Neither EM Soares nor other studies in the literature to the best of our knowledge was reported to introduce a quantitative method that helps the organizations to differentiate between software development suppliers, to select the potential ones based on measured data, and to monitor the execution of service level agreements established between the stakeholders and service providers according to CMMI-DEV v.2 specific goals of SAM process area.

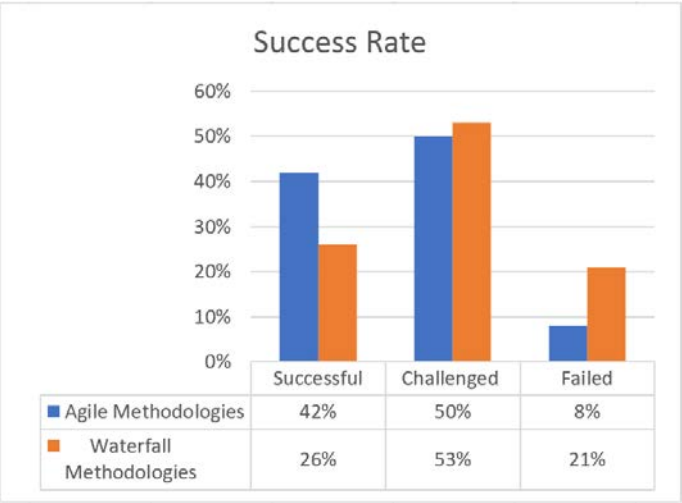


Fig. 1. Projects Success Rate.

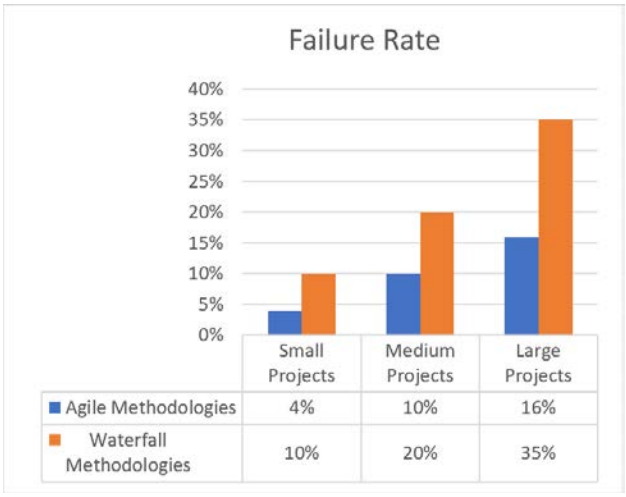


Fig. 2. Projects Failure Rate.

The main goal of this work is to propose a reference model that can be used to evaluate the supplier agreements quantitatively. And help the organizations to identify and select the potential suppliers. In the coming sections of this paper, a background about the supplier agreement management (SAM) process area and the selection process of potential suppliers is presented in a detailed manner. Then an introduction to the proposed model, its mathematical profile, and how it is in the evaluation process of suppliers is introduced. As well as, clarification of why the SQM model will be helpful for organizations to select the potential suppliers is demonstrated. Afterward, the verification and validation of the SQM model were discussed. Finally, a vision for further work in the future is proposed.

II. Background and Related Work

Supplier agreement management (SAM) is a CMMI process area at maturity level 2, it falls in the project management category and aims to manage the acquisition process of products and services from suppliers [21]. SAM is applicable only when organizations deal with a third party (supplier) to acquire products or services for software development. The specific goals of this process area were formulated to determine the acquisition type, to select suppliers, to establish and maintain supplier agreements, to execute the service level agreements (SLAs), to monitor the processes of the selected supplier, to accept the acquired product, and to ensure the successful delivery of products.

The supplier selection process comes directly after the identification of the acquisition type required from the vendor/supplier. Suppliers are selected based on an evaluation of their ability to meet specific predefined requirements and founded criteria. In the evaluation criteria, it is crucial to define the critical factors for the project such as costs, the geographical location of the supplier, quality services, supplier's performance record, prior experience, etc. to be able to identify and select the potential suppliers, then to evaluate their proposals [20,22,23]. Afterward, the risks associated with each proposed supplier and his ability to perform work must be assessed through the evaluation of their prior experience and prior performance [22].

Selecting the right supplier who can deliver a high-quality and reliable product at the desired time and within the defined budget is not an easy task, and unfortunately, not much literature was dedicated to define a measurable method with quantitative terms to select potential suppliers. Nevertheless, authors always report the evaluation and selection process of suppliers in a plain way as mentioned above, for example,

G.O'regan [24] reported that in the identification process of suppliers, organizations may search or depend on recommendations from colleagues and previous work relationships to select the candidate suppliers. Then an evaluation team will be responsible to rate each one of them according to the founded criteria. Afterward, a shortlist of potential suppliers will present their proposals followed by a Q & A session, where the final decision will be made depending on this discussion.

Chaudhary et al. [25] stated that candidate suppliers can be chosen based on their capability or experience in delivering the desired product or services. After selecting the potential candidates, the final selection will be performed depending on their proposals, and the advantages and the disadvantages of each in the light of the factors of the predefined criteria.

EM Soares et al. [20] developed a catalog of the best practices to manage the software acquisition process by employing agile methodologies. They build up their work depending on the framework activities of Furtado [26], where they showed that these activities are compatible with the specific goals of the SAM process area in CMMI-DEV. They used this catalog to define finely the steps to evaluate and select suppliers based on the agile practices using the formal alternate evaluation method. Where they made a survey and collected information about suppliers to have a list, then they analyzed these data to limit this list to the most suitable candidates. Afterward, the potentials were selected depending on a checklist of the project requirements and the features of the supplier. Finally, the supplier who meets all the requirements or is closest to them will be selected.

Although these references give detailed and precise ways to evaluate and select suppliers, to the best of our knowledge, no study was reported in the literature to propose a method or model that facilitates the differentiation and evaluation of suppliers and the risk associated with each one of them in a measured way.

Since the identification of potential suppliers and the risk associated with each one of them is considered a crucial role in the success of a software acquisition process, the present study proposes a model that facilitates the evaluation and estimates the risk associated with each proposed supplier.

The developed model calculates the supplier's qualification through a quantitative score (SQScore) based on the established service level agreements (SLAs), enables organizations to use this quantitative score (SQScore) to distinguish between suppliers depending on their historical accumulative SQScores.

III. Introducing Supplier Qualification Model (SQM)

A. The Need for a Qualification Model for Selecting the Right Supplier

According to the CMMI SAM process area reference guide, many factors control the selection process of suppliers such as budget (money), quality, project business value, supplier reliability, compliance, the risk of failure, supplier experience, and previous work [21].

In the light of the above factors, when an organization must select a supplier from a list of professional suppliers, before the final selection, a set of questions should be answered to define the potential (right) supplier:

- Which supplier could deliver better business value when delivering the project?
- Does the supplier have experience working with the organization?
- What is the rate of associated failure of each supplier?
- What is the percentage of failure of each supplier?
- What is the risk of assigning the project to each supplier?

The supplier qualification model (SQM) will help the organizations to answer these questions, not only to select the supplier but also to monitor the selected supplier during the project execution for a better future qualification process.

B. Idea behind the SQM Model

According to SAM specific goal (SAM.SG.1), Establish Supplier Agreements [27], organizations established some service level agreements (SLAs). Each particular SLA mentioned in the supplier agreement has terms and conditions that suppliers should fulfill to satisfy this SLA. For example, an organization can establish a service level agreement to determine the level of quality for a specific number of works (Wn), the maximum amount of budget for a set of specified features, and/or the level of skilled resources that can handle certain types of work.

The proposed SQM is considering these factors depending on a set of SLAs execution measures. When the supplier fails to satisfy the terms and conditions of an agreed SLA, this is termed as a violation, hence the total number of violations is given as (Vn), while the percentile of deviation from the original goal of the SLA refers to the violation percentage (Vp). Since the degree of importance of an established SLA varies depending on the nature of the SLA and the outstanding project, in the present model, we were concerned to measure two SLA weights, (i) the compliance weight (ComplianceWt) which identifies how much compliant is a supplier with all the founded SLAs, and (ii) the risk weight (RiskWt) to estimate the associated risk of the supplier.

Organizations can control the weights of the SLAs, where in some cases they can decide to give a low score for any violation of a certain SLA regardless of the percentage of the violation, while for other SLAs the percentage of failure is

more important. Since the two weights are mutually exclusive, the summation of both weights per SLA should be 100 points of weight.

C. Mathematical Profile of the SQM Model

The values of the ComplianceWt and the RiskWt are defined according to specific rules related to the project. This could differ from one SLA to another. As an example, for some SLAs, if the cost of failure to comply with a given SLA is very high, the compliance weight (CompilanceWt) should have a high value. Nevertheless, for other SLAs, the cost of failure will not be as high as long as the percentage of failure is low, Hence the SLA should have a high-risk weight (RiskWt) and low-Compliance weight (ComplianceWt) as will be explained in the following example.

If we assume that an organization is assigning a project to Supplier 1 and that 8 SLAs were established with each SLA has a number of work items (Wn). Each SLA has designated specific RiskWt points and ComplianceWt points (please note that the summation of each pair is 100). Each SLA has been violated a number of times (Vn), each time the violation is incurred with a specific percentage of violation (Vp). These values are depicted in the following set of matrices.

$$RiskWt \coloneqq \begin{bmatrix}20\\40\\15\\80\\50\\90\\30\\23\end{bmatrix} \quad ComplianceWt \coloneqq \begin{bmatrix}80\\60\\85\\20\\50\\10\\70\\77\end{bmatrix} \quad Vn \coloneqq \begin{bmatrix}6\\2\\1\\2\\2\\0\\4\\1\end{bmatrix} \quad Wn \coloneqq \begin{bmatrix}6\\7\\1\\4\\6\\5\\3\\2\end{bmatrix} \quad Vp \coloneqq \begin{bmatrix}10\ 40\ 20\ 5\ 12\ 7\\20\ 20\ 0\ 0\ 0\ 0\\30\ 0\ 0\ 0\ 0\ 0\\10\ 20\ 0\ 0\ 0\ 0\\3\ 10\ 0\ 0\ 0\ 0\\0\ 0\ 0\ 0\ 0\ 0\\3\ 10\ 20\ 30\ 0\ 0\\10\ 0\ 0\ 0\ 0\ 0\end{bmatrix}$$

Hence, if the number of SLAs (n) is given as,

$$n \coloneqq \text{rows}(Wt) = 8 \tag{1}$$

moreover, the index (i) is,

$$i \coloneqq 0..n-1 \tag{2}$$

The SLA Risk Weight Index (SlaRiskWI) and the SLA Compliance Weight Index (SLACWI) that measure the importance of each SLA relative to the whole project can be calculated using equations (3) and (4). For both weight indices, the total value of each is 100.

$$\boldsymbol{SlaCWI}_i \coloneqq \boldsymbol{ComplianceWt}_i \Big/ \sum_{i=0}^{n-1} \boldsymbol{ComplianceWt}_i * 100 = \begin{bmatrix}17.699\\13.274\\18.805\\4.425\\11.062\\2.212\\15.487\\17.035\end{bmatrix} \tag{3}$$

$$\boldsymbol{SlaRiskWI}_i \coloneqq \boldsymbol{RiskWt}_i \Big/ \sum_{i=0}^{n-1} \boldsymbol{RiskWt}_i * 100 = \begin{bmatrix}5.747\\11.494\\4.31\\22.989\\14.368\\25.862\\8.621\\6.609\end{bmatrix} \tag{4}$$

Using the above equations (3 & 4), for (I = 3) the SlaRiskWI_3, which is the SLA risk percentage for the fourth SLA out of all risk weights for the whole project) will be 22.989%, and the SlaCWI_3, which is the SLA compliance percentage of the fourth SLA out of all compliance weights for the whole project is 18.805%. Then using the violation percentage (Vp), the Violation Severity (VS) can be calculated as follows:

$$\boldsymbol{VS} = \sum_{r=0}^{cols(\boldsymbol{Vp})-1} \frac{\boldsymbol{Vp}^{(r)}}{100} = \begin{bmatrix}0.68\\0.4\\0.3\\0.3\\0.13\\0\\0.63\\0.1\end{bmatrix} \tag{5}$$

Dividing the violation severity (VS) by the number of work gives the average violation percentage (Average Compliance) for all work assigned to the supplier (see equation (6)).

$$\boldsymbol{AverageCompliance}_i \coloneqq \frac{\boldsymbol{VS}_i}{\substack{\boldsymbol{if}(\boldsymbol{Wn}_i>0)\\ ||\boldsymbol{Wn}_i\\ \boldsymbol{else}\\ ||\boldsymbol{1}}} = \begin{bmatrix}0.113\\0.057\\0.3\\0.075\\0.022\\0\\0.21\\0.05\end{bmatrix} \tag{6}$$

1) Evaluating Risk vs. Compliance: While using Average Compliance is helpful in many cases, using this measure only for qualifying suppliers, could be misleading. That is why we need in some cases to have an additional measure that considers risk associated with each supplier. For instance, If we assumed a case study of two suppliers, supplier A and supplier B, have the same violation severity, and supplier B has fewer number of violations than supplier A., In this case, the number of violations of supplier B (who has fewer violations) will indicate how much risk is associated with this supplier. For more clarification let us assume that both suppliers are assigning five projects and their violation percentages were as follows:

Vp Supplier A = [30,30,30,30,30] &

Vp Supplier B = [0,0,0,75,75].

Using equations (5) and (6), the violation severity of both suppliers is 150, and their average compliance is 30 respectively. This means that both suppliers have the same failure percentage for the whole assigned work.

For supplier A, since the violation percentage of each project is small, the risk is therefore low. On the other hand, for supplier B, nevertheless, he assigned three projects successfully, the associated risk would be high because of the high violation percentage of the last two elements of violated projects array (75,75).

Based on the numbers of the original example of (5) & (6), the average violation percentage of violated work (which is the AverageRisk) will be calculated by dividing the violation severity (VS) by the number of violations (see equation 7).

$$\boldsymbol{AverageRisk_i} \coloneqq \frac{\boldsymbol{VS_i}}{\substack{\boldsymbol{if(Vn_i>0)} \\ \boldsymbol{||Vn_i} \\ \boldsymbol{else} \\ \boldsymbol{||1}}} = \begin{bmatrix} 0.113 \\ 0.2 \\ 0.3 \\ 0.15 \\ 0.065 \\ 0 \\ 0.158 \\ 0.1 \end{bmatrix} \quad (7)$$

Equations 6 and 7 calculate the average violation percentage for the whole assigned and violated work. Nevertheless, the success factor of not making compliance failures, and not incurring risk factors should be rewarded for each supplier. Therefore, the success value must be considered by calculating the average compliance success, and the average risk success as follows:

$$AverageComplianceSuccess_i \coloneqq 1 - AverageCompliance_i = \begin{bmatrix} 0.887 \\ 0.943 \\ 0.7 \\ 0.925 \\ 0.978 \\ 1 \\ 0.79 \\ 0.95 \end{bmatrix} \quad (8)$$

$$AverageRiskSuccess_i \coloneqq 1 - AverageRisk_i = \begin{bmatrix} 0.887 \\ 0.8 \\ 0.7 \\ 0.85 \\ 0.935 \\ 1 \\ 0.843 \\ 0.9 \end{bmatrix} \quad (9)$$

2) Calculating SQScore for each SLA: The Compliance Score for each SLA (SlaComSc) is calculated by subtracting the product of the $AverageCompliance_i$ and $SlaCWI_i$ of the violations, from the product of $AverageComplianceSuccess_i$ and $SlaCWI_i$ (see equation (10)). Based on this equation, when the supplier achieves successes more than failures, the $SlaComSc_i$ will be a matrix of positive numbers otherwise, it will be negative numbers.

$$SlaComSc_i \coloneqq (AverageComplianceSuccess_i * SlaCWI_i) - (SlaCWI_i * AverageCompliance_i) = \begin{bmatrix} 13.687 \\ 11.757 \\ 7.522 \\ 3.761 \\ 10.583 \\ 2.212 \\ 8.982 \\ 15.332 \end{bmatrix} \quad (10)$$

Similarly, the Risk Score for each SLA (SlaComSc) is calculated using equation 11.

$$SlaRiskSc_i \coloneqq (SlaRiskWI_i * AverageRiskSuccess_i) - (SlaRiskWI_I * AverageRisk_i) = \begin{bmatrix} 4 \\ 6 \\ 1 \\ 1 \\ 2 \\ 5 \\ 5 \end{bmatrix} \quad (11)$$

3) Calculating SQScore: Therefore, the Supplier Compliance Score for all SLAs and the Supplier Risk Score for all SLAs, per each supplier, will be calculated from the sum of the SlaComSc and the sum of the SlaRiskSc respectively as seen in equations 12 and 13.

$$\sum_{i=0}^{n-1} SlaRiskSc_i = 78.712 \quad (12)$$

$$\sum_{i=0}^{n-1} SlaComSc_i = 73.837 \quad (13)$$

Finally, the SQScore of the supplier will be the average of both SlaRiskSc and SlaComSc.

$$QScore \coloneqq \left(\sum_{i=0}^{n-1} SlaRiskSc_i + \sum_{i=0}^{n-1} SlaComSc_i\right) \Big/ 2 = 76.27 \quad (14)$$

Equation 14 shows that the supplier scored a success rate of 76.274. Now we can say that using this model organizations will be able to differentiate easily between suppliers based on their historical projects.

4) Evaluating suppliers: If we assume that an organization has the following historical SQScore for six different suppliers:

Supplier 1	***Supplier 2***	***Supplier 3***	***Supplier 4***	***Supplier 5***	***Supplier 6***
90	76	50	33	55	80

The organization needs to categorize these suppliers into three ranges a high, medium, and low depending on their SQScores. Hence we have to calculate the standard deviation. To do so, firstly the mean is calculated by dividing the sum of suppliers score by the number of suppliers:

If n = index of the supplier:

$$meanSQScore\ (Average) \coloneqq \frac{\sum_{i=0}^{n-1} x_i}{n} = 64 \quad (15)$$

Then calculate the difference from the mean for each supplier,

Supplier 1	***Supplier 2***	***Supplier 3***	***Supplier 4***	***Supplier 5***	***Supplier 6***
26	12	-14	-31	-9	16

Afterward, we get the variance from the sum of the square of each difference, then dividing by the number of suppliers.

$$Variance \coloneqq \frac{\sum_{i=0}^{n-1} difference_i^2}{n-1} = 462.8 \quad (16)$$

Now the Standard Deviation (**σ**) is the square root of the variance.

$$\sigma \coloneqq \sqrt{Variance} = \mp 21.513 \quad (17)$$

Adding equations 15, 16, and 17 we get:

$$\boldsymbol{\sigma} \coloneqq \sqrt{\frac{1}{n-1}\sum_{i=0}^{n-1}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^2} \quad (18)$$

Using the standard deviation and the mean, suppliers 2, 3, 5, and 6 can be categorized as average suppliers (the period of values between the $meanSQScore \mp \boldsymbol{\sigma}$), while supplier 1 is above average, and supplier 4 is below average as illustrated in Fig. 3.

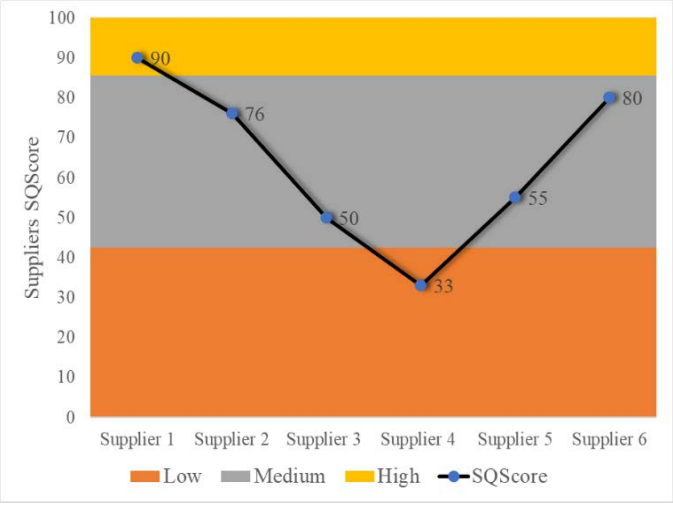


Fig. 3. The High, Medium, and Low Ranges of Suppliers SQscores.

IV. SQM Verification and Validation

A. Model Verification

To verify the applicability of the proposed SQ Model, an extension tool has been built to automate applying the proposed SQM for software teams. In this study, the extension is based on the Microsoft Team Foundation Server (TFS), which is one of the worldwide leading software engineering platforms according to Gartner [28].

The verification process was executed in two steps, (i) check the possibility of extending an agile tool (in this study is the TFS) to calculate the total SQscore of a supplier and of each established SLA using the proposed model, (ii) verify the possibility of extending the TFS client (Visual Studio) that allows organizations to create SLAs based on the SQ model, and monitor and track the SLA execution and the supplier SQ score.

The TFS Scrum process template provides organizations with the artifacts, processes, and workflows which are necessary to adopt the scrum method. The proposed TFS SLA Server Extension adds the following artifacts to the TFS Scrum process (i) SLA Configuration, (ii) SLA Violation Work-Items. The SLA Configuration will store the Service Level Agreements information such as SLA Risk Weight, SLA compliance weight, SLA threshold, and SLA percentage (see Fig. 4). Nevertheless, the SLA violation will store the violation information such as the violation percentage. Organizations can establish and maintain Service Level Agreements by creating as many SLA configuration items as needed to define flexible rules such as work deadline information and relation with other work. The SLA configurations are implemented dynamically to allow customizations that fit many different business scenarios. For example:

- Escalating work that in-progress for more than n number of hours/days.
- Escalating work with deadline configured on fields.
- Escalating all work related to a specific feature, epic, or product backlog items.
- Escalating work that takes more than n% of parents.
- Escalating assignment of specific work to a specific resource with a specific persona.

Once the organization created and activated SLAs for a particular project, the SLA is executed, and the SQscore is calculated by the TFS extension. Hence, the organization can monitor and track the execution and satisfaction of the SLAs.

The TFS SLA Client extension introduces a new capability to Visual Studio which allows organizations to monitor and track SLAs and measures the SQScore of a specific Supplier.

If Project X (see Fig. 5) with 8 SLAs has been established and executed by the supplier; each SLA has been violated several times, the extension calculates the Supplier Qualification Model Score by calculating the SQScore for each SLA using the model described above.

For example, SLA no. 320 has been violated 6 times. Therefore the Risk score is 4.44, the Compliance score is 13.69, and the SQScore is 9.1, and the supplier SQScore for the project is 76, as shown in (Fig. 5).

B. Model Validation

The TFS SLA Server and TFS SLA Client extensions have been published to Microsoft Visual Studio Gallery [29] and have scored 426 (for both versions) usages for the TFS SLA Server and 614 (for both versions) usages for the TFS SLA Client (check Ref. [29]) as illustrated in Fig. 6.

The SLA has been used by many organizations in the public and government sectors in multiple countries, which proofs the validity of the proposed SQM Model.

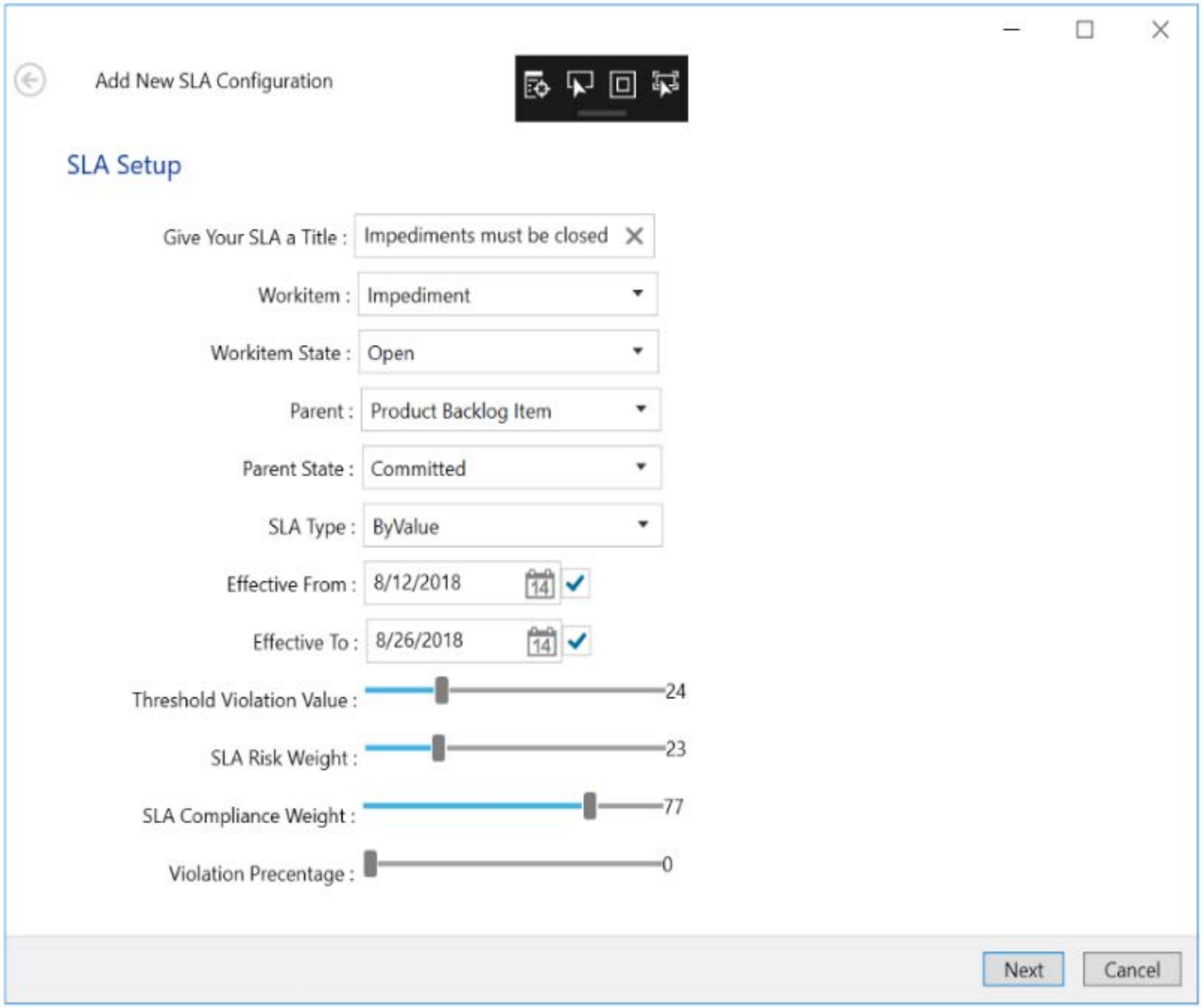


Fig. 4. SLA Configuration.

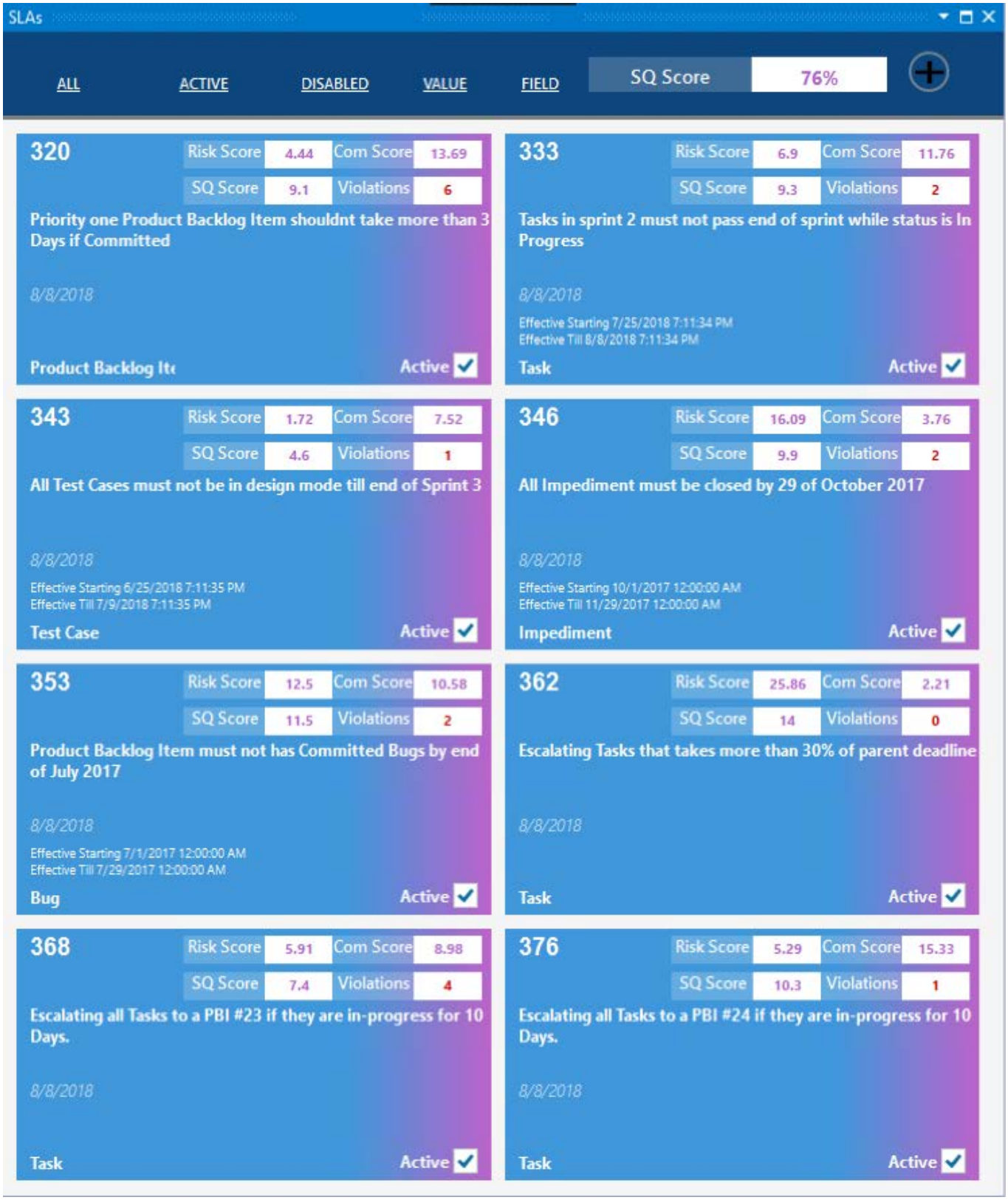


Fig. 5. TFS SLA Client Extension.

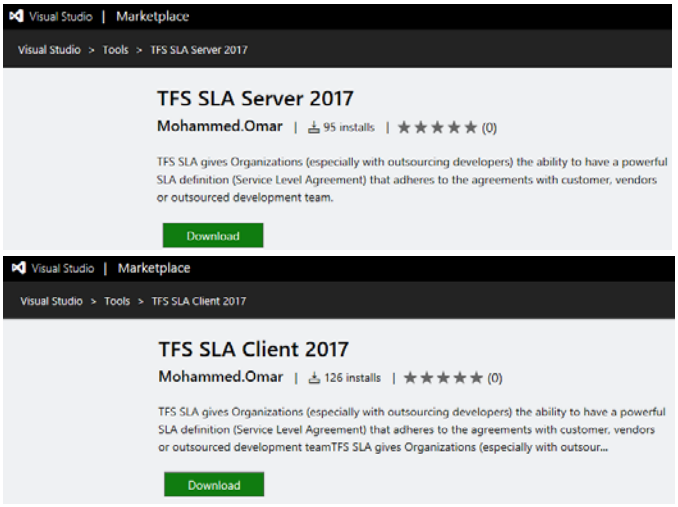


Fig. 6. TFS SLA usages [29].

V. Future Work

The proposed model is an efficient suit for an organization that wants to qualify their vendors and categorize them into categories. It works effectively for the organizations that have vendors with previous data, nevertheless, in the case of new vendors, organizations will not have previous data, every organization will have its own SLAs. To help overcome this problem, future research can define and unify a set of common SLAs for vendors who want to get appraised and qualified by the proposed SQM. A global system that uses predefined SLAs should be created to give appraisal and SQScore for each vendor, therefore organizations could use this global SQScore to validate vendors.

VI. Conclusion

Evaluating and selecting service suppliers is a hard task especially that up to now, no method or model was reported to evaluate suppliers in a quantified manner according to the CMMI's SAM process area and its corresponding specific goals. The present work has proposed a quantitative model to help teams and organizations to evaluate and select potential suppliers based on the previous history of each one of them. The proposed model has introduced some measures to evaluate the success and failure of each supplier of being compliant with the agreed SLAs. Besides, it presented some other measures to evaluate the risk of dealing with each supplier. Taking these two factors in consideration, a set of mathematical equations are used to come up with the SQScore which could help in categorizing the suppliers into three categories (low, medium, and high performing suppliers) based on the average value of the SQScores of all suppliers and the corresponding Standard Deviation. This quantitative model has been verified by implementing it in a practical extension tool. Furthermore, it has been validated by using this tool in about 600 companies which proves it is suitable.

Acknowledgment

We would like to acknowledge Eng. Hosam Kamel from Microsoft for his help during building the first version of the TFS SLA extension.

References

[1] F. Soares, S. Meira, An Agile Maturity Model for Software Development Organizations, in IARIA XPS Press, Venice, Italy,2013. https://www.iaria.org/conferences2013/ICSEA13.html.

[2] D. O'Neill, CMMI The Agile Way, CROSS TALK J. Def. Softw. Eng. 29 (2016) 1–9.

[3] M. Chaudhary, A. Chopra, CMMI for Development: Implementation Guide, Apress, 2017. //www.springer.com/de/book/9781484225288 (accessed July 31, 2018).

[4] L.S. Plácido, L. Araújo, S. Sampaio, M. Marinho, R. Godoi, A. Luna, Q. B. da Silva, Organizational Motivations for Adopting CMMI-based Software Process Improvement: An Extended and Updated Systematic Review, in: Fortaleza, Brasil, n.d.

[5] P. Mcmahon, P. Pem Systems, Extending agile methods: A distributed project and organizational improvement perspective, (2005).

[6] History: The Agile Manifesto, (n.d.). http://agilemanifesto.org/history.html (accessed July 17, 2018).

[7] Scott W. Ambler, Scaling Agile: An Executive Guide, (2010). ftp://public.dhe.ibm.com/software/emea/de/rational/agility/Scaling-Agile--An-executive-guide.pdf .

[8] T. Dingsøyr, S. Nerur, V. Balijepally, N.B. Moe, A decade of agile methodologies: Towards explaining agile software development, J. Syst. Softw. 85 (2012) 1213–1221. doi:10.1016/j.jss.2012.02.033.

[9] L. Williams, A. Cockburn, Guest Editors' Introduction: Agile Software Development: It's About Feedback and Change, Computer. 36 (2003) 39–43. doi:10.1109/MC.2003.1204373.

[10] D. Duka, Agile Experiences in Software Development, in: Opatija, Croatia, 2012. https://www.balcon-project.eu/events/mipro2012.

[11] K. Beck, Manifesto for Agile Software Development, (2001). http://agilemanifesto.org/principles.html.

[12] David F. Rico, What is the ROI of agile vs. traditional methods? An analysis of XP, TDD, Pair Programming, and Scrum, (2008). http://davidfrico.com/rico08b.pdf.

[13] Version One ,3rd Annual 'State of Agile Development' Survey results., (2008). http://www.versionone.com.

[14] VersionOne, 12th Annual State of Agile Report, VersionOne 12th Annu. State Agile Rep. (2018). https://explore.versionone.com/state-of-agile/versionone-12th-annual-state-of-agile-report.

[15] CHAOS Report: Decision Latency Theory (2018) Package - The Standish Group, (n.d.). https://www.standishgroup.com/store/services/10-chaos-report-decision-latency-theory-2018-package.html (accessed July 22, 2018).

[16] D. Batra, Modified Agile Practices for Outsourced Software Projects, Commun ACM. 52 (2009) 143–148. doi:10.1145/1562164.1562200.

[17] 10 reasons why outsourcing software development works!, (n.d.). https://www.orientsoftware.net/software-outsourcing/why-outsourcing/ (accessed July 31, 2018).

[18] S. Gantman, IT Outsourcing in the Public Sector: A Literature Analysis, in: Glob. Sourc. Serv., World Scientific/NOW Publishers, 2016: pp. 83–134. doi:10.1142/9789813109315_0003.

[19] W. Humphrey, Managing the Software Process, Addison-Wesley Publishing Company,Inc, 1989.

[20] E.M. Soares, S.R.B. Oliveira, M.C. dos Santos, A.M.L. Vasconcelos, A Catalog of Best Practices about Supplier Agreement Management and Agile Practices, in: IARIA XPS Press, Rome, Italy, 2016. https://www.iaria.org/conferences2016/ICSEA16.html.

[21] CMMI for Development, Version 1.3, Software Engineering Institute (SEI), 2010. https://resources.sei.cmu.edu/asset_files/TechnicalReport/2010_005_001_15287.pdf.

[22] CMMI - Supplier Agreement Management (SAM) Process Area, CMMi - Supplier Agreem. Manag. SAM. (n.d.). http://www.software-quality-assurance.org/cmmi-supplier-agreement-management.html (accessed August 10, 2018).

[23] Guidelines for Supplier Agreement Management in CMMI, CMMI Consult. Blog. (2013). https://www.cmmiconsultantblog.com/cmmi-faqs/guidelines-for-supplier-agreement-management-in-cmmi/ (accessed August 10, 2018).

[24] G. O'Regan, Supplier Selection and Management, in: G. O'Regan (Ed.), Concise Guide Softw. Eng. Fundam. Appl. Methods, Springer International Publishing, Cham, 2017: pp. 123–130. doi:10.1007/978-3-319-57750-0_8.

[25] M. Chaudhary, A. Chopra, CMMI DESIGN, in: CMMI Dev. Implement. Guide, Apress, 2017: pp. 9–69. www.springer.com/de/book/9781484225288 (accessed October 1, 2018).

[26] J. C. C. Furtado, Spider-ACQ: An Approach to the systematization of Products and Services Acquisition Process Based Quality Multimodels., Masters Dissertation, PPGCC in Federal University of Pará (UFPA), 2011.

[27] Supplier Agreement Management (SAM) (CMMI-DEV), (n.d.). https://www.wibas.com/cmmi/supplier-agreement-management-sam-cmmi-dev (accessed July 24, 2018).

[28] Magic Quadrant for Enterprise Agile Planning Tools, (2017). https://www.gartner.com/doc/reprints?id=1-3Z6ZBTD&ct=170501&st=sb.

[29] TFS SLA Server 2017 - Visual Studio Marketplace, (n.d.). https://marketplace.visualstudio.com/items?itemName=MohammedOmar.TFSSLAClient2017 (accessed August 16, 2018).

Feature-Based Sentiment Analysis for Arabic Language

Eng. Ghady Alhamad[1], Dr. Mohamad-Bassam Kurdy[2]
Master in Web Science, Syrian Virtual University, Hama, Syria[1]
Ph.D. in Mathematical Morphology, Syrian Virtual University, Dijon, France[2]

***Abstract*—In light of the spread of e-commerce and e-marketing, and the presence of a huge number of reviews and texts written by people to share views on products, it became necessary to give attention to extracting these opinions automatically and analyzing the feelings of the reviewers. The goal is to obtain reports evaluating products and contribute to improve services at a glance. Sentiment Analysis is a relatively recent study that deals with the processing of natural texts published in web sites and social networks. However, the processing of texts written in the Arabic language is one of the challenges that specialists face because people do not rely on standard Arabic, writing people in spoken/colloquial languages and use various dialects. This paper will present feature-based sentiment analysis for Arabic language which works on text analysis technique that breaks down text into aspects (attributes or components of a product or service), and then allocates each one a sentiment level (positive, negative or neutral).**

Keywords*—Sentiment analysis; feature-based; colloquial Arabic; opinion mining; natural language processing

I. Introduction

Sentiment analysis is an active research area since 2003 [1] and, it refers to the process of mining the texts in order to identify the tone of the passage written by the reviewers [2]. These tones are the focus for the decision makers to assess customer satisfaction with their products, which have been categorized into different poles. The most significant polarizations were absolutely in many studies, such as [3], [4] and others were usually three tones: positive, negative, and neutral. Sentiment analysis, which is also called opinion mining, is the computational study of people's opinions, sentiments, and attitudes about topics, entities, people and events, that are expressed in texts [5].

Recently the number of internet users has increased significantly in the Middle East and people are becoming more and more interested in buying online. According to new statistics [7] which have resulted that the number of internet users in the Arab countries has reached 157 million people, according to the Arabic Network for Human Rights Information. Internet buyers are distributed in the Middle East in several countries, reaching 10.6 million in Saudi Arabia, 6.8 million in the UAE, 2.4 million in Kuwait and 15.2 million in Egypt and around other Arab countries at different rates. The mobile phone is also the best-selling product online in the Arab world, according to the director of Souq. (Source: payfort) [7].

The Arabic language is one of the fastest growing languages on the web [6]. The main challenge in this study that sentiment analysis is for Arabic which is considered a poor area for this language. In addition to the peculiarity of the Arabic language whether in the Standard Arabic or in terms of the diversity of its dialects. The Arabic language is a Sematic language which consists of 28 letters. It is a cursive language, in which word formation consists of connecting letters to each other. As opposed to the English language. Arabic writing starts from right to left and has no capitalization [6].

Human can easily read texts and recognize reviewer's sentiment by understanding context, but for computers it is not normal process. Therefore, the main task in this study is to make computers recognizing the reviewer's sentiment and this achieved by Natural Language Processing (NLP). NLP is a framework to support an interaction between computers and human languages [8].

In this paper, based on the market need in the Arab world, and in light of the lack of Arab studies in this field with the wide spread of Arabic texts on the web written in various non-standard Arabic dialects, it was necessary to fill the gap and present a theory in this field. Since mobiles are the best-selling products, they will be the focus of this study. This theory exhibits a proposed method for recognizing Arabic sentiment phrases for mobile phones with consideration of each feature of phone like: camera, battery, memory … etc. The opinion phrases identified by building grammatical analyzer which is defining several forms for these phrases. Grammatical analyzer needs a lexical analyzer as input to define opinion tokens. opinion tokens could be mobile features, entities names and opinion words. Which could be positive names, negatives names, positive verbs, negative verbs, positive adjectives, negative adjectives, modifiers and negation words. This process called Parts of Speech Tagging (POS) that will be presented in this study. POS tagging has been used for a long time in text classification and NLP. POS tagging differentiates syntactic meaning of words in a sentence by using some specific tags, such as tags for noun, pronoun, verb, adjective, adverb, conjunction and others [8].

Also, after identifying opinion phrases the study will classify the opinion into five polarities in range [-2, 2]: {Strong positive, positive, neutral, negative, strong negative}. Finally, the summarization is necessary in order for decision-makers to gain knowledge.

The rest of this paper is organized as follows. Section 2 overviews related work. Section 3 describes the methodology followed with examples showing exactly how the study could achieve the goal. Section 4 presents the results of the experimental analysis and evaluation. Finally, in section 5, conclusions and possible future work are discussed.

II. Related Works

This section exhibits a number of related previous studies as this paper adopts some of their approaches and overcome the absence of some points for Arabic in others.

Bing Liu [9] is one of the most famous studies that cited by most researches in this field. He used rules for recognizing opinions and made for them [1]Backus-Naur Form (BNF). BNF is a meta syntax notation for context-free grammars, often used to describe the syntax of languages used in computing. This study adopted his approach for Arabic language. In Mohammad N. et al. [1] recognized opinions using lexicon, they concerned in Modern Standard Arabic (MSA) and colloquial for example: "Khaliji". In addition to Chetashri B. et al. [10] discussed the lexical and machine learning approach. Mongkol Seansuk et. al. [11] exhibited traditional methodology and evaluated opinions logically for each sentence; they considered opinion is positive by comparing sentences and the result will be positive only if both of them are positive, else negative.

Asad Ullah R. K. et al. [12] retrieved comments from YouTube to analysis sentiment about Android and iOS; they used General Architecture for Text Engineering ([2]GATE) component and build plugin. GATE is necessary component for NLP, as this paper used it to achieve multiple ideas. Weishu Hu. et al. [13] presented how to mine product features in opinion sentences. It made use of SentiWordNet based algorithm to find opinion of the sentence. Samir A et al. [14] presented a novel solution for Arabic Named Entity Recognition (ANER) problem, which aimed to boost the identification of extracted named entities. They utilized a machine learning technique using pattern recognition to classify name entities (NE).

Sana A. et al. [6] proposed study for Twitter sentiment analysis model that based on supervised machine learning and semantic analysis. They are divided their approach to two phases training and testing, in the training phase, they needed to learn from a set of labeled tweets for classifier. Then they used to classify unlabeled tweets in the testing phase. Mohamad H. et al. [16] also focused on studying sentiment analysis for Arabic language that collected from Twitter, Facebook and YouTube. Taysir .H et al. [15] focused on mining social networks for sentiment analysis of colloquial Arabic comments. The approach concerned with Egyptian terminology as it provided a structure to define the standard meaning of the word and the informal terms associated with this word. Alaa El-Dine A. H. et al. [17] also used classification methods to analyze users' comments and detected the comments that agree, disagree or is natural with respect to a post. The data collected from Facebook.

Sawsan C. et al. [18] adopted in their approach ontology for detecting Arabic Emotion. They detected language or dialect that belonged to with the help of GATE. They arrange the emotional vocabularies into intensities belonging to the integer numerical domain [-10, +10]. Whereas other studies detected specific dialect of Arabic language like Abdullah D. et al. [19]. Arabic Levantine tweets are a corpus of the study, they implemented different methods to automatically classify text messages of individuals to infer their emotional states.

Abdul-Mageed et al. [20] presented a subjectivity and sentiment analysis system (SAMAR) based on a Support Vector Machine (SVM) classifier for different Arabic social media applications: Web forums, chat, Wikipedia Talk Pages, and Twitter. They studied different features including word n-grams, POS tagging, and word stems. Also, many stylistic features related to social media applications were investigated. The results showed that the classifier performance relied on the type of the dataset and feature used.

III. Methodology

This section presents method for feature-based sentiment analysis for Arabic language. Mobile phone is the target product. Therefore, the study exhibits analyzing people's sentiment for mobile phones for each feature. As well as it presents entity recognition for mobile names. The study consists of the process as shown in Fig. 1.

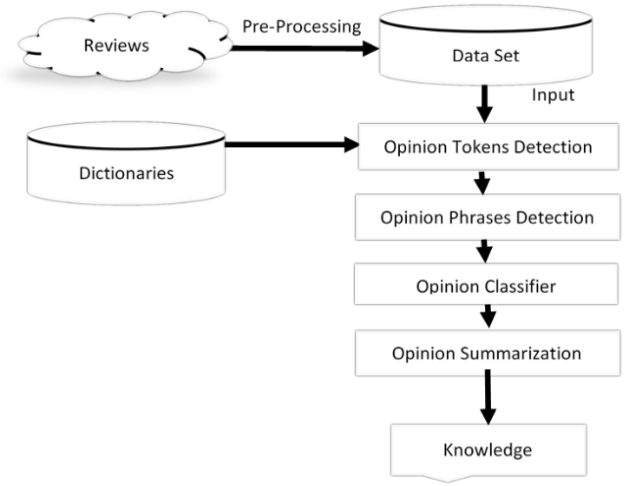


Fig. 1. Feature-based Sentiment Analysis for Arabic Language.

[1] *https://en.wikipedia.org/wiki/Backus%E2%80%93Naur_form*
[2] *https://gate.ac.uk/*

A. Dictionaries

This approach includes three dictionaries for features, sentiment words and entities. The data collected as a sample for training data. With the possibility of feeding these dictionaries later dynamically with flexibility, prior experience is not required. The dictionaries data collection source details in pre-processing section.

1) Features dictionary: Features are domain-based of sentiment analyzer and in this study the domain is mobile phones. This dictionary composes 84 words as a sample data and it is scalable.

E.g. for mobile features: camera; "كاميرا", memory; "ذاكرة", battery; "البطارية" ...etc.

2) Sentiment words dictionary: Sentiment words contribute to the quality of sentiment classifier. They are domain-independent unlike the features, but they are related to the terminology of the Arabic language in all its dialects. They were collected by relying on experiments from people's reviews, and space was also allowed for scalable.

Sentiment words are classified into several categories: five positive categories, five negative categories, negations category, and strong words (or modifiers) category.

The structure of sentiment words is shown in Table I. with number of words for each category and examples.

TABLE I. SENTIMENT WORDS STRUCTURE

Sentiment Words Categories		
Positive	***words***	***Example***
Strong Positive Adjectives	25	wonderful - "رائع"
Positive Adjectives	46	nice - "حلو"
Positive Names	22	advantage - "ميزة"
Positive Verbs	19	advice - "أنصح"
Positive Comparitive	13	better - "أحسن", "أفضل"
Negative	***words***	***example***
Strong Negative Adjectives	7	disappointed - "مخيب"
Negative Adjectives	23	weak - "ضعيف"
Negative Names	23	disadvantage - "عيب" problem - "مشكلة"
Negative Verbs	20	hate - "أكره"
Negative Comparitive	8	worst - "أسوأ"
Other	***words***	***example***
Negations	12	not - "ليس", "ما"
Strong Words (modifiers)	13	very - "كثير", "وايد", "يزاف"

3) Entities dictionary: Entities represent product names, which was one of the challenges this approach really face. Because there is no standard way to write mobiles names. Most of the mobile phone brands are not Arab. The main problem is when someone tries to write mobile name in Arabic alphabet. In addition, it may not be strange if others mixed spelling the name between Arabic and Latin letters.

An example of the different cases that reviewers write for the mobile name of the "Samsung Galaxy S6".

a) سامسونج S6. (Mixed, not full name).

b) Samsung S6. (Review in Arabic,mobile name in English, not full name).

c) جلاكسي اس6. (Arabic only, not full name with other parts).

d) جلاكسي S6. (Mixed, Not full name).

e) S6. (Version only, not Arabic).

f) اس6. (Version only, Arabic).

Also, the same word may have several spellings in Arabic, that the stemmer unable to stem them because that are not Arabic and have no meaning.

- "جلاكسي", "جالكسي", "جلكسي", "غلاكسي" (for Galaxy word).
- "سامسونج", "سامسونغ" (for Samsung word).

Therefore, this approach defines specific structure for mobile names as a hierarchy. Each level has multiple keywords to include all different spellings for the same name.

Fig. 2 shows the entity structure categorized into three levels with examples.

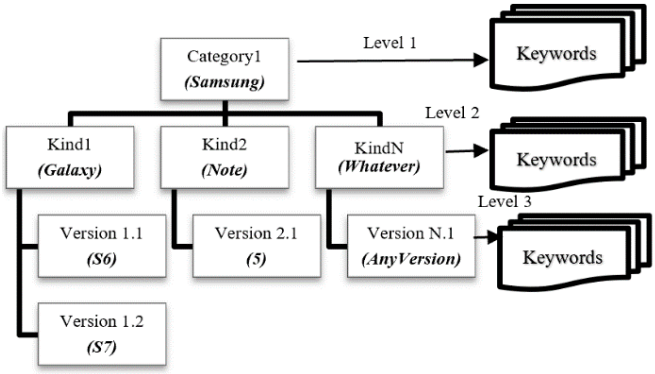


Fig. 2. Entity Data Structure.

This dictionary includes three mobile categories (brands), each category has several kinds and each kind has several versions. Each level has keywords list for different forms of the same word.

B. Pre-processing

1) Data collection: As sentiment analysis depends on the training data which labelled. The 85 posts of mobile phones are collected as a dataset, they include 1024 comments, which include 570 replies obtaining from mobile pages; [3]souq.com and [4] mobihall.com pages on Facebook. Most of posts are advertisement about mobiles therefore, the comments and replies are the target reviews.

2) Reviews structure and format: Since the reviews has been collected from different sources, the standard structure

[3]*http://souq.com*
[4]*http://mobihall.com*

became required. The appropriate format chosen to represent reviews is the "eXtensible Markup Language" (XML). The reviews include ratings as likes or stars, created date time and review id. Fig. 3 shows sample of reviews data.

```
<posts>
 <post>
 <id>261</id>
 <url>mobihallsocial</url>
 <socialmedia>Facebook</socialmedia>
 <postisadvertisment>false</postisadvertisment>
 <createdtime>16/04/2016</createdtime>
 <eval>
 <value>25.0</value>
 <type>LIKES</type>
 </eval>
 <message> سامسونج جلاكسى اس 6 موبايل انيق والمعالج الاكسيون اثبت
كفائة عالية ال3 جيجا بايت ليست كافية ابدا الكاميرا ممتازة تصميم اكثر من رائع
العيوب اللى قابلتها عدم وجود فتحة لكارت الميمورى وشريحة واحد
اتصال</message>
 <comments>
 <comment>
 <id>2763</id>
 <createdtime>16/04/2016</createdtime>
 <eval>
 <value>10.0</value>
 <type>LIKES</type>
 </eval>
 <message>وفي عيب اخر و هو البطاريه</message>
 <replies>
 <reply>
            <id>585</id>
            <createdtime>16/04/2016</createdtime>
            <eval>
            <value>0.0</value>
            <type>LIKES</type>
            </eval>
            <message> انا عندى نفس المشكله????</message>
 </reply>
 </replies>
 <comment>
 </comments>
 </post>
</posts>
```

Fig. 3. Sample of Reviews Sata.

As well as the dictionaries data that collected by reliance on the same sites, they are not only for reviews that represent dataset but also for all reviews of whole sites pages.

3) Arabic stemmers: This approach needs two stemmers for reviews and dictionaries lists:

a) Light stemmer is built-in by this study for noise elimination or normalization:

- ✓ Standardize Hamza "أ".
- ✓ Eliminate Tashkeel ً, ٓ, ٔ, etc.
- ✓ Standardize "ة", "ه".
- ✓ Standardize "ي", "ى" in the end of the word.
- ✓ Remove repeated letters "جداااااااا".
- ✓ etc.

b) Advance stemmer to extract root words. "Khoja" and "Arnlp" are the most famous stemmers for Arabic language. This approch used "Arnlp" because in addition to finding the root word, it works to find the stem word. The stem word may be more meaningful and reduce the confusion that occurs due to the presnce of one root for opposing words.

C. Sentiment Analyzer

The approach achieves natural language processing with GATE component. The process of sentiment analyzing consists of these steps:

1) Opinion tokens detection: The detection of opinion tokens is considered as the lexicon in this study. Opinion tokens include dictionaries lists; features, sentiment words and entities. The opinion tokens detection implemented using GATE Gazetteer. The GATE Gazetteer matches words in lists with the possibility of annotating each matched word. These annotations are very useful data for next steps. Therefore, Gazetteer includes following lists:

a) Features list includes: feature, feature identifier.

b) Sentiment words list includes: sentiment word, sentiment category, polarity, sentiment id..

c) Entities list for all keywords of mobiles names in one list, includes: keyword of mobile name, full mobile name, product id, level name.

Fig. 4 shows opinion tokens detection example. The lexical analyzer detects opinion tokens and classifies them based on its semantic meaning as kind of POS tagging.

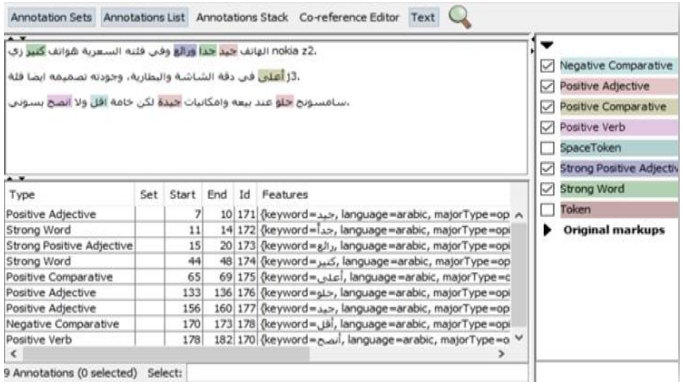


Fig. 4. Opinion Tokens Detection Example.

2) Opinion phrases detection: In this stage, the grammar analyzer is identifying opinion phrases syntax. The opinion phrases detection is performed using GATE JAPE transducer. The GATE JAPE transducer defines rules for forms of opinion phrases. It takes opinion tokens annotation as input and the output is the opinion phrases annotation. Before identifying the opinion phrases rules, using JAPE the reviews should be split into sentences by GATE Sentence Splitter to detect phrases for each sentence separately. Opinion phrases rules described with [1]BNF meta syntax notation that used to describe the syntax of phrases.

Fig. 5 shows the suggested syntax by this approach of opinion phrases for Arabic language. The BNF identifies six rules (or cases) for the forms of opinion phrases rules, and two rules for compare two products.

<OpinionRule_1> ::= (Product | Feature)? (Negation)? (SentimentWord) (Modifier)?

<OpinionRule_2> ::= (Feature) (Product) (Negation)? (SentimentWord) (Modifier)?

<OpinionRule_3> ::= (Feature | Product)(((Modifier)? (SentimentWord)) | ((Negation)? (SentimentWord) (Modifier)?))+

<OpinionRule_4> ::= (Feature) (Product) (((Modifier)? (SentimentWord)) | ((Negation)? (SentimentWord) (Modifier)?))+

<OpinionRule_5> := (Negation)?
(((Modifier) (SentimentWord)) | ((SentimentWord) (Modifier)) | (SentimentWord)))
((Product) | (Feature) (Modifier)?)+

<OpinionRule_6> ::= (Negation)? (((Modifier) (SentimentWord)) | ((SentimentWord) (Modifier)) | (SentimentWord)) ((Feature) (Product))+

<CompareOpinionRule_1> ::= (Feature1)? (Product1 | Category1 | Kind1) (Negation)? (SentimentWord: Comparative) (Feature2)? (Product2 | Category2 | Kind2)

<CompareOpinionRule_2> ::= (Product1 | Category1 | Kind1) (Negation)? (SentimentWord: Comparative) (Feature1)

Fig. 5. Opinions Phrases BNF.

Notes about BNF:

a) Product is the entity with full mobile name.

b) Category is the first part of mobile name, that often represents company name.

c) Kind is the second part of mobile name.

d) Version is the third part of mobile name.

Fig. 6 shows opinion phrases detection example. The grammatical analyzer detects opinion phrases to test eight rules. The bottom table in the figure shows necessary information, opinion words, polarity, mobile features, mobile name and rule identifier.

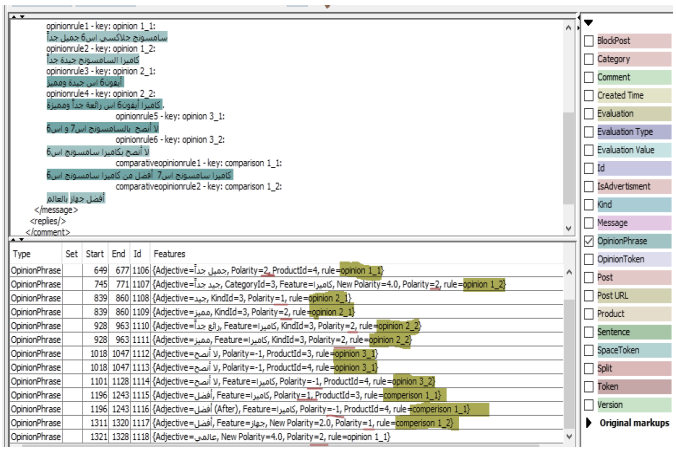


Fig. 6. Opinion Phrases Detection Example.

3) Opinion classifier: This stage classifies opinions into polarities. The polarities divided into five categories; {Strong Positive, Positive, Neutral, Negative, Strong Negative}.

The classification begins for each of the opinion phrases that are defined in the BNF. The opinion phrase cannot be neutral at all, but it is possible that the entire review is neutral when there is balance between the positive and negative opinion phrases in the same review. The polarities have been defined mathematically in the following ranges values between [-2 and +2]:

a) Strong Negative: [-2, -1.5].

b) Negative:] -1.5, -0.5].

c) Neutral:]-0.5, 0.5].

d) Positive:] 0.5, 1.5].

e) Strong Positive:] 1.5, 2.0]

Since the review consists of one or more opinion phrases and polarity value is fuzzy in the range [-2, +2], the polarity will be calculated by average function for polarities of opinion phrases.

$$\textbf{\textit{Polarity}} = \frac{\sum Polarity_{opinion\_phrase} \times (weight + 1)}{(weight + 1)} \quad (1)$$

Weight either represents the number of opinions for specific polarity or the number of likes. Likes mean if someone has copied the same opinion and gets the same polarity. The polarity value is multiplied by weight + 1. +1 represents the opinion itself and avoids dividing by zero.

Suppose 9 positive opinion phrases, 1 strong positive opinion phrase, 1 negative opinion phrase, and 4 strong negative opinion phrases. Where Strong Positive 2, Positive 1, Negative -1, Strong Negative -2. The (2) shows for polarity calculation.

$$\textbf{\textit{Polarity}} = \frac{9\times(1)+1\times(2)+1\times(-1)+4\times(-2)}{9+1+1+4} = \frac{2}{15} = 0.1333 \quad (2)$$

Final result based on ranges that are shown in the beginning of this section where 0.1333 ϵ]-0.5, 0.5], therefore, it is Neutral.

The Table II shows example for each rule defined in the BNF. It must be pointed out that in Arabic grammar, the noun comes before the adjective, in contrast to English grammar, where adjective precedes the noun that is being described, in addition to some other differences, therefore the translation of examples is only for illustration and it is not necessary that is correct for English. For example, "Red Flower", in Arabic, it is written as "Flower Red" – "الزهرة حمراء". Therefore, the illustration respects BNF rules and word order.

TABLE II. EXAMPLE FOR BNF RULES (OPINION PHRASES)

Opinion Phrases Examples			
Rule Name	***Example***		***Polarity***
OpinionRule_1	سامسونج جلاكسي اس6 جميل جداً	Samsung Galaxy S6 is nice so much	2
OpinionRule_2	كاميرا السامسونج جيدة جداً	Camera of Samsung is good very much	2
OpinionRule_3	جـ يد 6 اس أيـ فون وممـ يز	Iphone 6s is good and distinctive.	{1, 2}
OpinionRule_4	كاميرا أيفون اس6 رائعة جداً ومميزة.	Camera of iPhone 6s is wonderful very much and distinctive	{2, 2}
OpinionRule_5	لا أنصح بالسامسونج اس7 و اس6	I do not advise you with Samsung s7 and s6	-1
OpinionRule_6	لا أنصح بكاميرا سامسونج اس6	I do not advise you with camera of Samsung s6.	-1
CompareOpinionRule_1	كاميرا سامسونج اس 7 أفضل من كاميرا سامسونج اس 6	Camera of Samsung s7 is better than camera of Samsung s6	1 for product1 -1 for product2
CompareOpinionRule_2	أفضل جهاز بالعالم	The best device in the world	1

4) Opinions summarization: This method summarizes results of sentiment analyzer in several ways:

a) Feature-based: The method summarizes the results of opinion polarity for each feature separately.

b) ReviewDate-based: The method summarizes the results of opinion polarity during specific time periods.

c) Polarity-based: The method summarizes the results of opinion polarity as a ratio for each polarity.

d) Product-based: The method summarizes the results of all opinions polarity.

IV. EXPERIMENTS AND RESULTS

This section presents the experiments and results. The experiments are performed to analyze the quality of the proposed methodology whereas in the results will present the results of this study with examples. The experiments are achieved with precision, recall and f-measure for opinion tokens detection and opinion phrases detection.

As for opinion tokens detection are evaluated by the dictionaries size and by stemmer for matching words. The Table III shows a test with 20 tokens extracted from several reviews consists of 100 words.

TABLE III. OPINION TOKENS DETECTION EXPERIMENT

	Positive (Retrieved)	***Negative (Not Retrieved)***
True	18	77
False	3	2

The Precision for opinion tokens is defined by the formula:

$$\boldsymbol{P} = \frac{TP}{TP+FP} = \frac{18}{18+3} \approx 0.86 \quad (3)$$

The Recall for opinion tokens is defined by the formula:

$$\boldsymbol{R} = \frac{TP}{TP+FN} = \frac{18}{18+2} \approx 0.9 \quad (4)$$

From (3) and (4) results, the F-measure is defined by the formula:

$$\boldsymbol{F} = 2.\frac{\mathrm{R}\times P}{\mathrm{R}+P} = \frac{2\times(0.9\times 0.86)}{0.9+0.86} \approx 0.88 \quad (5)$$

As well as the opinion phrases detection are evaluated by measuring the quality of opinion rules that defined in BNF in the METHODOLOGY section. The Table IV shows a test with 35 phrases extracted from several reviews consists of 60 phrases.

TABLE IV. OPINION PHRASES DETECTION EXPERIMENT

	Positive (Retrieved)	***Negative (Not Retrieved)***
True	31	20
False	5	4

The Precision for opinion phrases is defined by the formula:

$$\boldsymbol{P} = \frac{TP}{TP+FP} = \frac{31}{31+5} \approx 0.86 \quad (6)$$

The Recall for opinion phrases is defined by the formula:

$$\boldsymbol{R} = \frac{TP}{TP+FN} = \frac{31}{31+4} \approx 0.89 \quad (7)$$

From (6) and (7) results, the F-measure is defined by the formula:

$$\boldsymbol{F} = 2.\frac{\mathrm{R}\times P}{\mathrm{R}+P} = \frac{2\times(0.89\times 0.86)}{0.89+0.86} \approx 0.87 \quad (8)$$

The results show in followed figures some models for opinions summarization to build knowledge that can benefit decision makers: Fig. 7 shows bar chart for feature-based statistics about comparison of two mobiles Sony Xperia Z5 and Sony Xperia Z3. It shows polarity for each feature in range [-2, 2].

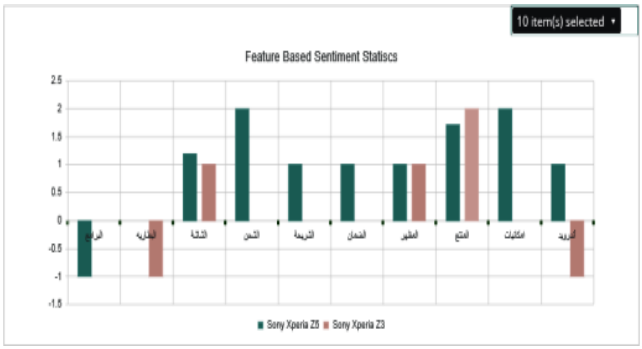


Fig. 7. Feature-Based Summarization.

Fig. 8 shows pie chart that summarizes the percentage rate for each polarity.

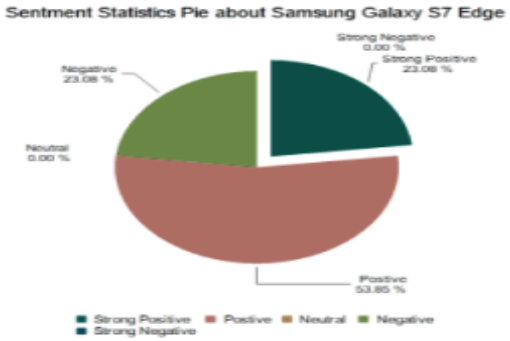


Fig. 8. Polarity-Based Summarization.

Fig. 9 shows line chart about comparison of two mobiles that summarizes the polarity for each mobile through specific periods. It illustrates statistics from 2014 to 2016. Polarity in range [-2,2] for mobiles Apple iPhone 5s and Sony Xperia Z5.

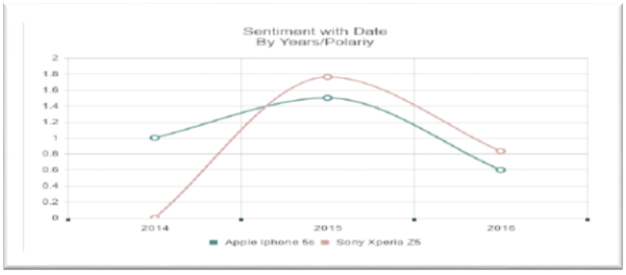


Fig. 9. ReviewDate-Based Summarization.

Fig. 10 shows gauge chart about comparison of two mobiles that summarizes the final polarity for each mobile. Polarity in range [-2,2] for mobiles Apple iPhone 5s and Sony Xperia Z5.

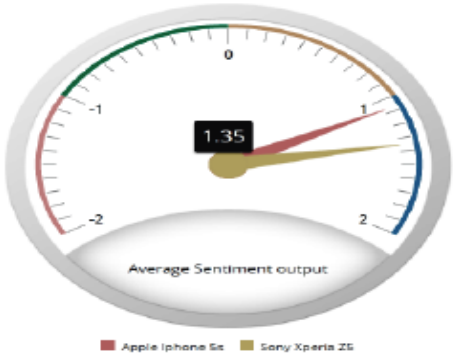


Fig. 10. Product-Based Summarization.

V. Conclusion

This paper proposes feature-based sentiment analysis for Arabic language, the target product is mobile phone. The results of this mining are demonstrated as the degree of strong positive, positive, neutral, negative and strong negative. This result is useful for both consumers and companies. This study presents an approach in active area for Arabic language. The f-measure rate from experimental result is 88%. The study presents an effective method for identifying opinion phrases by building Arabic grammatical analyzer with good result and expandable. The future works will be focusing on entering new categories of products and services, support grammatical analyzer with new rules, expand dictionaries, in addition to include other platforms of social media. The sentiment of emoji is one of the future works.

References

[1] M. N. Al-Kabi, A. H. Gigieh, I. M. Alsamadi, H. A. Washeh and M. M. Haidar, “Opinion Mining and Analysis for Arabic Language,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 5, no. 5, p. 15, 2014.

[2] Dr.S. Murugavalli, U. Bagirathan, R. Saiprassanth and S. Arvindkumar, “Feedback analysis using Sentiment Analysis for E-commerce,” International Journal of Latest Engineering Research and Applications (IJLERA), vol. 02, pp. 84-90, 30 March 2017.

[3] F. J. A. P. Mattosinho, “Mining Product Opinions and Reviews on the Web,” Master Thesis, W. M. Medieninf, Rer. Nat. Habil, H. C. Alexander Schill (Advisors), Chair of Computer Networks, 2010, July,.

[4] M. El-Masri, N. Altrabsheh, H. Mansour, A. Ramsay, “A web tool for Arabic sentiment analysis,” Procedia Computer Science, vol. 117, pp. 38-45, 2017.

[5] B. Pang, L. Lee, “Opinion mining and sentiment analysis,” Foundation and Trends in Information Retrieval, Vol. 2 No. 1-2, 1-135, 2008.

[6] S. Alowaidi, M. Saleh, O. Abdulnaja, “Semantic Sentiment Analysis of Arabic Texts,” International Journal of Advanced Computer Science and Applications (IJACSA), Vol 8, No. 2, July 2017.

[7] iweb digital advertising agency, “E-commerce stats,” iweb123.com, accessed in 2020-10-3 website.

[8] S. Siddiqui, M. Abdul Rahman, S. Daudpota, A. Waqas, “Opinion Mining: Approach to Feature Engineering,” International Journal of Advanced Computer Science and Applications (IJACSA), Vol 10, No. 3, 2019.

[9] B. Liu, "Sentiment Analysis and Opinion Mining," Morgan & Claypool, Chicago, 2012.

[10] C. Bhadane, H. Dalal, H. Doshi, “Sentiment analysis: Measuring opinions,” Procedia Computer Science, vol. 45, pp. 808-814.

[11] M. Seansuk, P. Songram and P. Chomphuwiset, “Feature-Based Opinion Mining On Smart-Phone Reviews,” Proceedings of the 3rd IIAE International Conference on Intelligent Systems and Image Processing, p. 5, 2015.

[12] A. U. R. Khan, M. Khan and M. B. Khan, “Naïve Multi-label classification of YouTube comments using comparative opinion mining,” ELSEVIER, Procedia Computer Science 82, vol. 82, pp. 57-64, 12 May 2016.

[13] Weishu Hu, Zhiguo Gong, JingzhiGuo, “Mining Product Features from Online Reviews,” IEEE International Conference on E-Business Engineering, 2010.

[14] S. AbdelRahman, M. Elarnaoty, M. Magdy, A. Fahmy, “Integrated Machine Learning Techniques for Arabic Named Entity Recognition,” International Journal of Advanced Computer Science and Applications (IJACSA), vol 7, issue 4, No. 3, July 2010.

[15] T. Hassan, M. Ali M, A. Hedar, M. M. Doss, “Mining Social networks’ Arabic Slang Comments,” Proceedings of IADIS European Conference on Data Mining, 22-24 July.

[16] M. Hammad, M. Al-awaidi, “Sentiment Analysis for Arabic Reviews in Social Network Using Machine Learning,” Information Technology, Springer, 2016, pp. 131-139.

[17] A. El-Dine Ali Hamouda, F. El-zaharaa El-taher, “Sentiment Analyzer for Arabic Comments System,” International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 4, No. 3, July 2013.

[18] S. Cassab, M. B. Kurdy, “Ontology-based Emotion Detection in Arabic Social Media,” International Journal of Engineering Research & Technology (IJERT), Vol. 9 issue 08, August 2020.

[19] A. Daood, I. Salman, N. Ghanem, "Comparison study of automatic classifiers performance in emotion recognition of Arabic social media users," Journal of Theoretical and Applied Information Technology (JATIT), Vol. 95 No. 29, Octobor 2017.

[20] M. Abdul-Mageed, S. Kuebler, M. Diab, "SAMAR: A System for Subjective and Sentiment Analaysis of Social Media Arabic," 3rd Workshop on Computional Approaches to Subjectivity And Sentiment Analysis (WASSA), 2012.

AUTHORS' PROFILE

Eng. Ghady Alhamad Born in Syria/Hama July/1987, She obtained bachelor degree in information systems engineering from Syrian Virtual University - Syria December, 2012, Master in Web Science from Syrian Virtual University - Syria Feb, 2013, She started to work as a Software Engineer from 2013. She prefered to work remotely because of her healthy profile with challanging all drawbacks. She worked between UAE and KSA companies.

She is supervisor of graduation projects in Information Systems Engineering Program (Under graduate students) and Master projects (Post graduate) in Web Science Program, Syrian Virtual University since 2015.

She did some special projects: Java compiler, e-commerce sites, web applications, messengers and in augmented reality field View Art in Room. She started in 2020 to publish her works in social media pages on LinkedIn page & Facebook page to leave imprint in this field.

Ph.D. Mohamad-Bassam Kurdy Born in Syria/Damascus July/1961, He obtained Master degree in information systems engineering from INPG - France 1986, Ph.D. in Mathematical Morphology from Mines ParisTech -France 1990, He worked at HIAST 1991-2013. He was Head of Computer Sciences at HIAST between 1997 and 2003. Country Manager for Syria of EC project EUMEDIS -Medforist. Actually Professor at SVU, ESC Dijon and ESC Rennes teaching: Advanced Data Mining, Big Data, Information Retrieval, cbIR (content based Image Retrieval) and Supervising many Master student projects (postgraduate).

Permission Extraction Framework for Android Malware Detection

Ali Ghasempour[1], Nor Fazlida Mohd Sani[2], Ovye John Abari[3]
Department of Computer Science, Universiti Putra Malaysia
43400 UPM Serdang, Selangor, Malaysia

***Abstract*—Nowadays, Android-based devices are more utilized than other Operating Systems based devices. Statistics show that the market share for android on mobile devices in March 2018 is 84.8 percent as compared with only 15.1 percent iOS. These numbers indicate that most of the attacks are subjected to Android devices. In addition, most people are keeping their confidential information on their mobile phones, and hence there is a need to secure this operating system against harmful attacks. Detecting malicious applications in the Android market is becoming a very complex procedure. This is because as the attacks are increasing, the complexity of feature selection and classification techniques are growing. There are a lot of solutions on how to detect malicious applications on the Android platform but these solutions are inefficient to handle the features extraction and classification due to many false alarms. In this work, the researchers proposed a multi-level permission extraction framework for malware detection in an Android device. The framework uses a permission extraction approach to label malicious applications by analyzing permissions and it is capable of handling a large number of applications while keeping the performance metrics optimized. A static analysis method was employed in this work. Support Vector Machine (SVM) and Decision Tree Algorithm was used for the classification. The results show that while increasing input data, the model tries to keep detection accuracy at an acceptable level.**

Keywords*—*Malware detection; android device; operating system; malicious application; machine learning

I. Introduction

In recent time, Android is the most famous platform for mobile devices [1]. Saving and storing confidential data such as banking information or contact numbers is part of every mobile device. Hence providing a barrier between these information and an attacker is highly needed. Nowadays antiviruses are greatly developed and provide a wide range of options based on user needs. However, studies show most Android users do not rely on antiviruses to secure their phones [2]. So, this can be a great motivation for the attacker to focus on the Android platform. Kaspersky, which is one of the biggest security solutions company released information related to malicious activity on the Android platform in 2018. Based on the statistics, 5,730,916 malicious packages were detected by their lab [3]. Thus, protecting Android devices from misuse or any malicious application is important and needed.

Mobile malware can be classified into three (3) groups [4]. These are Malware, Grayware, and Spyware. The Malware focuses on gaining access to personal data or damaging to the hardware (mobile device). One of the main aims of Malware attacks is unprivileged access to user personal information. Malware attacks are SMS, Bluetooth, GPS, and Root attacks. SMS attack is mostly related to phishing and adware. In Bluetooth attack, an attacker can steal user personal information or location services. GPS attacks can compromise GPS devices on mobile phones to steal user location. In Root attack, an attacker can get privileged access to the phone operating system to install or remove applications. Grayware is an attack that mostly focuses on the marketing side without any damaging to phone or user data. Spyware is the most frequent type of attack related to the mobile phone which steals user data and sends it to an unwanted application rather than the original one. Android malware can further be specifically categorized into Worm, Trojan, Back-doors, Botnet, Spyware and Ransomware [5].

To analyze the behavior of application in the Android platform, two methods of analysis are commonly used. These are Dynamic and Static analysis [5]. Behavioral or dynamic focus on application behavior during run time. For example, requested permissions while an application is running can determine application intend. Additionally, a system call is one of the main features which can be measure for application behavior in runtime. To address why system call can be named as dynamic analysis, an Android operating system is based on Linux. A system call is provided by the application to request specific resources whether hardware or library from Linux kernel. Requesting irrelevant resources for an application can be marked as suspicious activity. [6].

The static analysis is mostly applicable for analyzing application by disassemble package and extracting source. Like dynamic analysis, different approaches exist in static analysis. Some researches focused on finding a sequence of op-code that reflects the malicious activity. The reason why op-code can be named as a feature is that every instruction in the computer program follows a specific structure. In this structure op-code determine the action of this instruction and the rest is an address in memory. Malicious packages may follow the same action in instruction, and so, op-code can be a good classifier for detection.

Permission is a famous feature for Android malware detection and it expose the exact intent of the application. It is designed to protect user privacy on the Android platform. Permission is requested by an application to authorize itself to access specific hardware or software resource. For instance, the sensitive user data (such as SMS or contacts) or system features (such as a camera or internet). Based on application

nature, the system may grant permission or ask the user to grant it. Basically, no application has the authorization to access or read other application data, system files, and user private data (such as SMS). Details related to permission are discussed in the next section. Fig. 1 illustrates how Android malware can be categorized based on different methods of analysis.

In Malware detection, after the feature selection phase, designing the model using real-world input sample data is the next phase. In the design phase, different statistical and mathematical techniques are used to prune unnecessary features from input data as well as to highlight selected features. A well-structured model can improve detection accuracy. For the classification phase, existing works mostly used machine learning, deep learning, and statistical methods to carried out the classification. Although most of the techniques are based on mathematics but the terms are different.

Machine learning is one of the basis and famous techniques used in the classification phase. It is a general term and contains many different algorithms inside. The main idea is to design a model based on known input data to predict unseen input data. In Machine learning, we try to predict as precious as possible for unknown input data which is called test data. Most of the algorithm follows two steps; train and test. In the training phase, the machine tries to make a relevant model based on labelled input data and whereas, in the test phase, the machine will predict based on its knowledge of previous learning. Moreover, machines may work on an unforeseen situation where the train phase does not exist. Therefore, machines try to find a pattern to distinguish between objects. Two main machine learning techniques are classification (Supervised learning) and clustering (Unsupervised learning) [7]. Supervised learning uses labelled data for the classification. Here, the features need to be defined by the operator, and results are described through the mapping of input to output by rules that are provided by the supervisor. Unsupervised learning focuses on finding a pattern in unlabelled or not fully labelled data. The results mostly are the grouped of input data. In clustering problems, machines try to group data with the same features instead of classifying them.

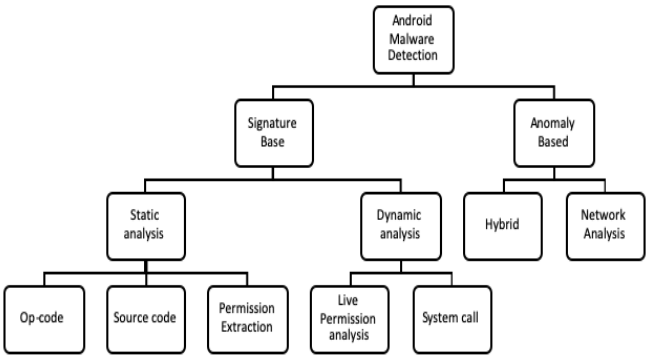


Fig. 1. The different Methods in Android Malware Detection.

Recently, deep learning has been used for classification. In deep learning, instead of focusing on linear analysis, we target multiple levels of abstraction. Data representation in each layer transfers to the next level to achieve a more accurate level of abstraction for input data. Also in this method, feature selection is used to improve the accuracy [8]. However, one of the drawbacks of the deep learning method is that it uses a lot of resources and takes a long time due to the high level of decomposition to complete the classification. The statistical method is used for the classification and to differentiate between applications. The statistical model uses less computational resources than machine learning techniques [9].

In this work, the authors focused on permission as an input feature. Because we have only one feature to analyse, the model is designed based on binary classification. The proposed model used different statistical techniques to boost computation while keeping accuracy high. Permission Support and Permission Ranking were used in the early steps of this model. Principal Component Analysis (PCA) was used to find different views of permission correlation. Dimensionally reduction using Kaiser and Cumulative techniques was implemented to select the most significant permissions. Selected permissions were compared to the results of Li [10] to select similar permission. Also, separate analysis has been done on our method without any overlap with the existing results. The last phase of this work focused on classification where Supervised Machine Learning approaches (Support Vector Machine and Tree-based algorithm) were used for the classification because of their popularity. In summary, the main contributions of this paper are as follows:

Proposed a multi-level permission extraction framework that improved malware detection accuracy in Android devices.

The framework designed was able to handle a large number of applications while keeping the performance metrics optimized.

The organization of the paper is as follows. Section 2 presents the literature review or related work. Section 3 gives the details about the methodology used. Section 4 gives the details about the proposed framework. Section 5 is the experiments setting and results. Section 6 gives a summary of the paper.

II. Literature Review

The important concepts in the understanding of malware detection in Android devices were discussed. The current trend and techniques used to develop an Android based malware detection model are also explained in this section.

A. Static based Analysis

The static based analysis includes the following methods**.**

- Source Code

[11] proposed a model to extract the source code of application using machine learning and clustering. Firstly, the authors extract the source code and then tokenized it with the n-gram method and used a bag of words (BoW) algorithm,

which is one of the NLP algorithms. Then they used both supervised and unsupervised machine learning techniques to get their result. The authors used ensemble learning by repeating the experiment 10 times combining with other methods and use the majority of results to make predictions. With the SVM method, they achieved a 95% detection rate. However, with the clustering method, they obtained only a 72% detection rate. Although the ensemble method showed better result with 95.6% detection rate [11].

- Op-Code

[12] developed a Deep Android Malware Detection to get a sequence of op-code and used a convolutional neural network (CNN) in detecting malicious applications. They used a rich dataset for the training of the model. The authors used a one-layer convolutional neural network over three types of datasets. For a small dataset, they achieved a 97% accuracy and for the large and extra-large dataset, they got 78% and 86% accuracy respectively. Their results show that the model produces better results as the dataset and training time are increasing [12].

- Permission

[11] proposed a model for Android malware detection using machine learning. The authors used the ensemble technique to get more accurate results. The results for classification with the SVM algorithm were 87.9% and using ensemble learning with C.45, Random Forest, Random tree, SVM, Logistic regression was 89.4%. Using the clustering method, they achieved a 64.6% detection rate [11].

- Hardware and Permission

Multimodal learning is one of the neural network techniques that use multi-input as network input. [13] applied static analysis to extract features from android packages. The features the authors considered in their work are permission and hardware features. They two features were chosen because the requested permission can define access to sensitive data, and the hardware permission can access physical resources in mobile devices such as cameras. The authors used these features separately to achieve a higher degree of accuracy. They further combined both features to improve and obtain a better performance result. Grid search was used over selected hyperparameters to optimize performance value. The analysis demonstrates that the overall accuracy achieved was 94.5%.

- API Tag and Permission

CENDroid was developed by [14] as a static-based Android malware detection based on API tags and permission. They formed a different combination of features for analysis behaviour and set up five machine learning algorithms for the classification. Afterward, three ensemble combination techniques were used. CENDroid was based on the clustering training dataset into an optimized number of clusters and then used ensemble learning for each of them. Ensemble model help to improve the performance as it can handle dissimilarity due to misclassification. The proposed model was trained and tested on different datasets and results revealed that linear SVM achieved a higher degree of 96.13% accuracy as compared to other classifiers. In ensemble learning, the weighted majority voting method hits the highest accuracy of 97.38% as compared to the staking and majority voting [14].

B. Dynamic based Analysis

Th system call method in a dynamic based analysis is discussed below.

- System Call

M0droid was proposed by [15] as a client/server architecture for detecting the malicious application using system call analysis. In this model, when an application is installed on a phone, the client check hash of APK file with his database to detect whether it is listed or not. If it is not listed, the server will send the APK file to do further investigation. The application will be run on an isolated environment and all of the system calls will be captured. After that, a vector that was normalized using z-score store how many times the system call is repeated. A signature is made from application and Spearman's rank correlation was used to compare the database and labelled the application. After setting the threshold value to 0.92, the result obtained was 60%.

C. Hybrid Analysis

[16] used a hybrid approach (a combination of analyzed source code and getting the requested permission) to proposed their model for malware detection. The developed model is capable of extracting three items. First, meta-data of application which contain permission of applications. Second, the binary of *Dex* code and library used in APK. Third, analysing the *Dalivk* assembly. A fuzzy fingerprint was used to make DNA for each application. Further, the authors proposed another malware detection, called ROAR for comparing the mentioned feature from the new application. Their framework is divided into three parts, family fingerprinting, peer fingerprinting, and merged fingerprinting. They achieved an accuracy rate of 95% on the last two methods and 85% for family fingerprinting.

[17] proposed a model by extracting nine (9) features from every application. The *DroidBox* is used to emulate the application in a virtual environment. Cryptography material, network operation, file operation, *Dexclass* load, information leak, sent SMS, phone call, service start, receive reaction and system call is used for malware detection. A Chi-Square algorithm was applied for dimension reduction and further used ensemble learning and combined different machine learning techniques. The best success rate among the different datasets and machine learning methods was 96.42%.

[18] discussed the importance of Android devices on the current Internet of Things (IoT) space. It was revealed that securing IoT devices that are running on Android is an issue. The authors used Factor Analysis of Information (FAIR) model to measure the risk associated with the IoT devices. In a situation of threat, they considered Situational Awareness (SA) to recognize the environmental elements. The proposed model contains three layers. The first layer, using machine learning, detect malicious behaviour of the attacker. The second layer mapped the selected feature in the first layer to

LEF factor of the FAIR model. The third layer, which is the decision-making layer finalized the process. The results demonstrate that with different machine learning techniques, different degrees of accuracy can be achieved. Based on the provided data, linear SVM has the highest accuracy of 99% as compared to other methods.

Author in [19] focused on different features to differentiate between Android applications and proposed a malware detection model. Op-code and API calls are choosing as desire features for analysing. The proposed model contains five layers. After feature extraction from APK files, Op-codes are tokenized by n-gram technique and API call by API frequency vector. A Convolutional neural network (CNN) was used as the main analyser in this model. Due to the high number of dimensions, principal component analysis (PCA) was used to distinguish significant features. The Neural network outputs are passed to feature fusion to achieve a higher degree of accuracy. SoftMax was used as the classification algorithm. The results focus on achieving accuracy while keeping runtime low. On 0.034 seconds, 95.1% accuracy achieved.

D. Significant Permission Identification

To achieve malware detection with high certainty, the Significant Permission Identification is used to achieve a high detection rate with the lowest possible permissions for analysing. This method focuses on permission which has a high chance to labelled as malware and omit permissions with low effect on our detection. To implement, three-layer data filtering is used for analysing real-time data. These are: (1) permission ranking with negative rate, (2) support based permission ranking and (3) permission mining with association rules. After applying these permissions, a supervised machine learning model is trained to detect further application.

- Data Pruning

The first step for analysing is to eliminate the necessity of considering all of the permissions for analysing. Due to the high number of applications in our dataset and each application consist of many permissions, it takes time to analyse all of them. We use Multi-Level Data Pruning (MLDP) to detect the most affectable permissions. The following are discussions about MLDP solution.

1) Permission Ranking with Negative Rate (PRNR): In every application, any requested permission reflects the need of the application. As far as we know the malicious application follows a specific subset of permission, then there will be no need to study all of the permissions. The algorithm only focuses on dangerous permissions and frequently requested permissions. Also, it can separate malicious applications from the benign applications with rare permissions requested by malware applications. For increasing the detection rate, the algorithm is omitting the same permission requested by both classes. To come up with one solution, permission ranking with negative rate is used to distinguish not only high-risk permission but also the permission used by benign applications. The goal is to differentiate normal and malware attacks.

We define two matrices, M and B. M is defined as the list of permission used by malicious application and B is the list of permission used by the benign application. M_{ij} says whether permission j^{th} is used by the i^{th} malicious application or not. If the answer is yes then 1 otherwise 0. The same goes for B_{ij} for benign applications.

In the first step, we need a balance between malicious applications and benign applications. In this work, the size of the benign application is 310926 while malicious is only 5494 applications, which caused skew in our model. Therefore, it is used in the equation 1 to find the support of each permission in a larger dataset and then scale down that permission to match smaller datasets. In this case number of benign B is more than malicious M.

$$S_B\left(P_J\right) = \frac{\sum_i B_{ij}}{Size(B_J)} * Size(M_J) \quad (1)$$

P_j describes j^{th} permission and $S_B(P_j)$ is the support of j^{th} permission in the B matrix. After rescaling, we can deploy PRNR in the equation 2:

$$R(P_j) = \frac{\sum_i M_{ij} - S_B(P_j)}{\sum_i M_{ij} + S_B(P_j)} \quad (2)$$

The $R(P_j)$ shows the rate of j^{th} permission. The range for an answer could be from [-1, 1] which 1 shows the permission P_j exists in the malicious dataset, so it can be categorized as danger permission. Also, if it is -1, it means that permission exists in the benign dataset which is low-risk permission. If 0, it means that it doesn't have a special impact on the detection. After that, two lists are generated from $R(P_j)$ in a form of ascending or descending order. To continue, the Permission Incremental System (PIS) is proposed for getting top permission from the benign and malicious list with the following metrics: True Positive, False Positive, Precision, recall, accuracy and F-measure. The aim is to find out those permissions which have much effects on the whole of the datasets. For the 310926 benign applications and 5494 malicious applications, all of the requested permissions are 135 and decreased to 95 with PRNR.

2) Support Based Permission Ranking (SPR): To narrow the result of PRNR and support each permission is considered that we focus on how many times one permission is repeated in the whole of the dataset or if permission exists only on a specific dataset. Therefore, we can omit permission with the lowest support. To implement this, the Permission Incremental System (PIS) is used to find the most supported permission. Our results show that 25 permission out of all the permissions is supportive.

3) Permission Mining with Association Rule (PMAR): After two steps pruning data, we get 25 most significant permission. To look closer to these permissions, it seems that some of the permissions are formatted as pairs. As an example, permission WRITE_SMS always comes with READ_SMS and so, we can consider them as one. To find this association between permissions, we used association rules. These rules are used for discovering relations between entities in the dataset. It only focused on high confidence

permissions. The algorithm which is used for detecting permissions is Apriori [20], which is one of the famous algorithms in association rules. With 96.5% confidence and 10% minimum support, 3 permission were found which can be removed due to its relative to another permission. Finally, 22 permissions were left as the most important permissions as shown in Table I.

- Machine Learning on Significant Permission

For the detection, a supervised machine learning method was used. Support Vector Machine (SVM) was applied to small data to test the Multi-Level Data Pruning model. SVM needs two classes from the training dataset to differentiate its hyperplane. One class is map to the malicious and the other class to the benign while the input data would be to unknown application. The application is then mapped to vector space to decide whether it is benign or malicious.

To check the applicability of MLDP, 67 machine learning techniques were used. The results of the learning algorithm over the original dataset were compared and applied algorithms on the MLDP dataset. Also, tree base techniques like decision tree have better result but due to high features among the data, this method consumed a lot of memory and thus greatly reduce the accuracy of the results.

Functional tree (FT) and Random Forest are more likely to have better Recall than the other algorithms. However, Random Forest consumes more memory and time to analyse data. One of the main issues with this algorithm is that while the number of datasets increasing, FPR and other performance metrics are slightly decreasing. Table II shows that as we moved from 2650 to 54694 applications, the false positive rate increases to 4.85%, and this is really high.

E. Comparison between Different Techniques

In this section, we conclude that each method has its own advantages and disadvantages. These pros and cons can be shown in terms of the accuracy of detection or overhead on computational resources. Therefore, there is no specific best method for investigation in the area of Android. Fig. 2 illustrates the comparison between different techniques in terms of accuracy.

TABLE I. MOST SIGNIFICANT PERMISSION BY MLDP

MLDP	
22 Permissions	
ACCESS_WIFI_STAT CAMERA CHANGE_NETWORK_STATE CHANGE_WIFE_STATE DISABLE_KEYGUARD GET_TASKS INSTALL_PACKAGES READ_CALL_LOG READ_CONTACTS READ_EXTRERNAL_STORAGE READ_HISTORY_BOOKMARKS	READ_LOGS READ_PHONE_STATE READ_SMS RECEIVE_BOOT_COMPLETE RESTART_PACKAGES SEND_SMS SET_WALLPAPER SYSTEM_ALRERT_WINDOW WRITE_APN_SETTING WRITE_CONTACTS WRITE_SETTINGS

TABLE II. RESULTS FOR DIFFERENT DATASETS

Num_of_Malapp	2650	5494	54694
Precision	98.83%	97.54%	95.15%
Recall	94.4%	93.62%	92.17%
FPR	1.17%	2.36%	4.85%
FM	94.97%	95.54%	93.63%
ACC	96.47%	95.63%	93.67%

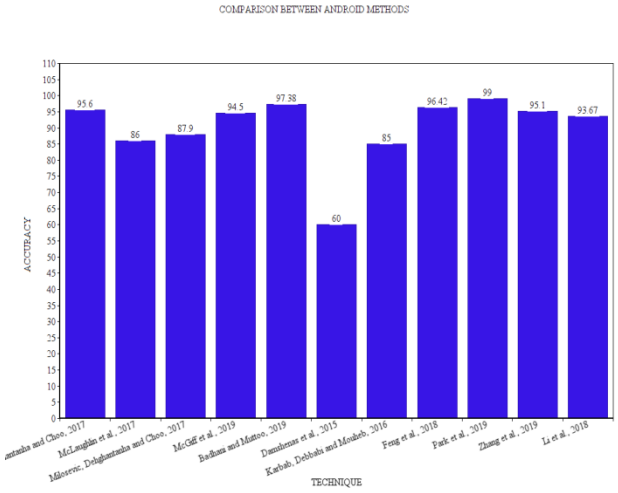


Fig. 2. Comparison between the different Methods in Android Malware Detection.

III. METHODOLOGY

Here, we discussed the simulation environments, data collection, and research methodology.

A. Simulation Environment

In this work, Python 3.6.1 is used as the base scripting language. Python provides a wide range of libraries and functions. Additionally, there are a lot of Android packages analysis tools that are written in Python. Operating systems vary during different phases of analysis. Windows 10 is our base operating system in this analysis but however in some cases, Linux CentOS 6.8 is used.

The data extraction platform needs strong computational power for the early stage of the study. Because of the high number of input data set, it is good to use a high-frequency CPU to extract specific data from APK files. In our work, we used two different platforms. First, to only inspect a small amount of test data in making the machine learning model, we run it on Intel i7-7500U, 2.70 GHz, 16 GB RAM. Second, most of the computation is conducted on Intel Xenon L5640 2.70 GHz, 10 GB RAM.

B. Datasets Collection

The scope of Android malware detection only focuses on real applications. These applications can be collected from different common markets like Google play store. Over the years, researchers collected some labelled datasets to ease their researches. Some of these datasets are kept confidential with related to specific projects and accessing them requires special permission. But some of these datasets are accessible for researchers.

In this research, M0droid project dataset [21] with overall 410 labelled applications is used. This dataset contains 200 malware applications and 210 normal applications. The other dataset used was collected from different projects in the University of Brunswick [22]. This dataset is highly categorized and labelled. Number of applications in this dataset is over 5000 and Android packages with 2000 normal applications and 3000 malware applications.

Also, the AndroZoo dataset [23] has granted permission to this research to use their dataset. This dataset is made up of about 8,000,000 applications. Android PRAGuard Dataset [24] is another dataset that was used with almost 10479 applications. Most of the applications are labelled. Android Malware Dataset (AMD) [6], is one of the largest malware datasets in Android. It contains 24553 applications that are grouped into 135 different groups. Up to 71 malware families have been seen in this dataset.

C. Android Packaging File (APK)

Application in the Android platform is delivered by APK file format. The focus of this study is on the APK file. Fig. 3 describes the architecture of the APK file. The *META-INF* includes manifest file, lib contain compiled code for any processor the device is using, res contains the non-compiled file, *assests* include application asset, *AndroidManifest.xm*l is used for the identification of applications and *classes.dex* contains the source code in *dex* format [25]. *Dex* files are only usable by *Dalvik* or *ART* compiler on Android devices. The APK file is compressed with *ZIP* format so we can get the desired file from unzipping the package. Different tools can be used to extract desire information.

D. Android Permission Structure

The proposed framework stands on Android permission. Permission is designed to add another level of security to protect user sensitive data such as SMS or contacts. In Android security architecture, no application by default has access to any action related to the operating system, other applications, or user private files. Permissions with low risk are automatically granted to the application. If application request high risk permission such as CAMERA, user acceptance is required to grant permission.

Dangerous permission can be asked in two methods:

1) Install time request: the system asks to grant all requested permission.

2) Runtime request: whenever services needed by an application, request permission from the user.

Permissions are highlighted by '*uses-permission*' in the Android manifest file. Some permissions are related to specific hardware feature and therefore, developer needs to mention '*uses-feature*' in the manifest file. This option helps the application runs on devices without the requested hardware. Google defined three protection levels for permission: *normal, signature,* and *dangerous.* These are discussed below.

- Normal Permission

Normal permission is a permission that is requested by application for accessing data or resources out of its box. Mostly, these permissions are granted by the system when the application is installed. According to Google on Android 9, the following as shown in Table III are normal permissions:

- Signature Permission

The system grants these permissions but the permissions need to be signed by the same application that requests permission. Table IV contains the list of signature permissions.

- Dangerous Permission

Dangerous permission is the type of permission that is requested to access user's private data or system files. When permission is declared as dangerous, user approval is necessary for allocating requested resources or data. Table V shows the dangerous permissions.

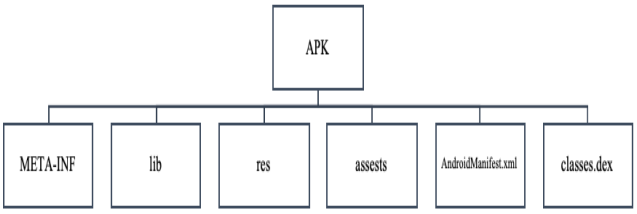


Fig. 3. APK File Structure.

TABLE III. NORMAL PERMISSION

ACCESS_LOCATION_EXTRA_COMMANDS	ACCESS_NETWORK_STATE	ACCESS_NOTIFICATION_POLICY	ACCESS_WIFI_STATE	BLUETOOTH	BLUETOOTH_ADMIN
BROADCAST_STICKY	CHANGE_NETWORK_STATE	CHANGE_WIFI_STATE	CHANGE_WIFI_MULTICAST_STATE	DISABLE_KEYGUARD	EXPAND_STATUS_BAR
FOREGROUND_SERVICE	GET_PACKAGE_SIZE	INSTALL_SHORTCUT	INTERNET	KILL_BACKGROUND_PROCESSES	MANAGE_OWN_CALLS
MODIFY_AUDIO_SETTINGS	NFC	READ_SYNC_SETTINGS	READ_SYNC_STATS	RECEIVE_BOOT_COMPLETED	REORDER_TASKS
REQUEST_COMPANION_RUN_IN_BACKGROUND	REQUEST_COMPANION_USE_DATA_IN_BACKGROUND	REQUEST_DELETE_PACKAGES	REQUEST_IGNORE_BATTERY_OPTIMIZATIONS	SET_ALARM	SET_WALLPAPER
SET_WALLPAPER_HINTS	TRANSMIT_IR	USE_FINGERPRINT	VIBRATE	WAKE_LOCK	WRITE_SYNC_SETTINGS

TABLE IV. SIGNATURE PERMISSION

BIND_ACCESSIBILITY_SERVICE	BIND_AUTOFILL_SERVICE	BIND_CARRIER_SERVICES	BIND_CHOOSER_TARGET_SERVICE	BIND_CONDITION_PROVIDER_SERVICE
BIND_DEVICE_ADMIN	BIND_DREAM_SERVICE	BIND_INCALL_SERVICE	BIND_INPUT_METHOD	BIND_MIDI_DEVICE_SERVICE
BIND_NFC_SERVICE	BIND_NOTIFICATION_LISTENER_SERVICE	BIND_SCREENING_SERVICE	BIND_TELECOM_CONNECTION_SERVICE	BIND_TEXT_SERVICE
BIND_TV_INPUT	BIND_VISUAL_VOICEMAIL_SERVICE	BIND_VOICE_INTERACTION	BIND_VPN_SERVICE	BIND_VR_LISTENER_SERVICE
BIND_WALLPAPER	CLEAR_APP_CACHE	MANAGE_DOCUMENTS	READ_VOICEMAIL	REQUEST_INSTALL_PACKAGES
SYSTEM_ALERT_WINDOW	WRITE_SETTINGS	WRITE_VOICEMAIL	*	*

TABLE V. DANGEROUS PERMISSION

CALENDAR	READ_CALENDAR	LOCATION	ACCESS_FINE_LOCATION	SMS	SEND_SMS
	WRITE_CALENDAR		ACCESS_COARSE_LOCATION		RECEIVE_SMS
CALL_LOG	READ_CALL_LOG	MICROPHONE	RECORD_AUDIO		READ_SMS
	WRITE_CALL_LOG	PHONE	READ_PHONE_STATE		RECEIVE_WAP_PUSH
	PROCESS_OUTGOING_CALLS		READ_PHONE_NUMBERS	STORAGE	READ_EXTERNAL_STORAGE
CAMERA	CAMERA		CALL_PHONE		WRITE_EXTERNAL_STORAGE
CONTACTS	READ_CONTACTS		ANSWER_PHONE_CALLS	*	*
	WRITE_CONTACTS		ADD_VOICEMAIL	*	*
	GET_ACCOUNTS		USE_SIP	*	*
	RECEIVE_MMS	SENSORS	BODY_SENSORS	*	*

IV. PROPOSED FRAMEWORK

The proposed framework contains different layers for data processing. It is called a multi-level permission extraction framework. The multi-level permission extraction framework contains methods from the previous researches as well as a novel method to improve the accuracy of the detection model. The general overview of the proposed framework is shown in Fig. 4.

A. Permission Extraction from APK Files

The first step in our framework is to extract permissions from APK files. As described previously, permissions exist in the Android manifest file. To read and extract desire information, specific tools are required which is discussed in the experimental section. After extracting permissions from both malware and benign datasets, data are mapped into a matrix. This matrix is called a feature matrix.

B. Feature Matrix

The feature matrix represents the existence of permission in the applications. In the feature matrix, each application's permissions are defined by Boolean variables. We defined two matrices, M and B. Matrix M is the list of permission used by the malicious application and matrix B is the list of permission used by the benign application. M_{ij} says whether permission j^{th} is used by the i^{th} malicious application or not. If the answer is yes then 1 otherwise 0. The same goes to B_{ij} for benign applications. Table VI is the sample matrix from the malware dataset.

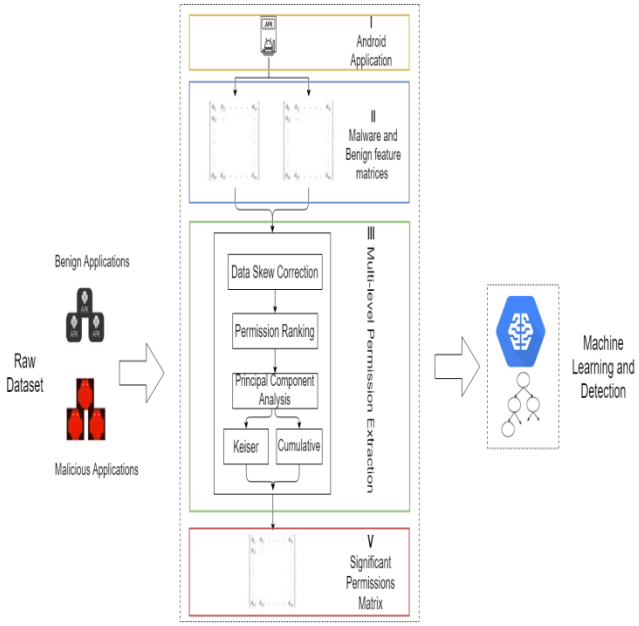


Fig. 4. Proposed Multi-Level Permission Extraction Framework.

C. Multi-Level Permission Extraction

Multi-level permission extraction level is the proposed framework. In this framework, different levels of permission pruning are applied to achieve the most significant permissions. The proposed framework contains four main phases; data skew correction, permission ranking, principal component analysis, and two different statistical algorithms

for identifying the most valuable permissions in the dataset. This framework hired some techniques from one of the states of arts research in Android malware detection [10].

- Data Skew Correction

In data analysis research, it is necessary that data in every category balanced to avoid any wrong or incorrect results. However, providing the same amount of data for every category is not possible. In this research, we used different sources for our datasets. The malware dataset size is around 25000 APK files as compare to 15000 APK files for the normal datasets. Also, we have around 8,000,000 APK files in both malware and benign dataset. This unbalances in malware and normal data category may lead to inaccurate results in the final steps.

In data skew correction layer, we used equation 3 to find support for each permission in the larger datasets and then scale down that permission to match smaller dataset. In this case, the number of malicious applications is more than the benign applications.

$$S_M(P_j) = \frac{\sum_i M_{ij}}{size(M_j)} \times size(B_j) \quad (3)$$

Where P_j describes j^{th} permission and $S_M(P_j)$ is the support of j^{th} permission in the malware matrix.

- Permission Ranking

In this step, permission ranking scheme has been used to hold the most frequent permissions in the dataset. Some permissions only happened in the context a few times, and hence these permissions have less support as compare to the common permissions. This scheme provides a more accurate view on what each permission represents in the datasets.

Term-Frequency Inverse Document Frequency (TF-IDF) technique is used to indicate the importance of each permission in all of the datasets. Term weighting methods are commonly grouped into supervised and unsupervised methods. The unsupervised or traditional term weighting methods are originated from the information retrieval field. The supervised weighting methods used the prior information of the training documents in predefined categories. The weight of a feature f with respect to a class represent the discriminating ability of f towards normal and attack classes. The higher the weight, the stronger the discriminating power of this feature in identifying the anomaly instances. Because *permission-based Vector* is by nature a bag of permission in the form of a vector, the *tf-idf* weighting method, where, $tf - idf$ is the weight of the feature f_i with regards to permission s_j is the product of $tf_{i,j}$ and $idf_{i,j}$ [26] and this is show mathematically in equation 4.

$$w_{f_i,s_j} = tf_{i,j}.idf_{i,j} \quad (4)$$

The matrices from the previous layer are passed into this layer and the value of all the rows is summed in one row. The new row represents how many times one permission is requested in a specific dataset. Then, the TF-IDF is applied to determine the impact of permission in the dataset. The weighting scheme result will omit the less frequent permissions and show the significant permissions in both datasets.

- Principal Component Analysis

Data transformation is required to come up with an efficient method to analyse a large amount of data. This method helps to reduce the dimension of the feature matrix to increase the efficiency of the model. Principal component analysis (PCA) is a statistical model which reduce large feature sets into smaller one while keeping most of the information intact. It transforms correlated features into some uncorrelated features called the principal component. This method is closed to a correlation technique that is applied to data with a wide difference of variance. PCA algorithm is mainly used for the algorithm with common share variance and focuses on a linear combination of a variable to extract maximum possible variance. In the PCA algorithm, eigenvector and eigenvalue are considered. The eigenvector shows a common variance and unique variance for producing correlation. Eigenvalue is the measure of all variance for a specific factor.

- Kaiser and Cumulative

The last layer of the proposed framework is to select the most significant permissions based on the eigenvalues and eigenvectors variables. Factor analysis is used in this step. Factor analysis is a statistical model to check variability between observed and unobserved factors. Kaiser's method sets a threshold between the maximum and minimum margin for the eigenvalue of 1. It means each eigenvalue above 1 demonstrates the desired factor and those below are not selected [27]. In this study, we applied Kaiser's method on the PCA results to find the most significant permissions.

Besides Kaiser, the cumulative technique was also used. Unlike Kaiser, there is no specific threshold for cumulative technique. Rather than a threshold, a percentage has been suggested. In most of the existing models, when 90% of the variance is reached, the model stopped. In our case, we also set the variance percentage to be 90% and applied the cumulative method on the PCA results to distinguish the most important permissions [28].

We compared Kaiser and cumulative results to obtain similar permission from the list. Similarly, this comparison has been done with Li [10] and Google dangerous permission list. The Final selected permissions are obtained from the intersection of all the results.

- Significant Permission Matrix

In the final step of the data processing, the desired feature matrix with the most significant permissions is formed. Also, the data label is added to the model to be used for classification in the machine learning and detection part.

V. Experimental Setting and Results

This section summarily discussed how the data are extracted and processed, the implementation that is done using machine learning approaches and the discussion on the evaluation of the proposed framework using some standard metrics.

A. Data Extraction

Permissions exist in the Android manifest file. To access permissions, extracting Android packages is required. For APK files in *ZIP* format, accessing them directly is not possible. There are two main methods to access the manifest file. First, extracting files manually and decoding AndroidManifest.xml file, and second, using a tool to extract them. In this work, the *Androguard* tool has been used. *Androguard* is a tool that extracts details from APK files [29]. When using this tool, we only need the permission extraction part. Thus, we extract all the required permissions and mapped them into the desired feature matrix to begin the data processing phase. The feature matrix store permissions as columns and applications as a row. The sample of the extracted data from the datasets is displayed in Table VIa.

B. Data Processing

Python is the main programming language used for the implementation. Scikit learn library is used for the PCA, Kaiser, and cumulative layers of our framework. Table VII shows the sample correlation among features. The main diagonal is 1 because the relation of each permission with itself is 1. Based on this matrix, each correlation variable is higher declared and these two permissions have much impact on the datasets.

C. Machine Learning

We used Scikit learn library to implement the malware detection model using SVM and decision tree algorithms. We used 80% of the datasets to train the model and 20% for the testing.

D. Experimental Results

We report the selected permission based on the multi-level permission extraction framework and its effectiveness on large datasets. To align with the main objective of this research, we are able to achieve a better optimized solution for large datasets as compared to the existing works.

- Significant Dangerous Permission

From our results, 16 permissions were highlighted as dangerous permissions. Table VI shows the list of high-risk permissions based on the proposed framework.

- Evaluation Metrics

The evaluation metrics considered in this research are precision, recall, and F-measure. Precision is the number of correct positive results by the number of positive predicted results. A recall is the number of correct positive results by the number of all datasets. F-measure is the balance between precision and recall rate. This rate tries to determine the accuracy of the classification. Mathematically, these metrics are defined in equations 5, 7, and 8.

$$Precision = \frac{True\ Positve\ Rate}{True\ Positive\ Rate+False\ Positive\ Rate} \quad (5)$$

$$Recall = \frac{True\ Positive\ Rate}{True\ Positve\ Rate+False\ Negative\ Rate} \quad (6)$$

$$F-Measure = \frac{2*Precision*Recall}{Precission+Recall} \quad (7)$$

Tables VIII and IX show the measurement metrics for our framework based on 20 selected permissions. From the tables, the input applications considered are 10000, 25000, and 60000. Although, the available datasets are higher than the number of applications, but we are able to get our desire threshold at 60000 based on the results.

Table VIII shows the performance metrics based on the SVM and Table IX contains the results based on the decision tree algorithm.

TABLE VI. DANGEROUS PERMISSION

CAMERA	RECEIVE_BOOT_COMPLETE
READ_CALL_LOG	ANSWER_PHONE_CALL
READ_CONTACTS	RECORD_AUDIO
READ_EXTERNAL_STORAGE	WRITE_SETTING
READ_PHONE_STAT	WRITE_EXTERNAL_STORAGE
READ_SMS	CHANGE_WIFI_STAT
WRITE_CONTACT	DISABLE_KEYGUARD
WRITE_SMS	SET_WALLPAPER

TABLE VIa : SAMPLE FEATURE MATRIX

	ACCESS	ACCESSORY	ACTIVITY	ANT	AUTHENTICATE	BATTERY	BIND	BLUETOOTH	BODY	BROADCAST	CALL	CAMERA	CHANGE	CLEAR	CONTENT	DEVICE	DISABLE	DOWNLOAD	EXPAND	FLASHLIGHT	GET	GOOGLE	INSTALL	INTERACT	INTERNAL	INTERNET	KILL	MANAGE	MAPS	MODIFY	MOUNT	NFC	PACKAGE	PERSISTENT	PROCESS	READ	RECEIVE	RECORD	REGISTER	REGISTRATION	REORDER	RESTART	SEND	SERVICE	SET	STORAGE	SUBSCRIBED	SYSTEM	TASK	UA	UI	UNINSTALL	UPDATE	USE	VIBRATE	WAKE	WRITE
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1
2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
4	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	1
5	1	1	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
7	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
8	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1
9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	1
11	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
13	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
14	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1
15	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

TABLE VII. CORRELATION BETWEEN PERMISSIONS

	ACCESSORY_FRAMEWORK	ACCESS_COARSE_LOCATION	ACCESS_FINE_LOCATION	ACCESS_KEY_APP	ACCESS_LOCATION_EXTRA_COMMANDS	ACCESS_MOCK_LOCATION	ACCESS_NETWORK_STATE	ACCESS_WEB_PAGES
ACCESSORY_FRAMEWORK	1	0.099614	0.097472	-0.00485	-0.01303	-0.00688	0.019474	-0.00485
ACCESS_COARSE_LOCATION	0.099614	1	0.717591	-0.04873	0.267477	0.141218	0.195498	-0.04873
ACCESS_FINE_LOCATION	0.097472	0.717591	1	-0.0498	0.205232	0.138181	0.199795	-0.0498
ACCESS_KEY_APP	-0.00485	-0.04873	-0.0498	1	-0.01303	-0.00688	0.019474	-0.00485
ACCESS_LOCATION_EXTRA_COMMANDS	-0.01303	0.267477	0.205232	-0.01303	1	-0.01848	0.052291	-0.01303
ACCESS_MOCK_LOCATION	-0.00688	0.141218	0.138181	-0.00688	-0.01848	1	0.027608	-0.00688
ACCESS_NETWORK_STATE	0.019474	0.195498	0.199795	0.019474	0.052291	0.027608	1	0.019474
ACCESS_WEB_PAGES	-0.00485	-0.04873	-0.0498	-0.00485	-0.01303	-0.00688	0.019474	1

TABLE VIII. PERFORMANCE METRICS BASED ON SVM

Number of applications	**10000**	**25000**	**60000**
Precision	98.20%	97.16%	95.17%
Recall	95.80%	93.75%	92.86 %
F-measure	96.98%	95.42%	94.00%

TABLE IX. PERFORMANCE METRICS BASED ON DECISION TREE

Number of applications	10000	25000	60000
Precision	98.99%	96.10%	92.11%
Recall	96.10%	93.20%	91.10 %
F-measure	97.53%	94.68%	91.60%

From the above tables, we say that the SVM performs better than the decision tree in all the measurement metrics. SVM shows promising results in false positive rate of 3.89 with 60000 applications. This is also shown in Fig. 5.

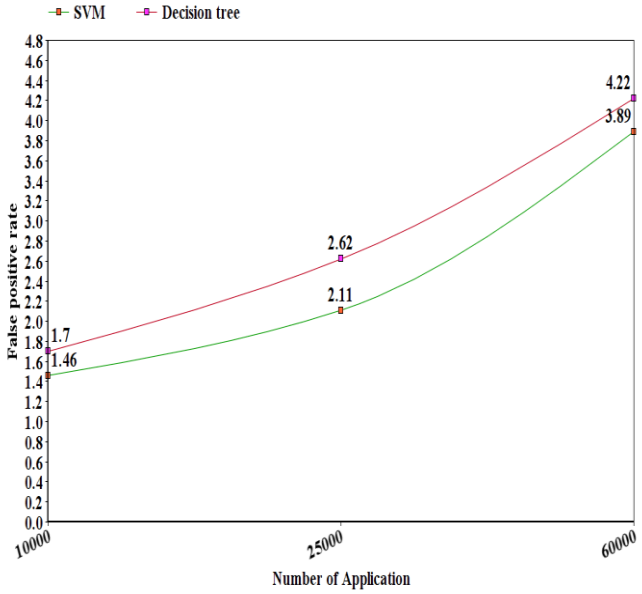


Fig. 5. False Positive Rate in SVM and Decision Tree.

- Discussion

Android malware detection is a crucial research area for the current digital world. Most of the handphones are running on Android. Due to the availability and wide usage of this platform, many attacks are carried out to take advantage of the user's personal data. Permission in Android is the main source of investigation on developer intent. Permission can be named as double edge blade which can be dangerous or helpful for users. In most cases users need to decide whether permission is granted or not. This decision is crucial when it comes to a wide range of applications on mobile phones. Google as the main developer of Android tried to facilitate this decision by adding extra labels to some permissions, however this barrier is still not enough.

The proposed framework successfully increased the detection rate by using only permissions. Although some researchers like [30] used an ensemble of features for detection but this method is complex and inefficient in a real-world implementation. In the work of [1], permission is used as the only feature for identification but their results show that as the data increases, the detection accuracy keep reducing. To maintain the detection accuracy when using a large number of datasets and one feature, we used different factor analysis methods such as Kaiser and cumulative to aggregate the significant permissions. Our results are compared with the results of [1] in Table X.

TABLE X. COMPARISON BETWEEN LI FRAMEWORK AND THE PROPOSED FRAMEWORK

Performance metrics	**Li Framework**		**Proposed Framework**	
	No. of Applications	Results	No. of Applications	Results
Precision	2650	98.83%	10000	98.20%
	5494	97.54%	25000	97.16%
	54694	95.5%	60000	95.17%
Recall	2650	94.4%	10000	95.80%
	5494	93.62%	25000	93.75%
	54694	92.17%	60000	92.86%
F-measure	2650	94.97%	10000	96.98%
	5494	95.54%	25000	95.42%
	54694	93.63%	60000	94.00%
False positive rate	2650	1.17%	10000	1.46%
	5494	2.36%	25000	2.11%
	54694	4.85%	60000	3.89%

During the analysis, some problems were faced. Some applications do not follow normal patterns. This anomaly can be code obfuscation or APK encryption. Some attackers encode the Android manifest file which makes it inaccessible for analysis. In this case, analysing the pattern of application in terms of network connection or requested permission while the application is running can help to distinguish malicious applications. But in most cases, this analysis is costly in terms of resource and time.

VI. Conclusion

Due to high demand and availability, the Android has become a famous platform for malicious activity. There are existing algorithms developed to avoid malware attacks on this platform but these algorithms are inefficient. Static, dynamic, and hybrid analysis are the three main techniques used to investigate malicious applications. The authors proposed a multi-level permission extraction framework to identify significant permissions to differentiate between normal and malicious applications in Android devices. The method used is based on static analysis as the researchers' focus is on the Android packaging file (APK) in a static environment. Permission was used as a feature to develop the proposed model and the model is able to achieve a better detection accuracy as compared to the existing works. To prune out unnecessary permissions, the researchers employed different mathematical steps in the proposed framework. The SVM and decision tree algorithms were used for the classification with different number of datasets. The results obtained are promising as with 60,000 applications, the model achieved 94% accuracy. This result is better when compared to other existing models. Hybrid method (a combination of static and dynamic permission analysis) would be adopted to see if better results can be achieved. Adding more features from applications and increasing the datasets will also be considered in our future work.

Acknowledgment

The authors would like to acknowledge the Universiti Putra Malaysia for supporting this research.

References

[1] J. Li, L. Sun, Q. Yan, Z. Li, W. Srisa-An, and H. Ye, "Significant Permission Identification for Machine-Learning-Based Android Malware Detection," IEEE Trans. Ind. Informatics, Vol. 14, No. 7, Pp. 3216–3225, 2018, Doi: 10.1109/Tii.2017.2789219.

[2] J. Walls and K.-K. R. Choo, "A Review of Free Cloud-Based Anti-Malware Apps for Android," in 2015 IEEE Trustcom/BigDataSE/ISPA, 2015, pp. 1053–1058, doi: 10.1109/Trustcom.2015.482.

[3] Roman Unuchek, "Mobile malware evolution 2017," 2018.

[4] L. Dua and D. Bansal, "TAXONOMY: MOBILE MALWARE THREATS AND DETECTION TECHNIQUES," pp. 213–221, 2014, doi: 10.5121/csit.2014.4522.

[5] R. Zachariah, M. S. Yousef, and A. M. Chacko, "Android Malware Detection A Survey," no. Iccs, 2017.

[6] F. Wei, Y. Li, S. Roy, X. Ou, and W. Zhou, "Deep Ground Truth Analysis of Current Android Malware," Springer, Cham, 2017, pp. 252–276.

[7] E. Alpaydin, Introduction to machine learning. MIT Press, 2010.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[9] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and Machine Learning forecasting methods: Concerns and ways forward," 2018, doi: 10.1371/journal.pone.0194889.

[10] J. Li, L. Sun, Q. Yan, Z. Li, W. Srisa-an, and H. Ye, "Significant Permission Identification for Machine Learning Based Android Malware Detection," IEEE Trans. Ind. Informatics, vol. 3203, no. c, 2018, doi: 10.1109/TII.2017.2789219.

[11] N. Milosevic, A. Dehghantanha, and K. K. R. Choo, "Machine learning aided Android malware classification," Comput. Electr. Eng., vol. 61, pp. 266–274, 2017, doi: 10.1016/j.compeleceng.2017.02.013.

[12] N. McLaughlin et al., "Deep Android Malware Detection," Proc. Seventh ACM Conf. Data Appl. Secur. Priv. - CODASPY '17, pp. 301–308, 2017, doi: 10.1145/3029806.3029823.

[13] J. Mcgiff, W. G. Hatcher, J. Nguyen, W. Yu, E. Blasch, and C. Lu, "Towards Multimodal Learning for Android Malware Detection," 2019 Int. Conf. Comput. Netw. Commun. ICNC 2019, pp. 432–436, 2019, doi: 10.1109/ICCNC.2019.8685502.

[14] S. Badhani and S. K. Muttoo, "CENDroid—A cluster–ensemble classifier for detecting malicious Android applications," Comput. Secur., Apr. 2019, doi: 10.1016/J.COSE.2019.04.004.

[15] K.-K. R. C. &Ramlan M. Mohsen Damshenas, Ali Dehghantanha, "No Title," J. Inf. Priv. Secur., vol. 11, no. 3, pp. 141–157, 2015.

[16] E. M. B. Karbab, M. Debbabi, and D. Mouheb, "Fingerprinting android packaging: Generating DNAs for malware detection," DFRWS 2016 USA - Proc. 16th Annu. USA Digit. Forensics Res. Conf., vol. 18, pp. S33–S45, 2016, doi: 10.1016/j.diin.2016.04.013.

[17] P. Feng, J. Ma, C. Sun, X. Xu, and Y. Ma, "A novel dynamic android malware detection system with ensemble learning," IEEE Access, vol. 6, pp. 30996–31011, 2018, doi: 10.1109/ACCESS.2018.2844349.

[18] M. Park, J. Han, H. Oh, and K. Lee, "Threat Assessment for Android Environment with Connectivity to IoT Devices from the Perspective of Situational Awareness," Wirel. Commun. Mob. Comput., vol. 2019, pp. 1–14, Apr. 2019, doi: 10.1155/2019/5121054.

[19] J. Zhang, Z. Qin, H. Yin, L. Ou, and K. Zhang, "A feature-hybrid malware variants detection using CNN based opcode embedding and BPNN based API embedding," Comput. Secur., vol. 84, pp. 376–392, 2019, doi: 10.1016/j.cose.2019.04.005.

[20] R. Agrawal and R. S&ant, "Fast Algorithms for Mining Association Rules," 20th VLDB Conf. Santiago, Chile, 1994.

[21] M. Damshenas, A. Dehghantanha, K.-K. R. Choo, and R. Mahmud, "M0Droid: An Android Behavioral-Based Malware Detection Model," J. Inf. Priv. Secur., vol. 11, no. 3, pp. 141–157, 2015, doi: 10.1080/15536548.2015.1073510.

[22] A. Mahindru and P. Singh, "Dynamic Permissions based Android Malware Detection using Machine Learning Techniques," in Proceedings of the 10th Innovations in Software Engineering Conference on - ISEC '17, 2017, pp. 202–210, doi: 10.1145/3021460.3021485.

[23] K. Allix, T. F. Bissyandé, J. Klein, and Y. Le Traon, "AndroZoo," in Proceedings of the 13th International Workshop on Mining Software Repositories - MSR '16, 2016, pp. 468–471, doi: 10.1145/2901739.2903508.

[24] D. Maiorca, D. Ariu, I. Corona, M. Aresu, and G. Giacinto, "Stealth attacks: An extended insight into the obfuscation effects on Android malware," Comput. Secur., vol. 51, pp. 16–31, Jun. 2015, doi: 10.1016/j.cose.2015.02.007.

[25] Google, "The Structure of Android Package (APK) Files," Nov. 2010.

[26] R. Banchs, Text Mining with MATLAB. New York, NY: Springer New York, 2013.

[27] J. Ruscio and B. Roche, “Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure.,” Psychol. Assess., vol. 24, no. 2, pp. 282–292, Jun. 2012, doi: 10.1037/a0025697.

[28] B. Williams, T. Brown Andrys Onsman, A. Onsman, T. Brown, P. Andrys Onsman, and P. Ted Brown, “Exploratory factor analysis: A five-step guide for novices Recommended Citation) "Exploratory factor analysis: A five-step guide for novices Exploratory factor analysis: A five-step guide for novices,” This J. Artic. is posted Res. Online, vol. 8, pp. 2010–990399, 2012.

[29] S. H. Anthony Desnos, “Androguard,” 2011.

[30] W. Wang, Z. Gao, M. Zhao, Y. Li, J. Liu, and X. Zhang, “DroidEnsemble: Detecting Android Malicious Applications with Ensemble of String and Structural Static Features,” IEEE Access, pp. 1–1, 2018, doi: 10.1109/ACCESS.2018.2835654.

Performance Impact of Genetic Operators in a Hybrid GA-KNN Algorithm

Raghad Sehly[1], Mohammad Mezher[2]
Department of Computer Science
Fahad Bin Sultan University
Tabuk, KSA

***Abstract*—Diabetes is a chronic disease caused by a deficiency of insulin that is prevalent around the world. Although doctors diagnose diabetes by testing glucose levels in the blood, they cannot determine whether a person is diabetic on this basis alone. Classification algorithms are an immensely helpful approach to accurately predicting diabetes. Merging two algorithms like the K-Nearest Neighbor (K-NN) Algorithm and the Genetic Algorithm (GA) can enhance prediction even more. Choosing an optimal ratio of crossover and mutation is one of the common obstacles faced by GA researchers. This paper proposes a model that combines K-NN and GA with Adaptive Parameter Control to help medical practitioners confirm their diagnosis of diabetes in patients. The UCI Pima Indian Diabetes Dataset is deployed on the Anaconda python platform. The mean accuracy of the proposed model is 0.84102, which is 1% better than the best result in the literature review.**

***Keywords*—Data mining; classification; K-NN; GA; Pima Indian Diabetes Dataset; UCI**

I. Introduction

The world is facing many prevalent and chronic diseases. Diabetes is one of them. According to statistics for 2019 from the International Diabetes Federation (IDF), about 463 million adults (20-79 years old) were suffering from diabetes that year [1]. Many methods exist for diagnosing diabetes. Doctors diagnose diabetes by measuring glucose levels in the blood using tests such as the Fasting Plasma Glucose Test (FPG), the Postprandial Glucose Test, the Random Blood Glucose Test, the Oral Glucose Tolerance Test (OGTT), and the Glycated Hemoglobin Test [2]. One measurement of blood sugar is not enough to diagnose diabetes, particularly if the patient is diabetic and has no other symptoms. In differential diagnosis, doctors must examine the medical records of previous patients with the same conditions. Data were taken from patients and experts is the most important factor [3]. If these data are classified and predicted in a precise way, the global health expenditure can be reduced by up to 10% (760 billion USD) [1]. Data mining is a logical process for discovering and predicting models with huge data sets to find useful information. Data mining has three steps: exploration, pattern identification, and deployment [4].

The first step of data exploration involves cleaning and transforming data. The second step involves determining the important variables and the nature of the data, depending on the problem. After the data are explored, refined, and defined for particular variables, it is necessary to form a pattern of identification. Identifying and choosing the patterns can contribute to enhancing prediction. Deploying patterns is implemented for the desired outcome.

However, classifying the dataset is one of the most popular techniques in data mining. It employs a set of pre-classified examples for developing a model that can classify the records population in general. The classification has many different models, such as the K-Nearest Neighbor (KNN), the Decision Tree Induction, Neural Networks, the Bayesian Classification, Support Vector Machines (SVM), and Classification Based on Associations. With the help of classification algorithms, diabetes can be diagnosed more accurately [4].

A. K-NN Classifier

KNN is a technique used to classify objects depending on the closest training examples in the space of feature. The K-NN classifier is a kind of instance-based learning where the function is convergent in a local way, and the computation is postponed until the classification is finished. In K-NN, the training samples are generally described by n-dimensional numeric attributes. The training samples are saved in an N-dimensional space. K-NN [5, 6] begins searching for the "K" training samples that are closest to the test sample or unknown sample when a test sample is given. Closeness is defined in terms of Euclidean distance. Some distance measures are used to determine closeness. Occasionally, one minus correlation value is taken as a distance metric. The following three distance measures are used for continuous variables: Euclidean distance measure, Manhattan distance measure, and Minkowski distance measure, the latter of which has been used in this paper. The Hamming distance should be used in the instance of categorical variables. The distance measures between x and y points, (x_1, x_2… x_n) and y (y_1, y_2… y_n), are calculated in below distance functions:

$$Euclidean = \sum_{i=1}^{k} (x_i - y_i)^2$$

$$Manhattan = \sum_{i=1}^{k} |x_i - y_i|$$

$$Minkowski = \left(\sum_{i=1}^{k} (|x_i - y_i|)^q \right)^{1/q}$$

The pseudocode of K-NN:

K-NN Classifier (X, Y, x) :

// where X: data, Y: decision labels

// x: test sample

for i = 1 to m **do** # i: data point

k-distance = compute the

Sort the k distances.

distance d (X_i, x)

end for

return k-nearest reference point for {Y_i where i ϵ I_m}

B. Genetic Algorithm (GA)

A programming method inspired by biological evolution; GA is used in problem-solving strategy to identify optimal solutions. Given that GA is a general algorithm, it works appropriately with any search space. GA is a useful tool for classifying the K-NN algorithm. Although the traditional K-NN algorithm employs Euclidian distance regularly, other measures can be used as well. GA can improve the performance of K-NN by using both Euclidian distance and cosine similarity to evaluate the optimal linear weights of features [7]. In search of an optimal reference point, GA is capable of optimizing additive measures to evaluate similarity using cosine measures. In [7], GA used the selection principles and evolution to find potential solutions to a given problem as shown in the pseudocode below:

1) Set a group of chromosomes (any applicable representation).
2) Generate an initial population out of (1).
3) Evaluate the generated population in (2).
4) Search Space until n generations.
a) Select fittest chromosome.
b) Mutate and recombine chromosomes.
c) Evaluate the generated (b) chromosomes.
d) Replace weak chromosomes with new chromosomes.

II. Literature Review

This section lists and reviews previous works that experiment with K-NN classification and GA algorithms. It also compares the algorithms based on accuracy performance.

Research paper [5] employed the Pima Indian Diabetes Dataset to measure the performance of combined K-NN and GA algorithms. The combination of K-NN and GA successfully improved the feature selection in the K-NN classifier. The accuracy of this model was 83.12%.

The authors of [8] conducted experiments to predict diabetes by using the Pima Indian Diabetes Dataset. For this evaluation, they chose to use the results of the 10-fold cross-validation. The accuracy of the K-NN algorithm was 71.84%.

Author in [9] evaluated the performance of the K-NN algorithm to classify the Pima Indian Diabetes Dataset. The authors used the data imputation, scaling, and normalization techniques to improve the classifier accuracy. A voting classifier was used to measure the performance of the Pima Indian Diabetes Dataset in predicting diabetes. The accuracy of the K-NN algorithm was 71.3%.

Author in [10] aimed to compare the performance of many classification algorithms to predict diabetes with the Pima Indian Diabetes Dataset. The authors compared many machine learning classifiers to classify patients with diabetes. The K-NN algorithm was one of these classifiers, and its accuracy was measured as 72.65% with using of 3-fold cross-validation.

The authors of [11] combined the GA algorithm with the K-NN algorithm to improve the feature selection of the Pima Indian Diabetes Dataset. The achieved accuracy of this model was 73.8%.

Table I shows the accuracy results of different research papers in which the best accuracies were for the hybrid models of GA and KNN.

K-NN and GA real-life applications:

K-NN algorithm: It is applied in many different ways in everyday life, including life, including:

Weather forecasting: The K-NN algorithm had been used to help with rainfall forecasts by using many weather factors, such as mean temperature, dew point temperature, humidity, sea pressure, and wind speed [12].

Economic forecasting: K-NN is a capable technique for economic prediction, especially concerning the financial distress of companies [13].

GA algorithm: It is used in many helpful practical applications, including:

Image Processing: GA is used to solve image segmentation, which is one of the most common problems in the image-processing area [14].

Bioinformatic: GA is a helpful model in interpreting huge data of bioinformatics more accurately and concisely [15].

TABLE I. Algorithms Comparison Results

Research Paper	Algorithm	Accuracy
[5]	K-NN + GA	83.12%
[8]	K-NN	71.84%
[9]	K-NN	71.32%
[10]	K-NN	72.65%
[11]	K-NN + GA	73.78%

III. Model and Implementation

The present researchers implemented the PIMA Dataset to calculate the accuracy of the combined K-NN and GA model. The Pima Indian Diabetes Dataset has eight attributes: Number of Times of Pregnancy (NTP), Plasma Glucose Concentration (PGC), Diastolic Blood Pressure (DBP), Triceps Skinfold Thickness (TSF), Body Mass Index BMI, Diabetes Pedigree Function (DPF), Age in years (AGE), and Binary Target (CLASS). This dataset has been divided into two sets: one for training the model (training set) and another for testing the model (testing set).

A. Data Preprocessing

Quality and quantity are the most effective factors in the classification model's accuracy and prediction. The medical databases are currently facing many obstacles such as noisy, inconsistent, and incomplete data, due to a large amount of data. These obstacles are leading to the low quality of the mining results. Thus, data quality should be upgraded with the assistance of suitable techniques to improve the results of data mining. The data preprocessing technique has a crucial role in improving the data quality and thereby increasing the accuracy of classification. Data preprocessing helps to detect anomalies in the data and remove the data that can lead to big payoffs for decision support. There are many methods of data preprocessing, including data integration, data cleaning, data reduction, data transformations, and data normalization [16].

B. Data Normalization

Data normalization is a process in data preprocessing intended to change the attribute value depending on a common scale or range to improve the performance of the machine learning algorithm. There are many techniques of data normalization, including min-max and z-score. The Python platform has a machine learning framework for data preprocessing such as sklearn. This framework has a large number of useful normalization techniques, including MinMaxScaler (MMS), MaxAbsScaler (MAS), StandardScaler (SS), RobustScaler (RS), and Normalizer (NM) [17]. The technique used in this paper is StandardScaler, which affected the model positively by increasing the accuracy percent.

C. Initial Population

Create a new population by repeating the next steps to completing the new population.

D. Fitness Function

The fitness f(x) of every chromosome x in the population is evaluated. Function fitness(X) is defined as in the equation.

$$Accuracy(x) = fitness(x)$$

E. Selection Operator

Select the chromosomes of the parents from a population depending on their fitness. GA has many different techniques that can be used for selecting the individuals who will be replaced by the next generation. The technique used in this paper is the Elitist selection, which worked effectively. Table II shows the comparisons of GA techniques and the usage of each technique.

TABLE II. Comparison of GA Techniques

GA Technique	Usage
Elitist selection	GA chooses the fittest individuals in every generation.
Fitness-proportionate selection	Chooses fitter members, but not certain, to be selected.
Roulette-wheel selection	Selects the individual depending on fitness level among competitors.
Scaling selection	Distinguishes between individuals with high fitness and those with small differences.
Tournament selection	Chooses the individuals by subgrouping them and then takes only one individual.

F. Crossover Operator

The parents are crossed over with a crossover probability (probability = 0.9) to produce a new offspring. If there is no crossover, the offspring will be an exact duplicate of the parents. In the GA algorithm, the crossover operator is one of the genetic operators often used in the GA lifecycle. After the individuals are selected, the next step is to produce the offspring. Crossover is a commonly used solution for this step, and among its many variant types the single-point crossover is the most used. As shown in Fig. 1, the single-point crossover solution involves selecting a place for locus to replace the remaining alleles between parents.

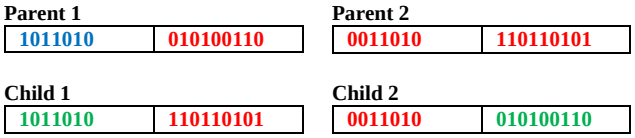


Fig. 1. Crossover Steps.

The children will receive just one section of the chromosome from their parents. The broken point of the chromosome is randomly selected by the crossover point. Because only one crossover point exists, this method is called single-point crossover. In some instances, only child 1 or child 2 is created, but in most cases, both offspring are created and located in the new population. However, the crossover does not occur always. In some cases, based on a probability set, the crossover does not occur, and the parents are copied into the new population directly. The range of the crossover occurrence probability is between 60% and 70%.

F. Mutation Operator

A new offspring will be mutated at every locus with the help of mutation probability (probability = 0.01). The second type of genetic operator used here and considered an exploitation operator is the mutation operator. When selection and mutation operators are used alone, a new individual will be mutated in some of its genes. Some genes are copied directly, and others will be mutated. To guarantee that the individuals are not completely the same, a mutation must be introduced. You loop over all the individuals' genes, and if that gene is chosen for mutation, you can replace it with a small amount, or you can replace it with a new value. The probability range of the mutation is usually less than 0.05. Fig. 2 shows the mutation process.

Before: 1101101001101110
After: 1101100001101110

Fig. 2. Mutation Steps.

The mutation is a simple operator in many ways. One simply must alter the selected alleles depending on what they feel is important and carry on. The mutation is necessary for ensuring the diversity of genes in the population [6].

G. Model Lifecycle

When the final condition is met, GA must stop, and return the best result in the current population [18]. There are two major types of parameter values setting, which are classified according to the behavior of the parameter values through the run. The first type is parameter tuning, a common technique based on experimenting with the diverse values of crossover and mutation. After selecting the value with the best results, the final run of the algorithm is carried out. There is no changing of the value during the course of this operation (fixed value).

The second type is parameter control based on the initial values for crossover and mutation, which are changed in some way during the run. Changes to the parameter values can be divided into three types:

1) Deterministic parameter control: It is used if the parameter value demands certain modifications using the same rule of the outcome. The parameter value should be tuned for producing the typical output with no results of the search process.

2) Adaptive parameter control: It is used if the specific type of feedback demanded from the search option helps in changing the parameter.

3) Self-Adaptive parameter control: With this type, the GA is capable of developing its parameters. The values of the parameters that are used are included in the individuals and persist during mutation and crossover.

These three types of parameter settings are commonly used by researchers in attempting to find optimal or near-optimal solutions with the best rates of the operators of crossover and mutation. This approach contributes to new strategies for controlling the rates of crossover and mutation. The proposed technique is classified as an adaptive parameter control [19].

Eight techniques of GA were used to compute the accuracy: Crossover and Mutation, Crossover only, Mutation only, crossover with adaptive parameter control (1), Crossover with adaptive parameter control (2), Crossover with adaptive parameter control (3), crossover with adaptive parameter control (4), and mutation with adaptive parameter control (5). Assume C: for Crossover, M: for Mutation, I: maximum of iterations, P: for population size, PF: for population fitness, MF: for Max Fitness,

$$C = (P * I)^2 \quad (1)$$

$$C = (P/I)^2 \quad (2)$$

$$C = \left(\frac{1}{2}\right)^{P*I} \quad (3)$$

$$C = \sum P - mean(PF) * MF - mean(PF) * \frac{1}{P-1} \quad (4)$$

$$M = \sum P - mean(PF) * MF - mean(PF) * \frac{1}{P-1} \quad (5)$$

The above four equations have increased the crossover to overpass mutation performance.

Fig. 3 depicts the steps used in the classified dataset. The first step is importing the Pima Indian Diabetes Dataset, after which the data must be preprocessed and normalized. Next, the five phases of GA are considered: initial population, fitness function, selection, crossover, and mutation, with the proposed four equations to improve the accuracy of the model by using adaptive parameter control because these operators play an important role in increasing the accuracy of GA. Finally, the K-NN is combined with the GA algorithm. The steps are repeated to produce a good result.

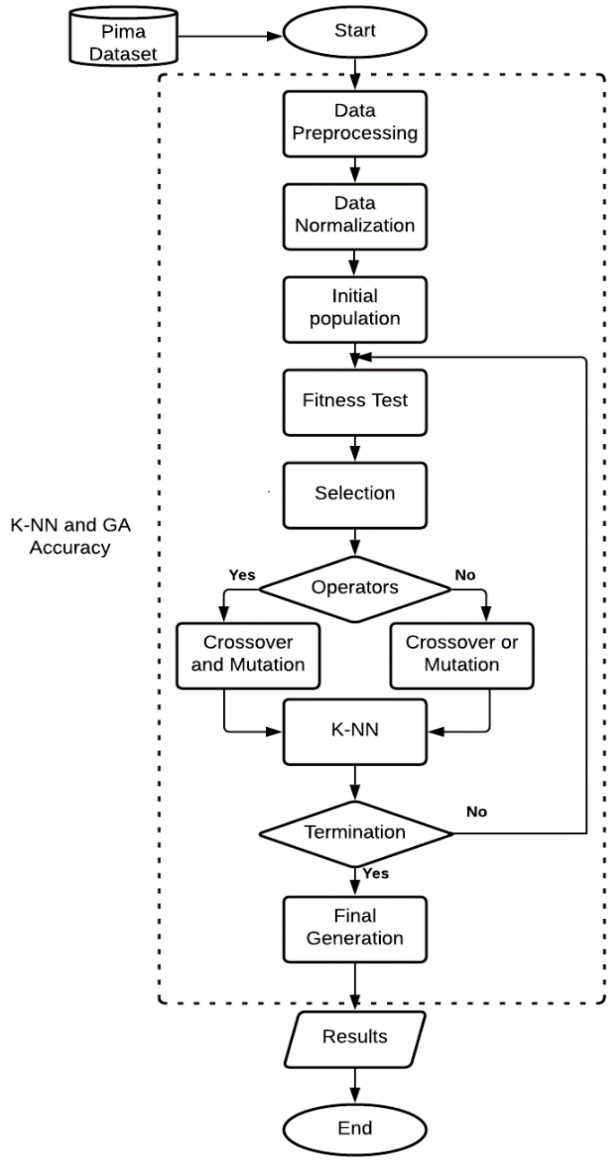


Fig. 3. The K-NN and GA Combination Model.

IV. Results and Discussions

This paper classifies the Pima Indian Diabetes Dataset into two groups: 0 or 1 (for negative '0', for positive '1') to compare the accuracy of combined K-NN and GA models using different genetic operators. Table III shows the Mean Accuracy of the eight techniques. Crossover operator with adaptive parameter control (4) achieved the best accuracy with 0.84, surpassing the performance of the mutation operator.

A. Scatter Plots Results

The mapping of scatter plots is similar to the mapping of line graphs in that they begin with mapping quantitative data points. The difference between the two plots is that the decision making on the scatter plots for the individual's point must not be connected directly with a line, but instead expresses a trend. This trend can be recognized directly through the point's distribution or the regression line [20].

Table IV shows the GA parameters and their values. The GA is configured to have a population size of 50 and was run for three generations. Crossover and mutation probability were 0.9 and 0.1, respectively.

Fig. 4 shows the scatter plot of the eight techniques in each group of generations.

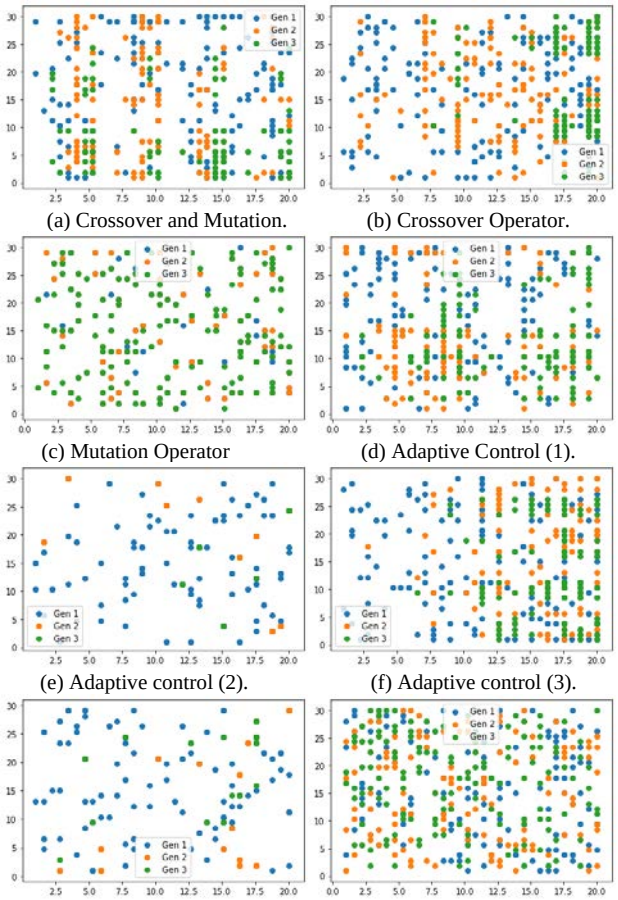
In experiment (a), a scatter plot shows that generation three has the best accuracy. In experiment (b), a scatter plot shows that generation two has the best accuracy. In experiment (c), a scatter plot shows that generation three has the best accuracy. In experiment (d), a scatter plot shows that generation two has the best accuracy. In experiment (e), a scatter plot shows that generation two has the best accuracy. In experiment (f), a scatter plot shows that generation three has the best accuracy. In experiment (g), a scatter plot shows that generation two has the best accuracy. Finally, in experiment (h), a scatter plot shows that generation two has the best accuracy.

TABLE III. Comparison of Mean Accuracy of the Different Techniques

Technique	Accuracy
Crossover and Mutation operators	0.71687
Crossover operator	0.71428
Mutation operator	0.72726
Crossover with adaptive parameter control (1)	0.74804
Crossover with adaptive parameter control (2)	0.75064
Crossover with adaptive parameter control (3)	0.72726
Crossover with adaptive parameter control (4)	0.84102
Mutation with adaptive parameter control (5)	0.81538

TABLE IV. GA Parameters and Values

Value	Parameter
Population size	150
No. of generation group	3
Crossover probability	0.9
Mutation probability	0.01
Max iteration	300



(a) Crossover and Mutation. (b) Crossover Operator.
(c) Mutation Operator (d) Adaptive Control (1).
(e) Adaptive control (2). (f) Adaptive control (3).
(g) Crossover Adaptive Control (4). (h) Mutation Adaptive Control (5).

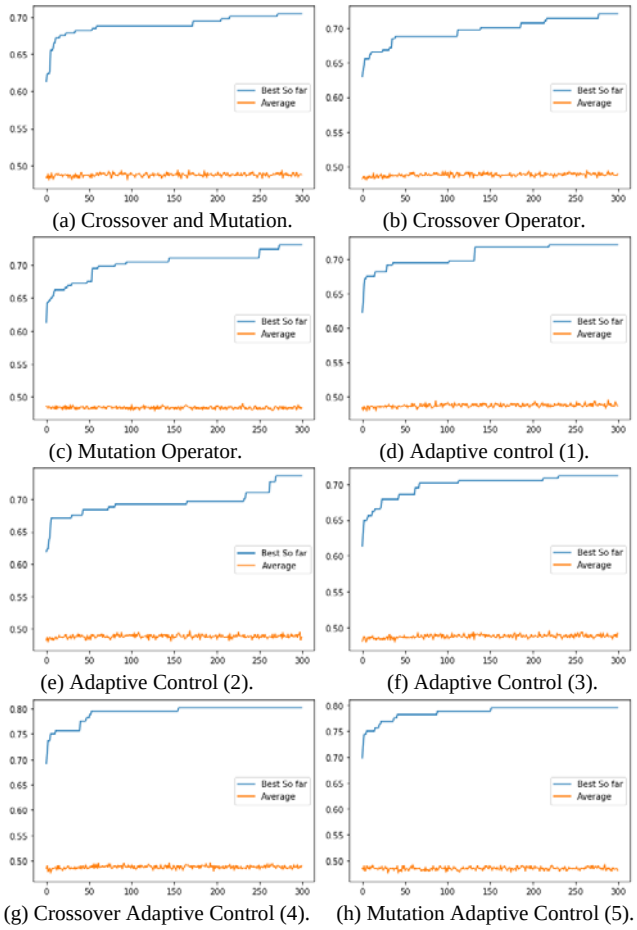
Fig. 4. Scatter Plots of the Eight Techniques.

B. Line Plots Results

Line plots are an excellent way of mapping quantitative variables. They can be either independent or dependent. If both variables are quantitative, the graph will have a slope that consists of the line segment. The latter can be visually interpreted relative to the slope of other lines or can be expressed as an accurate mathematical formula. [20] Fig. 5 shows the line plot of the eight techniques.

In experiment (a), the line plot of generation three achieved an accuracy level of 0.71. In experiment (b), the line plot of generation two achieved an accuracy level of 0.73. In experiment (c), the line plot of generation three achieved an accuracy level of 0.75. In experiment (d), the line plot of generation two achieved an accuracy level of 0.77. In experiment (e), the line plot of generation two achieved an accuracy level of 0.75. In experiment (f), the line plot of generation three achieved an accuracy level of 0.73. In experiment (g), the line plot of generation two achieved an accuracy level of 0.84. Finally, in experiment (h), the line plot of generation two achieved an accuracy level of 0.81.

Fig. 6 and 21, located in the appendix, show the results of the experiments according to the type of genetic operators used.



(a) Crossover and Mutation. (b) Crossover Operator.

(c) Mutation Operator. (d) Adaptive control (1).

(e) Adaptive Control (2). (f) Adaptive Control (3).

(g) Crossover Adaptive Control (4). (h) Mutation Adaptive Control (5).

Fig. 5. Line Plots of the Eight Techniques.

V. Conclusion

Diabetes comes in many forms. Among the general population, Type 2 Diabetes Mellitus (T2DM) is the most prevalent. The percentage of patients suffering from (T2DM) is approximately 90%. Thus, early and accurate diagnosis can help to reduce mortality rates. In this paper, a combined K-NN and GA model was conducted to improve the accuracy of the K-NN algorithm. The model was implemented by using the Anaconda Python platform on the Pima Indian Diabetes Dataset.

In terms of the accuracy of the proposed models of the two GA operators (crossover, mutation), the crossover with adaptive parameter control (4) is the best technique, achieving a mean accuracy of 84%, 1% better than the best result in the literature review. The combination of K-NN and GA is an effective choice for diagnosing diabetes accurately.

This paper aimed to enhance the performance of the K-NN algorithm. Many techniques have the potential to achieve this goal. This paper developed and used a hybrid GA-K-NN algorithm. Some suggestions for future work include:

- Modifying the mutation and crossover rates by using parameter tuning or parameter control (deterministic parameter control, adaptive parameter control, and self-adaptive parameter control) to achieve greater accuracy.
- Using different equations of parameter tuning or parameter control, which can modify the mutation and crossover rates to increase the accuracy percentage.
- Applying a hybrid GA-K-NN algorithm, which may achieve a good result if used in a real dataset to diagnose diabetes accurately.
- Merging another classification algorithm with GA, which might achieve an accurate diagnosis of diabetes.
- Combining the K-NN algorithm with other algorithms, which may enhance its performance.
- Applying GA with feature selection by using fewer input features, which may give rise to more accurate results.

References

[1] "International Diabetes Federation - Facts & figures", Idf.org, 2020. [Online]. Available: https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html. [Accessed: 06- May- 2020].

[2] M. Ngugi, J. Njagi, C. Kibiti, J. Ngeranwa and E. Njagi, "Diagnosis of Diabetes Mellitus", International Journal of Diabetes Research, 2012.

[3] M. Pradhan, G. Bamnote, V. Tribhuvan, K. Jadhav, V. Chabukswar and V. Dhobale, "A Genetic Programming Approach for Detection of Diabetes", International Journal Of Computational Engineering Research, vol. 2, no. 6, 2012.

[4] B. Ramageri, "Data Mining Techniques and Applications", Indian Journal of Computer Science and Engineering (IJCSE), 2010.

[5] R. Patil and S. Tamane, "Upgrading the Performance of KNN and Naïve Bayes in Diabetes Detection with Genetic Algorithm for Feature Selection", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJCSE), vol. 3, no. 1, 2018.

[6] A. Thengade and R. Dondal, "Genetic Algorithm – Survey Paper", International Journal of Computer Applications (IJCA), 2012.

[7] lmandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events Theoretical Background", international Journal of Engineering Research and Applications (IJERA),, 2013.

[8] Y. Christobel and P. Sivaprakasam, "A New Classwise k Nearest Neighbor (CKNN) Method for the Classification of Diabetes Dataset", International Journal of Engineering and Advanced Technology (IJEAT), vol. 2, no. 3, 2013.

[9] Y. Christobel and P. Sivaprakasam, "Improving the performance of k-nearest neighbor algorithm for the classification of diabetes dataset with missing values", International Journal of Computer Engineering and Technology (IJCET), vol. 3, no. 3, 2012.

[10] J. Kandhasamy and S. Balamurali, "Performance Analysis of Classifier Models to Predict Diabetes Mellitus", Procedia Computer Science, vol. 47, pp. 45-51, 2015.

[11] N. Bidi and Z. Elberrichi, "Best Features Selection for Biomedical Data Classification Using Seven Spot Ladybird Optimization Algorithm", International Journal of Applied Metaheuristic Computing, vol. 9, no. 3, pp. 75-87, 2018. Available: 10.4018/ijamc.2018070104.

[12] D. Gupta and U. Ghose, "A comparative study of classification algorithms for forecasting rainfall," 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, 2015, pp. 1-6, doi: 10.1109/ICRITO.2015.7359273.

[13] S. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background", International Journal of Engineering Research and Application, vol. 3, no. 5, 2013. [Accessed 23 November 2020].

[14] T. Alam, S. Qamar, A. Dixit and M. Benaida, "Genetic Algorithm: Reviews, Implementations, and Applications", International Journal of Engineering Pedagogy (IJEP), 2020. [Accessed 23 November 2020].

[15] [4]Z. Piserchia, "Applications of Genetic Algorithms in Bioinformatics", Master, University of California, 2018.

[16] M. Panwar, A. Acharyya, R. Shafik and D. Biswas, "K-Nearest Neighbor Based Methodology for Accurate Diagnosis of Diabetes Mellitus", in Sixth International Symposium on Embedded Computing and System Design (ISED)., 2016.

[17] L. Haldurai, T. Madhubala and R. Rajalakshmi, "A Study on Genetic Algorithm and its Applications", international journal of computer sciences and engineering (JCSE), vol. 4, no. 10, 2016. [Accessed 4 June 2020].

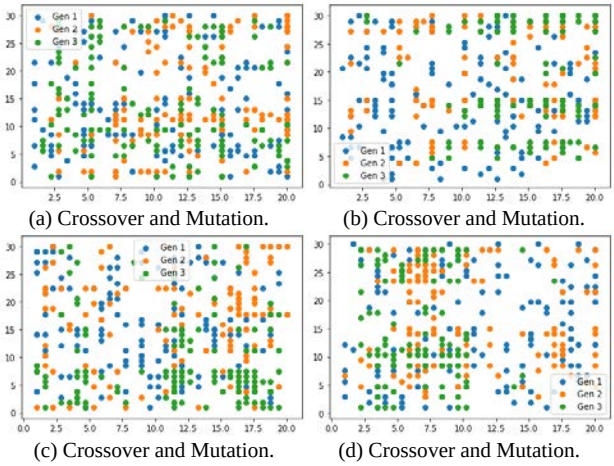
[18] K. Abd Halim, A. Jaya and A. Fadzil, "Data Pre-Processing Algorithm for Neural Network Binary Classification Model in Bank Tele-Marketing", International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 3, pp. 272-277, 2020.

[19] A. Hassanat, K. Almohammadi, E. Alkafaween, E. Abunawas, A. Hammouri and V. Prasath, "Choosing Mutation and Crossover Ratios for Genetic Algorithms—A Review with a New Dynamic Approach", Information, vol. 10, no. 12, p. 390, 2019. Available: 10.3390/info10120390.

[20] "Graphing - Line Graphs and Scatter Plots", Labwrite.ncsu.edu, 2020. [Online]. Available: https://labwrite.ncsu.edu/res/gh/gh-linegraph.html. [Accessed: 06- May- 2020].

APPENDIX

The scatter plot of the Crossover and Mutation technique:

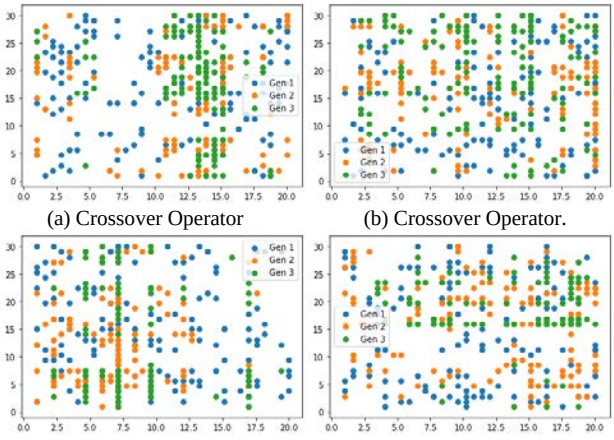


(a) Crossover and Mutation. (b) Crossover and Mutation.

(c) Crossover and Mutation. (d) Crossover and Mutation.

Fig. 6. Fitness Distribution Values of Crossover and Mutation.

The scatter plot of the Crossover only technique:

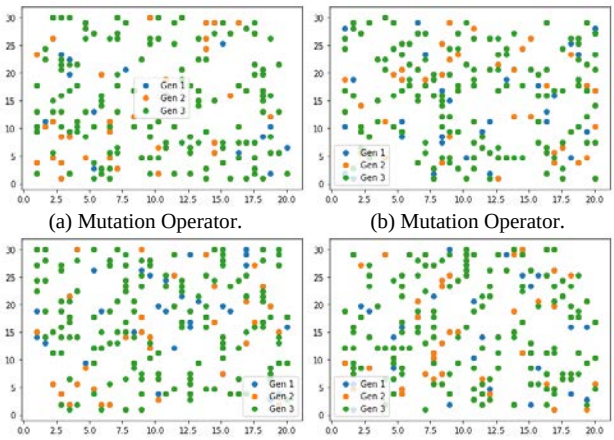


(a) Crossover Operator (b) Crossover Operator.

(c) Crossover Operator. (d) Crossover Operator.

Fig. 7. Fitness Distribution Values of Crossover Only.

The scatter plot of the Mutation only technique:

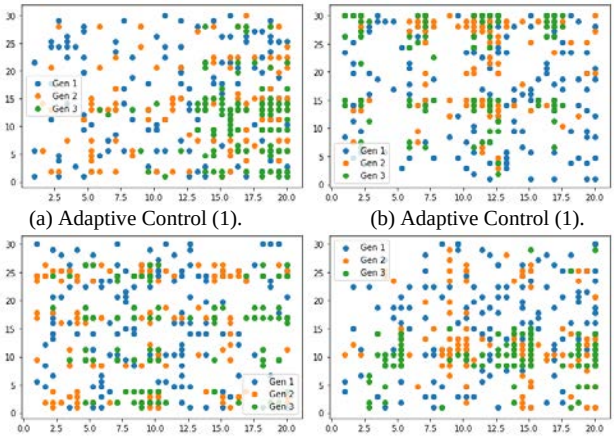


(a) Mutation Operator. (b) Mutation Operator.

(c) Mutation Operator (d) Mutation Operator.

Fig. 8. Fitness Distribution Values Mutation Only.

The scatter plot of the Proposed Crossover Equation (1):

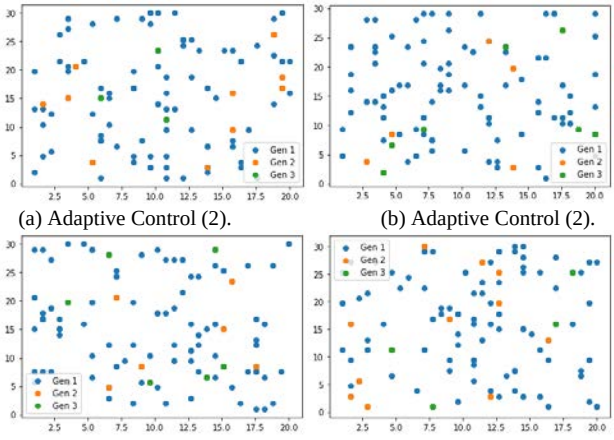


(a) Adaptive Control (1). (b) Adaptive Control (1).

(c) Adaptive Control (1). (d) Adaptive Control (1).

Fig. 9. Fitness Distribution Values of Crossover Equation (1).

The scatter plot of the Proposed Crossover Equation (2):

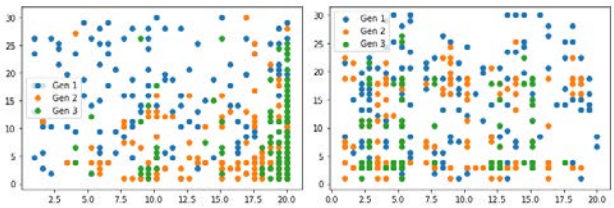


(a) Adaptive Control (2). (b) Adaptive Control (2).

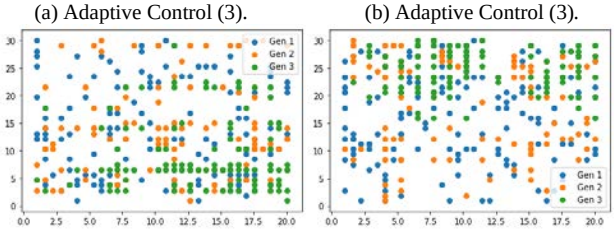
(c) Adaptive Control (2). (d) Adaptive Control (2).

Fig. 10. Fitness Distribution Values of Crossover Equation (2).

The scatter plot of the Proposed Crossover Equation (3):



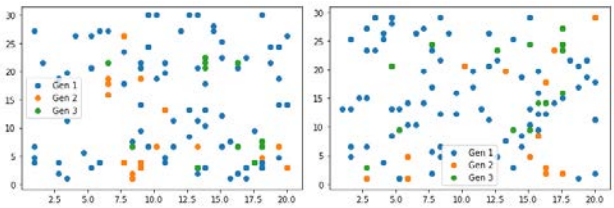
(a) Adaptive Control (3). (b) Adaptive Control (3).



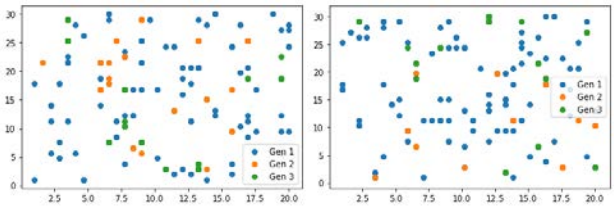
(c) Adaptive Control (3). (d) Adaptive Control (3).

Fig. 11. Fitness Distribution Values of Crossover Equation (3).

The scatter plot of the Proposed Crossover Equation (4):



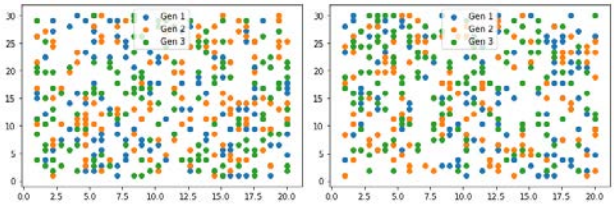
(a) Crossover Adaptive Control (4). (b) Crossover Adaptive Control (4).



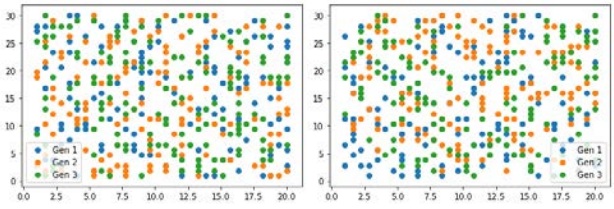
(c) Crossover Adaptive Control (4). (d) Crossover Adaptive Control (4).

Fig. 12. Fitness Distribution Values of Crossover Equation (4).

The scatter plot of the Proposed Mutation Equation (5):



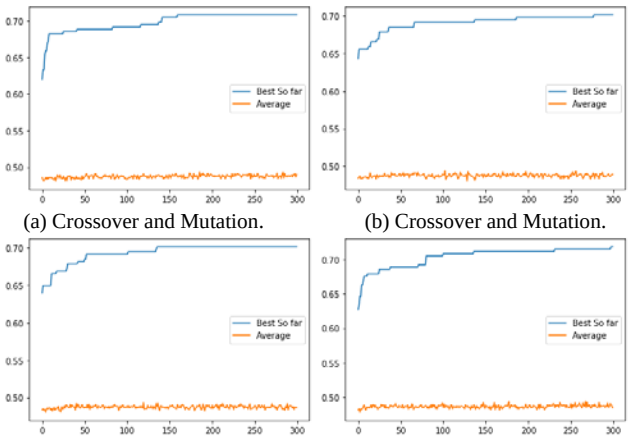
(a) Mutation Adaptive Control (5). (b) Mutation Adaptive Control (5).



(c) Mutation Adaptive Control (5). (d) Mutation Adaptive Control (5).

Fig. 13. Fitness Distribution Values of Mutation Equation (5).

The line plot of the Crossover and Mutation technique:

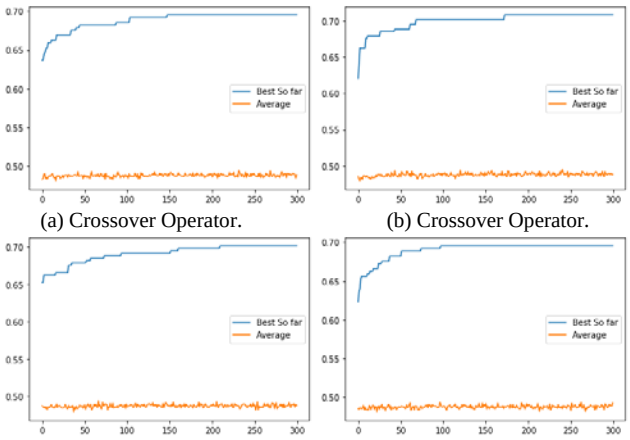


(a) Crossover and Mutation. (b) Crossover and Mutation.

(c) Crossover and Mutation. (d) Crossover and Mutation.

Fig. 14. Fitness Distribution Values of Crossover and Mutation.

The line plot of the Crossover only technique:

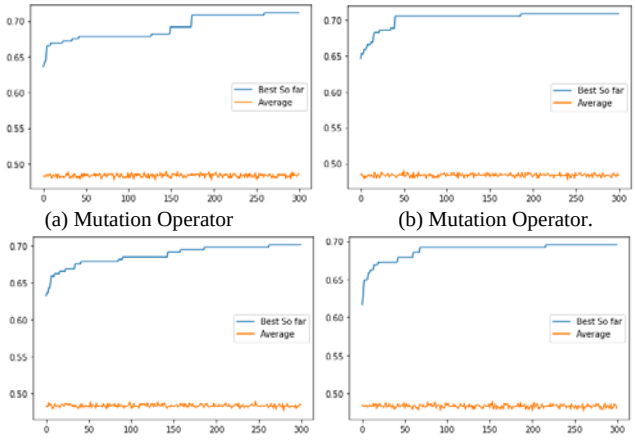


(a) Crossover Operator. (b) Crossover Operator.

(c) Crossover Operator (d) Crossover Operator.

Fig. 15. Fitness Distribution Values of Crossover Only.

The line plot of the Mutation only technique:

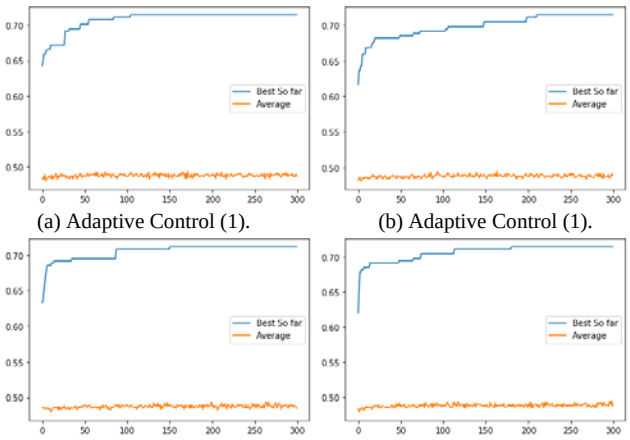


(a) Mutation Operator (b) Mutation Operator.

(c) Mutation Operator (d) Mutation Operator.

Fig. 16. Fitness Distribution Values of Mutation Only.

The line plot of the Proposed Crossover Equation (1):

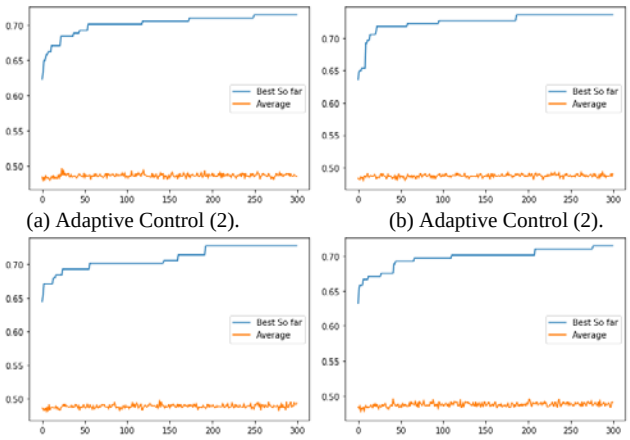


(a) Adaptive Control (1). (b) Adaptive Control (1).

(c) Adaptive Control (1). (d) Adaptive Control (1).

Fig. 17. Fitness Distribution Values Crossover Equation (1).

The line plot of the Proposed Crossover Equation (2):

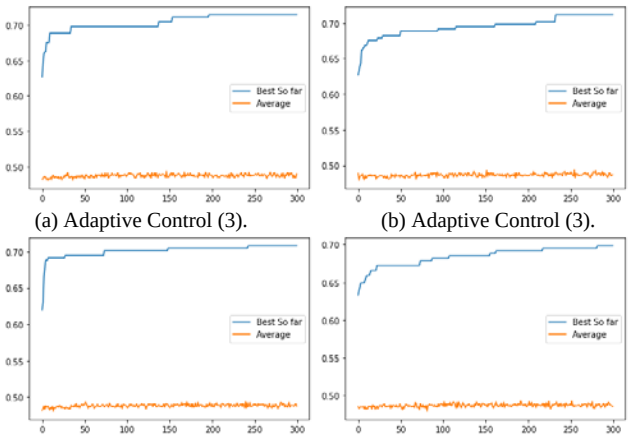


(a) Adaptive Control (2). (b) Adaptive Control (2).

(c) Adaptive Control (2). (d) Adaptive Control (2).

Fig. 18. Fitness Distribution Values Crossover Equation (2).

The line plot of the Proposed Crossover Equation (3):

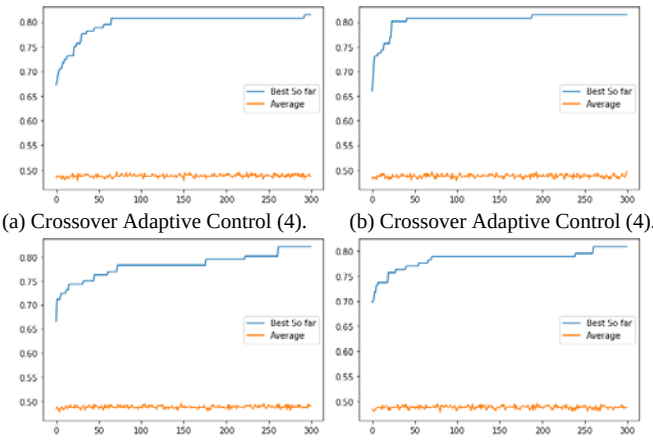


(a) Adaptive Control (3). (b) Adaptive Control (3).

(c) Adaptive Control (3). (d) Adaptive Control (3).

Fig. 19. Fitness Distribution Values Crossover Equation (3).

The line plot of the Proposed Crossover Equation (4):

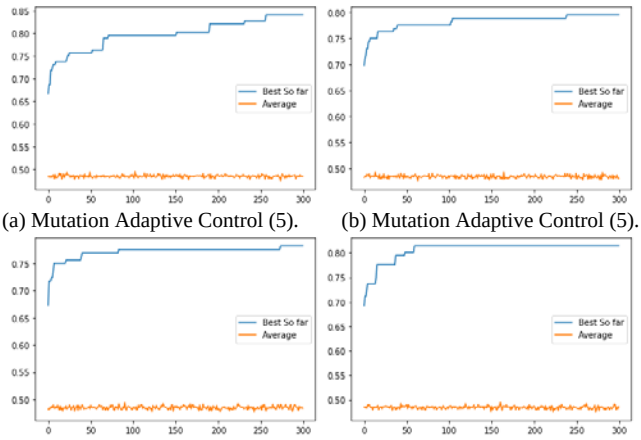


(a) Crossover Adaptive Control (4). (b) Crossover Adaptive Control (4).

(c) Crossover Adaptive Control (4). (d) Crossover Adaptive Control (4).

Fig. 20. Fitness Distribution Values Crossover Equation (4).

The line plot of the Proposed Mutation Equation (5):



(a) Mutation Adaptive Control (5). (b) Mutation Adaptive Control (5).

(a) Mutation Adaptive Control (5). (b) Mutation Adaptive Control (5).

Fig. 21. Fitness Distribution Values Mutation Equation (5).

Enhanced Method to Stream Real Time Data in IoT using Dynamic Voltage and Frequency Scaling with Memory

H. A. Hashim
Department of Computer Science
College of Computer Science and Engineering
Taibah University, Yanbu, KSA

***Abstract*—DVFS (Dynamic Voltage and Frequency Scaling) is a popular CPU (Central Processing Unit) level voltage frequency scale technology based on the application precedence. To motivate recurrence / voltage scaling as a feasible tool for energy productivity, i) basic workloads should ensure that memory recurrence scaling has an impact with insignificant degradation and (ii) that there is an enormous open door for reduction of power in this work. Therefore, if memory recurrence is that, two limiting forces on energy efficiency impact all items in an anomalous state. The competence depends on both power and runtime, because energy is the result of time and energy. The reduction in control alone will increase skills. However, further discussions at lower control work focuses are conducted, expanding operating times and energy in this way. There is a bloating edge that decreases the recurrence / voltage of memory in this way. This shows further that the recurrence of statically-scaling memory has little impact on many lower workloads because of recurrent effects only the idling of transmission interchange, part of the memory dormancy. This will be shown. Inspire in this paper, the scaling of memory recurrence will affect frame power (show a systemic model to simplify the scaling of voltage) and therefore electricity. It presents DVFS memory computing in real time. The DVFS technology is popular for measuring the frequency of voltage according to the CPU level applications. In this work, an enhanced DVFS with memory technique proposal is used to decrease energy use and improve performance at the memory level**.

Keywords*—Dynamic voltage; frequency scaling; central processing unit

I. INTRODUCTION

Data Streaming in most environments is very tedious and different factors has to be concentrate to keep the quality of service.

A. Big Data Stream Computing (BDSC)

BDSC, a long-lasting vision for ‘high performance computing’ and ‘high real time computing’, has opened a new age of upcoming computing due to Big Data, a dataset that is big, large, unstructured, scattered and beyond the capacity of software and devices to access, obtain, analyze, etc. [1] Stream computing system is a computer standard which reads data from hardware or software sensor collections in a stream format and computes continuous real-time data streams, where results and feedback should also be made available in a reliable data stream. A stream of data is an infinite series of datasets in the network and has more parallel streams processing a single stream at the same time [2].

B. IOT Application Data Streams

The IoT notes that innovation is a paradigm in which the inevitable numbers of sensors in the trillions will have the capacities to slowly track physical conditions, individuals and virtual substances; to deal with real as well as on-line perceptions; to carry out activities that improve the efficiency and efficacy of the environment or society's way of life and consolation [3]. Over a span of ten years, the building blocks of IoT engineering have evolved with sensor systems work and inevitable processing experience. The existing IoT inflection point has all been contributing to the ongoing development in capacity of a quick-port system (e.g. 2G/3G/4 G) and impromptu (e.g. Bluetooth) systems, affordability, smartphones and crowd-sourced data collection [4], big data analytics and the cloud-based data center platforms. IoT applications are now also present in vertical areas, such as failure management and demand control in intelligent grids, as well as sleep and exercise tracking, and health band and intelligent watch recommendations.

In particular, these streams were explored in huge scales (billions of sensors, several opportunities per second) to request transmission of large-scale computational assets through transmitted sensors. Distributed computing offers a typical stage for adapting the perceptions in all server farms conveyed and sending critical answers to the edge of the IoT framework.

Collection of IoT Data streams, interfacing, and improving these data streams with the relevant contextual metadata such as time stamps as well as location data, including too many IoT sources (such as devices linked to Internet). The collection process, as previously outlined, traditionally requires a solution to the heterogeneity of their IoT data streaming and the heterogeneity of interfaces to data formats and sources. IoT application authentication and validation of data source and origin format. Application validation. The method requires a precision, continuity and credibility validation. IoT data interoperability and convergence, which deal with IoT data streams homogenizing and uniting, capture IoT data streams, interfaces and enhances these streams with the use of relevant contextual metadata, including timelines and location

data. Semantic IoT Data interoperability and fusion as previously outlined, the collection process classically requires the heterogeneity of its IoT data streams, the interfaces of data formats and data source heterogeneity. IoT application validation and data source and format validation. IoT data validation. The method requires a precision, quality and competence evaluation. The protocol Interoperability and integration of the semanticization of IoT data sources Semantic IoT data interoperability and convergence [5].

C. D-RAM

For initialization operations and bulk data copying, a sample of such wastage often happens on which a copy page is initialized or added to a value. If the processor doesn't explicitly need the initialized or copied data then the processor can substantially save energy by changing the bandwidth and time within the DRAM (with minor changes to the DRAM) [6].

In this, DRAM chip can be detected internally on a large data set. The core DRAM structure can be exploited and a page copy can be done in the DRAM without any data being fetched from the DRAM chip, as shown in recent work. If the source and destination page are within the same DRAM subarray, the results show a page copy that is improved by more than a demand in magnitude, which is important to reduce the energy by ~74 times and no bandwidth wastage for a DRAM data bus [7].

The main idea is that 1) deactivated by the source row content of the knowledge amplifier, 2) by the source row, which offers very small hardware costs, which is smaller than 0.03% of DRAM chip space, and 3) straight activates the destination row, allowing the information amplifiers to push its contents through. The main idea is to deactivate the source row.

A few definitions must first be made to clarify the system of fall-off workload. The main idea is to enforce the use of the memory bandwidth, then change the memory frequency / voltage, to reduce the power with minimal DVFS memory loss as in Fig. 1.

The main contributions of our research work include:

- Evaluate the scaling of memory frequency / voltage to increase energy efficiency and reduce storage energy.
- Observation the frequency-dependent part of the memory system power has a control algorithm, which reduces the memory frequency and reduces the performance impacts.
- The important point is that the lowering of memory frequency does not significantly change memory access latency when bandwidth is used in low capacity.
- The propose control algorithm increases the memory frequency when the usage reaches a threshold, reducing the output effect, by monitoring the memory bandwidth utilization.

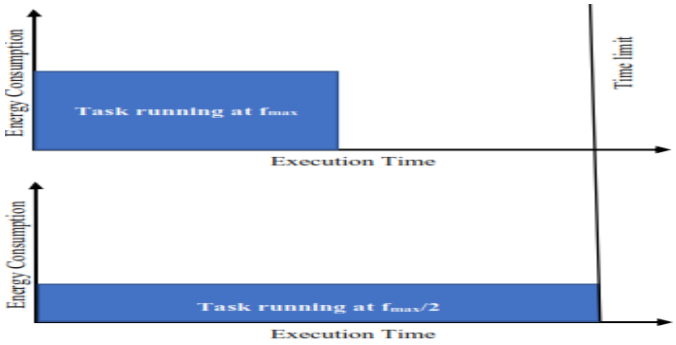


Fig. 1. Dynamic Voltage Frequency Scaling.

II. Related Work

Big Data Stream Computing is running in the memory entirely. Apply DVFS technology to the massive cluster of knowledge and the scientific evidence that gives Sun et al workloads for every on / off chip doling [8]. The duty to perform a job is defined as the entire CPI of all covers towards pathway of undertaking. The workload depends on various component parameters, such as the number of log jam on-chip as a result of data / control dependency or the missed branch requirement, and log jam cycle off-chip control as a consequence of I / D storage errors or I / D TLB failures. The CPU is located in the middle of an off-chip until the memory is needed. In the middle of an off-chip, the CPU clock cycles are therefore lost [9].

The use of DVFS in Baskiyar's main memory [10] is yet another interesting way to reduce energy consumption in DRAM. David et al [11] note that using a lower memory bandwidth does not substantially change memory access latency by reducing memory voltage / frequency. Comparatively current works have shown the importance of saving energy both in real simulated systems from the Sharifi and Shahrivari [12] systems, by adjusting the memory frequency and voltage based on predicted use of memory bandwidth. DVFS memory level can successfully allow dynamic heterogeneity in DRAM channels, leading to new mechanisms for optimization and personalization. The study of more fine-grained methods of management of power within the DRAM chips and the idle and active modes of low power are also very promising. In addition to the valuable space on the chip cache, the existing systems energy consumption considerably waste DRAM time and bandwidth by moving data from main memory to processors caches, often inappropriately.

This constraint on the application of data streaming in real time has been applied to new methods for data mining. The Sudipto [13] project variations in the k-means grouping algorithm. In Sudipto (2019), the authors present the HP Stream, a method for the grouping of streaming data. Yixin Chen and Li Tu [14] understand that the grouping process is distributed online to a variable that periodically stores and collects a detailed overview statistics, also an offline element which is only used for summary statistics. Yixin Chen and Li Tu [15] present density grouping methods for streaming data.

In an ever-growing environment, a classification process may involve in real time model testing and construction. Pedro Domingos and Geoff Hulten [16] present decision tree-based methods which include only the data stream via a single pass. The on-demand classification method in Charu et al (2016) suggests dynamically selecting tools from the correct window of previous training information to be used for classification building. The problem of routine data-streaming pattern mining is investigated. The loss counting methodologies and sampling methods for holding approximate counts over a window slide with a space with restricted access are available in Gurmeet Singh Manku and Rajeev Motwani [17].

Wang et al [18] suggest PW Join, is a three-operation algorithm designed to add binary windows to activities that are value-driven and can be found in streaming data. In Daoud [19] authors suggest GrubJoin, a multi-purpose, adaptive window data stream connection to efficiently complete CPU load flaking associations.

III. Existing Method

The existing method there is no concentration of memory. In BDSC platforms, DVFS-level fonts are accessible. It increases the energy for small businesses. The equally recent needs of the fundamental association of memory are threefold. To begin with, diverse innovation: an innovation that is quite adaptable to resources, cost and constraints, as defined by the past, is a new requirement for the outcome. Because DRAM, which continually stretches over 100-nm to 30-nm from innovation hubs, the need for more mobile innovation did not stand out. From now on, DRAM has not been done as much as the 30 nm hub with the enormous device or circuit scaling errands.

Secondly, the QS execution and accuracy in a common simple memory setting has been similarly forgiven for a new requirement. As one center system was a memory and the current limit and transmitting ability were a very small measure of an asset sharing, a much less prevalent or apparent need for planned implementation was present. Today's top methodological constraints have become increasingly common with more specialists / centers sharing memory frames on a chip and the aggregate costs of storage execution, reasonableness of memory, organizational union and methods to control memory impedance. Thirdly, the technique produced for the fundamental memory framework requires significantly higher data transmission capacity / control / energy effectiveness.

Propelled energy productivity to allow the frameworks of data transmission capacities, power and electricity is far more flexible where shared principle memory is shared between various operators, and it can allow new applications in all areas where processes are used. This is no innovative precondition at the moment, it is conceivably an additional cutting edge technology constraint which has not remained the usual calculation of cost, limitations and implementation.

A. Drawbacks

The techniques used by the DVFS are used to degrade and *measure* processor operating frequency and input voltage. It is a dynamic approach to reducing the use of energy. The lower the power state, the more prominent Baskiyar (2010) points out the energy reserve funds. Nevertheless, the execution time is longer. The time to complete the execution is littler and the task is completed before the date for which the errand remains running at higher express capacity. At the moment that the company is completed, the part is run in a stand-off gear, some amount of energy is also required. When the business continues to run in a small power express, the execution time is longer but not past the mark. When an assignment is completed in time, only less energy is required.

B. Proposed Method Dvfs with Memory Technique

Add frequency scaling algorithms for the energy efficiency boost. It reduces the application's energy consumption and improves application efficiency. The software basis is the frequency-based control system. The control strategy of data streams is displayed in Algorithm. The computing platform in large data is managed when IoT application data is being computed. It is therefore difficult to look for additional loopholes to overcome this problem. In Stream Computing Storm platform is open to meet the full requirements of the current data streaming dimension.

Round Robin is used by default in this scheduling strategy. The data circulation is not adequate at the point slow processing and the power use very many failures to this end. Data generated by computers. In the Storm platform shown in Fig. 2 the mechanism runs internally.

Three different types of data sensing devices (2) communication on request, Billard et al (1993), (3) event driven communication from Meng et al (2014). In the real time of computing the data streams, three different types of data sensing devices are available in IoT. It requires more computational power to process this type of data. Modifying the timetable approach to suit the various data sets offers highly effective management of data. The algorithm updated notice Algorithm 5.1 which calculates the time complexity of the algorithm. For further analysis, steps are identified with a double slash in the algorithm itself. Three strategies are used for the process mention in algorithm. Secondly, the DVFS referred to in Sveur et al (2010) solutions to energy reduction refine their stream maps. Secondly, there is hot-swapping technique to reprogram your worker nodes online. The third is to assign one and two approaches to performance scaling.

Input: Real-Time IoT streaming data

Output: Optimized Solution with frequency based task allocation

Step 1: Bandwidth required data stream
Step 2: Bandwidth available to process task
Step 3: Average Bandwidth thread assigning task
Step 4: if *B*andwidth < *Tf*b then
Step 5: set memory frequency to *f*b
Step 6: else if *Tf*b = *B*andwidth < *Tf*c
then set memory frequency to *f*2
Step 7: else if *Tf*c = *B*andwidth then
Step 8: set memory frequency to maximum frequency *f*a
Step 9: end if
Step 10: end while

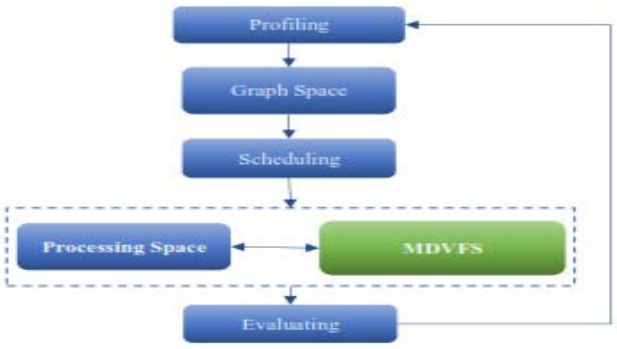


Fig. 2. Internal Work Flow of Storm.

C. Evaluating the Performance Time of a Task

The application of the DVFS technique to the huge group of gushing information and the scientific confirmation of every on / off chip doling process referred to by Sun et al (2015). The workload of an order is described as the total of all covers of the CPI to the stream of path. Task workloads depend on different component parameters such as on-board data / control reliability stoppage number and off-off-board log jam period checks due to I / D shop missing or I / D TLB missing data / control depending on branch missing forecasting. In the middle of an off-chip the CPU returns until the memory trade has been requested.

CPU clock cycles are therefore adjusted in the middle of an off-chip. This must first provide a few examples in order to clarify the rotary workload structure.

Sun et al (2015) shows Energy Reduction by graph construction methods, where Storm is a spelling structure and it is unbelievably interesting in relation to recently suggested package planning systems. The energy reductions are given in response time and high energy calculations.

It also generates a DAG of centers which enables the record to move between centers. It is ideal to vary from Hadoop MapReduce from Bhandarkar (2011) [20], storm and spark streaming (spouting structure rendered over Spark). All consider a DAG of activities covering the retaliation of spouts, but then imperceptibly process the DAG in particular. Storm defines the DAG of the center and sets out separate centers for each activity of the DAG. Streak Streaming does not reassign, but using Spark's basic instruments to delegate commands to available resources vigilantly. This gives various kinds of performance properties.

D. Enhanced Data Stream Adapting Dvfs with Memory

For BDSC implementations, the in-memory state is always in use. This uses the same memory capacity at all measurement speeds of data streams. Control frequency levels and improve performance for low and high task topologies.

BDSC for IoT workload analysis evaluates factors such as latency, jitter, performance, CPU and the use of memory. When running the program, three separate frequency ranges show that the transmission capacities will change as much as the static recurrence decrease in low data transfer capacity benchmarks, when retaining the remarkable decrease in capability (due to the extended runtime and the aggregation swells along these lines). The Bandwidth approach (0.5, 2) decreases memory control by 6,15W (11,41W maximum) over SPEC CPU 2006, for 0.18% normal (1.72% maximum) lull. All things considered, over the SPEC CPU 2006. If the data transfer capacity (i.e. with transmitting capacity generally less than 1.2 GB / s per channel) is sorted on one side of the gcc, all considerations of the lower memory frame control by 9.81W are taken into account.

1) Dram structure background: This is important when the standard 1363MHz memory structure is 65.1W. The bandwidth (0.5, 2) in the entire setting gives 2.43% normal (maximum 5.25%) change in energy efficiency. It is important to note that I reductions in control and thus improvements in energy productivity due to memory DVFS that give the approach CPU-based DVFS a similar degree of power as in Table I.

A device arrangement composed of a DIMM has a rank. In a rank, the different banks provide free DRAM storage items with relevant translators and intelligence speakers on each unit. Both parties share the I / O (collectors and drivers) equipment to the edge of DDR transport. Each group is a cell lattice, divided into columns. The column support can be used at any time to keep the individual dynamic line (or page). An initial call puts a line in the wire, which enables read / compose instructions to enter segments in the line.

A pre-charge charge returns information to the display and is ready for the subsequent execution. Itemized memory process representations can be found in detailed data sheets. Various strategies are the way a memory controller handles these charges. The page-open and page-shut solutions are two common methods. Page-open retains the last line dynamic (page) in the line pad. It allows the following access to a similar line, as no precise information is necessary. Page-shut plays a pre-load when the load is complete. Despite the fact that this brings outline hits, it decreases idleness in a missed line, because no preload is necessary, only an actuate. Take note that in this paper a page shut strategy is embraced. This is informed by the vision of multi-center frames as in Table II.

TABLE I. ENERGY PER OPERATION ON DRAM WITH FREQUENCY OF 850MHZ AND 1060MHZ THROUGH BASELINE FREQUENCY IS 1363MHZ

Command	Energy (nJ) @ 1363MHz	Energy (nJ) @ 850MHz	Energy (nJ) @ 1060MHz
Write (array)	22	22	21
Read (array)	17.5	17.3	12.5
Write I/O (1 DIMM/channel)	4.2	7	3.2
Read I/O (1 DIMM/channel)	1	1.7	2
Activate +Pre-charge(page open-close)	24.9	25.1	21.3
I/O additional termination (2 DIMMs/channel)	12.3	20.9	15.2
Average energy/write, page-close policy, 2 DIMMs/channel	60.9	72	49.8
Average energy/read, page-close policy, 2 DIMMs/channel	55.2	64.8	41.6

TABLE II. BACKGROUND POWER OPERATION: EVALUATING ALL THE PARAMETERS OF THE POWER CONSUMPTION VALUES

Power-down state	Exit latency to read Command	Power @ 1363	PIL, Out. Clk	IBT, ODT	DLL	Clk. Tree	Input Buf	Self Refresh
Self-Refresh	512 tCK	0.6W	0	0	0	0	0	1
Self-Refresh	tMRd+tXPDLL	0.98W	1	0	0	0	0	1
Pre-Charge Slow Down	tMRD+tXPDLL	1.35W	1	0	0	0	0	0
Charge Slow Down	tXPDLL	1.64W	1	1	1	0	0	0
Pre-Charge fast power	tXP+tRCD	2.79W	1	1	1	0	0	0
Active Paower	tXP	3.38W	1	1	1	0	0	0
Pre-Charge Standby	tRCD	4.76W	1	1	1	1	1	0
Active standby	0	5.46W	1	1	1	1	1	0

Fig. 3, the power consumption of various segments in the DRAM system is depicted as the basis for this work.

2) I/O Power: This portion of equipment power includes information cushions, read / write handles, DLL (delay bolted circles), drivers for the information transport and logic of control and is consumed when the DRAM is still sat (not shut down) or when the order is executed effectively.

I / O capacity is subordinate to memory recurrence: with less recurrence, it declines. The I / O control field, due to the dynamic control execution of transport scales, has a circuitous effect while the energy effect of recurring scales is taken into account, which has been stated below. In memory shutdown, I / O power is reduced.

3) Registered power: A registered DIMM consists of clock and order / address line information / produce registers; enlist control includes these parts, as well as associated justification and a stage bolted cycle (PLL). Like I / O control, enlist power is recognized as recurring subordinate with the transport interface. It also correlates with low-control countries.

4) Transmission power: Termination Power: Finally, at present day DRAM device incorporate On-Die Termination (ODT) to appropriately end the transport amid dynamic operation. End power is scattered in on kick the bucket resistive components and is changed in accordance with transport electrical attributes, contingent upon DIMM number. With per channel 2 DIMMs, DDR3 end power can achieve 1.8-2.2W for each DIMM.

E. Experimental Setup

In this model, the new proposed frequency control algorithm was introduced based on DVFS memory level policies to boost energy efficiency through the introduction of a Storm device step and accurate tests. Creates a Storm platform simulation environment which has been built in full parallel fault tolerance, distributed with Storm0.10.0 the latest version software. A 4 core Intel 13 processor 2.00 GHz 64 Bit processor, 16 GB of memory, and 1 Gbps network connectivity is built for virtual machines. 4 core 2 machines with an external 10 TB storage capacity are connected to each other.

The Linux server runs on each individual machine (Ubuntu server version 14.01). The following component of software usually is designed and used in conjunction with Java 1.8, zoo 3.4.0, python 3.0. In addition, all updated scheduling techniques refer to the replacement of the existing energy-efficient traffic scheduling plan for the IScheduler in Storm network. Observing the StormUI output. The average loading time for tuple metrics is used.

In this method the default time function (Storm) was used to monitor each tuple's processing time. StormUI can collect this information, but it was an average 10-minute display. This method took averages of 1 minute in this testing and implementation, which gave us far better accuracy in real-time estimates of results. Ubuntu Linux uses the NTP protocol specifications to synchronize worker nodes for the duration of experiments. Rest of all test values listed in Table III.

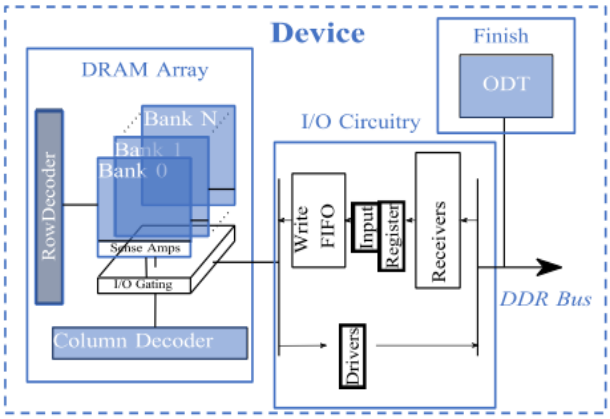


Fig. 3. General Overview of DRAM Device Structure.

TABLE III. EXPERIMENTAL VALUE ASSIGNING

S. NO	Bounds	Values
1	Estimation period and monitoring load	40 sec.
2	Estimation of coefficient (α)	0.5
3	Period of Schedule fetching $p(s_f)$	20 sec.
4	Period of Schedule generation $p(s_g)$	400 sec.
5	Each Experiment's Running Time E_{RT}	15000 sec.
6	No. of worker node available	15

IV. Results and Discussions

The BDSC evaluation of factors like latency, jitter, speed, CPU, and memory use for the IoT workload discussion. During implementation three different frequency ranges are established and demonstrate that the data transfer strategy can improve both energy productivity and static recurrence, while retaining a strategic distance from the broad performance declines, for lower transmission capacity thresholds (due to extended runtime and along these lengths). Normal framework control decreases: decreases are critical for low data transmission, compared with time spent at lower frequencies. Generally speaking, bandwidth (0.5, 2) decreased memory power by 6.15W (11.42W maximum) (DC control) for a standard (1,71% maximum) log jam by 0.18% over SPEC CPU 2006. In this method, the memory system power is reduced by overall 9.62W by including benchmarks and on one side of the gcc when sorted by data transfer capacity (thus with a transmission capacity usually below 1,12GB / s per channel). This is important when the standard 1363MHz memory frame power is 65.1W. In an entire network environment, bandwidth (0.5, 2) gives a standard change in energy efficiency of 2.43 percent (maximum 5.23 percent). It is important to note that I control reduction and changes in energy efficiency as a result of DVFS memory will give CPU-based DVFS comparative power losses and (ii) reduction accompanies negligible performance decreases.

Fig. 4 seems by all accounts, to be troublesome for undertakings that arrangement a high throughput, perhaps illustrative the memory spent by messages holding up in line somewhat than spent by the assignment objective itself, not operational power. As such, the power the device expands paying little respect to which or what number of charges it is executing, subordinate just on its present state and recurrence. Current DDR3 device bolsters an assortment of shut down states to spare power when idle. So as to evaluate DRAM control amid a framework's execution, must measure the power use in each state, and after that ascertain normal foundation control weighted when spent in each state.

With a specific end goal to comprehend the possible for voltage scaling in open DDR3 device, tests performed on 8 DIMMs in proposed assessment framework while physically controlling the memory voltage controller yield. The outcomes appear in Fig. 5. At 1363, 1060, and 850MHz individually, watched normal least constant voltages of 1.173V, 1.203V and 1.280V separately, in addition, a greatest over the 8 DIMMs of 1.24V, 1.27V and 1.35V distinctly. In this fashion, gather the required supply voltage lessens with recurrence. Afterward, in alluded from Benkhelifa et al (2014), will display control diminishment by conservatively expecting Vdd of 1.35V, 1.425V and 1.5V. Least Stable Memory Voltage as a Function of Frequency Distribution is shown in Fig. 5. Take note of that these voltages are well over the base stable voltages for the tried DIMMs.

By fluctuating the memory frequency and thus the highest transfer capacity, the memory failure seen with customer workloads is modified. As memory idleness expands, an off-demand center is less ready to cover up the inertness and slow down time, resulting in lower execution. Eventually, this effect is based on the quality of the application. Nevertheless, as mentioned below, know the data workloads in SPEC CPU2006, which includes CPUs and severe memory applications. This is because a transmission capacity-based scaling device with legally defined edges is shifted to a greater recurrence, with a lower dormancy and a higher immersion level, when tailing delays are obviously remarkable.

The SPEC-CPU2006 operating memory at 850 MHz and 1060 MHz (corruption by 1363 MHz) is shown in Fig. 6. Standard usage of normal transfer speeds is sorted for benchmarks. After all, metrics with higher transfer pattern rate experience more execution corruption at lower memory rates, provided that the data transmission of the measuring instrument is equal to or beyond the lower recurrence of the pinnacle transfer rate. As illustrated in Fig. 6, dormance increases significantly as the most extreme data transmission methodologies for use. These benchmarks show that genuine data transfer capacity has decreased at lower memory frequencies due to memory throughput limitations. As the previous area became apparent, benchmarks with the lower data transfer demand are generally tolerant for lower memory recurrence. This can minimize normal power by reducing the memory recurrence if it has no impact; because the runtime is not impaired or because energy efficiency can improve in less favorable conditions.

To do so, simply choose a fixed edge to relay for all recurrence movements. The calculation of controllers takes place occasionally in established ages and measures the normal use of data transfer for the last age. In view of this calculation, it chooses the recurrence of the comparative memory. This pragmatism is the product of equation 1 of tests, two sides have to be indicated: the shift between 850 and 1060 MHz and a shift between 1060 MHz and 1363 MHz. Assess two edge settings: Bandwidth (0.5,1) moves to 1066 and 1363MHz at 0.5GB / s and 1.0GB / s per channel, separately, and Bandwidth (0.5,2) moves at 0.5GB / s and 2.0GB / s. Both thresholds are mild because they are under the knees in Fig. 5. Or perhaps, because of normal transmitting capacity estimates per differential, they are picked in Fig. 5. Such parameter choices have a negligible impact on the implementation of the proposed model performance.

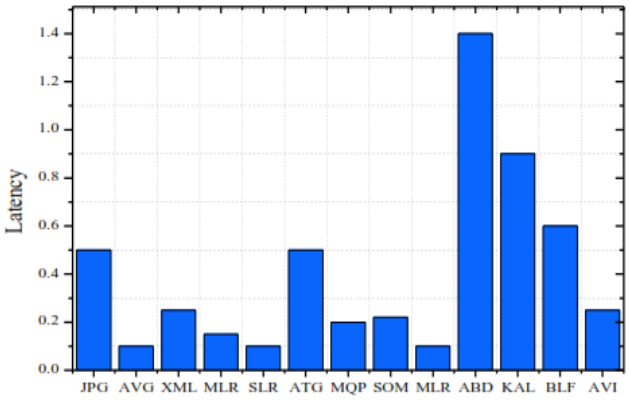


Fig. 4. Performance Analysis of IoT's Application Data Streams Applying MDVFS.

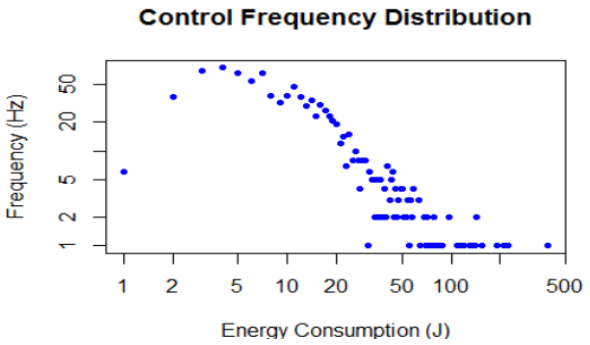


Fig. 5. Least Stable Memory Voltage as a Function of Frequency Distribution.

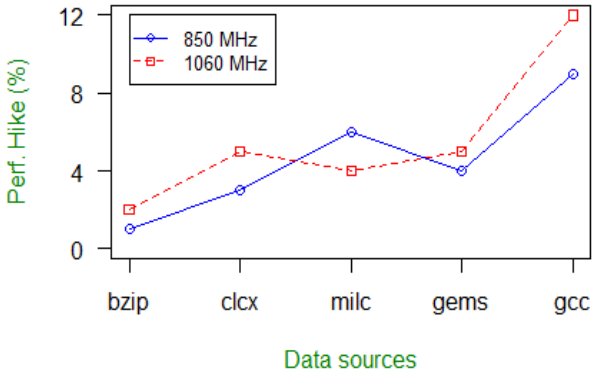


Fig. 6. Performance Varying as Per Memory Bandwidth Range from 1060MHz and 850MHz to Baseline 1363MHz.

V. Conclusion

The main barriers in memory recurrence scaling were presented in this approach and a fundamental assessment was made using a clear and natural calculation. More work remains to be completed, however. Initially there is a simple framework and a large plan space to determine and predict the impact of memory recurrence changes and to foresee the future effect.

In this paper, the proposed model is to evaluate the scaling of memory frequency / voltage to increase energy efficiency and reduce storage energy. Beginning with the observation that the frequency-dependent part of the memory system power has a control algorithm, which reduces the memory frequency and reduces the performance impacts. The important point is that the lowering of memory frequency does not significantly change memory access latency when bandwidth is used in low capacity. The proposed control algorithm increases the memory frequency when the usage reaches a threshold, reducing the output effect, by monitoring the memory bandwidth utilization. In this method, DVFS can be an efficient technology of energy efficiency, with particular when memory bandwidth is low.

References

[1] Fox R. Griffith A. Joseph R. Katz A. Konwinski G. Lee D. Patterson A. Rabkin I. Stoica "Above the clouds: A Berkeley view of cloud computing" Rep. UCB/EECS vol. 28 2009.

[2] Q. Zhang L. Cheng R. Boutaba Cloud computing: state-of-the-art and research challenges Journal of Internet Services and Applications vol. 1 pp. 7-18 2010.

[3] A. Thusoo J.S. Sarma N. Jain S. Zheng P. Chakka Z. Ning S. Antony L. Hao R. Murthy "Hivea petabyte scale data warehouse using Hadoop" Proceedings of IEEE 26th International Conference on Data Engineering (ICDE) pp. 996-1005 2010.

[4] M. Ezhilarasi and V. Krishnaveni , "An Optimal Solution To Minimize The Energy Consumption in Wireless Sensor Networks,"International Journal of Pure and Applied Mathematics", Vol. 119, Issue 10, (2018), pp. 829-844

[5] S. Chen Y. Sun D. Ulas K.L. Huang P. Sinha G. Liang X. Liu N. B. Shroff "When Queueing Meets Coding: Optimal-Latency Data Retrieving Scheme in Storage Clouds" IEEE INFOCOM pp. 1042-1050 2014.

[6] G. Liang U. Kozat FAST CLOUD: Pushing the Envelope on Delay Performance of Cloud Storage with Coding IEEE/ACM Trans. Networking vol. 22 no. 6 pp. 2012-2025 Nov 2013.

[7] S. L. Garfinkel "An evaluation of Amazons grid computing services: EC 2 S3 and SQS" in Tech. Rep. Harvard University 2007.

[8] V. Alieksieiev O. Gaiduchok "About the problem of data losses in real-time IoT based monitoring systems" Proceedings of International Scientific Conference "Mathematical Modeling" (Borovets Bulgaria December 13-16 2017) STUME "Industry 4.0" Sofia Bulgaria Year I vol. 1/1 pp. 10-11 2017.

[9] M. Ezhilarasi V. Krishnaveni "A Survey on Wireless Sensor Network: Energy and Lifetime Perspective" Taga Journal vol. 14 pp. 3099-3113 ISSN 1748-0345.

[10] Nagarajan, M. "A New Approach to Improve Life Time Using Energy Based Routing in Wireless Sensor Network. International Journal of Science and Research (IJSR), 2012

[11] V. Alieksieiev G. Ivasyk V. Pabyrivskyi N. Pabyrivska Big data aggregation algorithm for storing obsolete data" Proceedings of International Scientific Conference "High Technologies. Business. Society 2018" (Borovets Bulgaria March 1-15 2018) STUME "Industry 4.0" Sofia Bulgaria Year II iss. 1 (3) vol. I "High Technologies pp. 113-115 2018.

[12] N. V. Myasnikova M. P. Beresten M. P. Stroganov "Approximation of multi extremum functions and its applications to technical systems" Herald of higher education institutions. Volga region. Engineering sciences vol. 18 no. 2 pp. 113-119 2011.

[13] M. Nagarajan and S. Karthikeyan, "A New Approach to Increase the Life Time and Efficiency of Wireless Sensor Network", IEEE International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME), (2012), pp. 231-235

[14] H. Sundmaeker P. Guillemin P. Friess S. Woelfflé "Vision and challenges for realising the internet of things" Cluster of European Research Projects on the Internet of Things European Commision 2010.

[15] M Shanmugapriya, Nagarajan Munusamy, "An Approach to Increase Energy and Life Time Using Power Control Optimization in Wireless Sensor Networks", International Journal of Pure and Applied Mathematics, Volume 119 Issue.15, 2018, pp.1171-1181, ISSN: 1314-339

[16] M. Kovatsch S. Mayer B. Ostermaier "Moving application logic from the firmware to the cloud: Towards the thin server architecture for the internet of things" IMIS. IEEE pp. 751-756 2012.

[17] Ezhilarasi, M. & Krishnaveni, V. "An evolutionary multipath energy-efficient routing protocol (EMEER) for network lifetime enhancement in wireless sensor networks" Soft Computing, (2019). https://doi.org/10.1007/s00500-019-03928-1

[18] D. O. Olguin P. A. Gloor A. S. Pentland "Wearable sensors for pervasive healthcare management" Proc. IEEE. The 3rd International Conference on Pervasive Computing Technologies for Healthcare PervasiveHealth 2009 pp. 1-4 Apr. 2009.

[19] L. Gnanaprasanambikai and N. Munusamy, "Survey of genetic algorithm effectiveness in intrusion detection," 2017 International Conference on Intelligent Computing and Control (I2C2), Coimbatore, 2017, pp.1-5. doi: 10.1109/I2C2.2017.8321877

[20] R. C. Taylor "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics" BMC Bioinformatics vol. 11 no. 12 2016

Energy Efficient Cluster based Routing Protocol with Secure IDS for IoT Assisted Heterogeneous WSN

Sultan Alkhliwi
Assistant Professor at Faculty of Science
Northern Border University, Arar, Kingdom of Saudi Arabia

***Abstract*—Currently, wireless sensor networks (WSNs) and the Internet of Things (IoT) have become useful in a wide range of applications. The nodes in IoT assisted WSN commonly operate on restricted battery units, meaning energy efficiency is a major design issue. Clustering and route selection processes are commonly utilized energy-efficient techniques for WSN. Although several cluster-based routing approaches are available for homogeneous WSN, only a limited number of studies have focused on energy efficient heterogeneous WSN (HWSN). Moreover, security poses a major design issue in the HWSN. This paper introduces an energy efficient cluster-based routing protocol with a secure intrusion detection system in HWSN called EECRP-SID. The proposed EECRP-SID technique involves three main phases: cluster construction, optimal path selection, and intrusion detection. Initially, the type II fuzzy logic-based clustering (T2FC) technique with three input parameters are applied for cluster head (CH) selection. These parameters are residual energy level (REL), distance to the base station (DTBS), and node density (NDEN). In addition to CH selection, the salp swarm optimization (SSO) technique is utilized to select optimal paths for inter cluster data transmission, which results in energy efficient HWSN. Finally, to achieve security in cluster based WSN, an effective intrusion detection system (IDS) using long short-term memory (LSTM) is executed on the CHs to identify the presence of intruders in the network. The EECRP-SID method was implemented in MATLAB, and experimental outcomes indicate that it outperformed the compared methods in terms of distinct performance measures.**

Keywords*—Wireless Sensor Networks (WSN); Clustering; Routing; Type II fuzzy logic; salp swarm algorithm; long short-term memory (LSTM)

I. Introduction

Progressive development of the Internet of Things (IoT) and wireless sensor networks (WSNs) has proven beneficial for a broader set of real-time data gathering and tracking applications. Further, WSN has gained considerable attention from developers of various applications such as automated irrigation organization, target monitoring, landslide tracking, clinical observation, forest fire prediction, and disaster management. The major benefit of WSN is that it is embedded with a massive number of effective sensor nodes (SN), which help to monitor climatic disasters emerging in remote or harsh regions. Then, the SN also investigates atmospheric factors such as temperature, pressure, humidity, moisture content, and sound that represent the catastrophic symptoms. Once the sensing operation is completed by SN, the details are then collected and sent to the base station (BS). The sensor unit and data communication unit in SN consume maximum energy which is an energy utilization module. Fig. 1 shows the structure of WSN. When all the energy is depleted, then it is considered expired or unable to process. A node is only referred to as dead when it is not suitable for replacement or it cannot be recharged by any other power source. Therefore, it is more essential to balance the power utilized by SN.

To resolve these issues, the clustering technique has been applied by many developers [1]. This is one of the best topologies for accomplishing an extended network duration and excellent efficiency. Moreover, it helps to preserve the power of SN under the development of several reliable clusters. The number of clusters developed might be either temporary or permanent, depending on the clustering mechanism used. In addition, clustering also segregates the jointly placed nodes into clusters. This distribution is determined according to the similarity metrics such as distance from the BS (DBS), transmission radius, and cluster density. Once the clusters are developed, a node inside a cluster is chosen as a head node termed the cluster head (CH). The CH has the responsibility of organizing the data collected from cluster members (CM) and forward them to the BS node. For instance, if a single cluster has a single node, then it is mandatory to communicate with the sink node rather than the SN communicating with the BS. Finally, it is applicable to reduce power utilization.

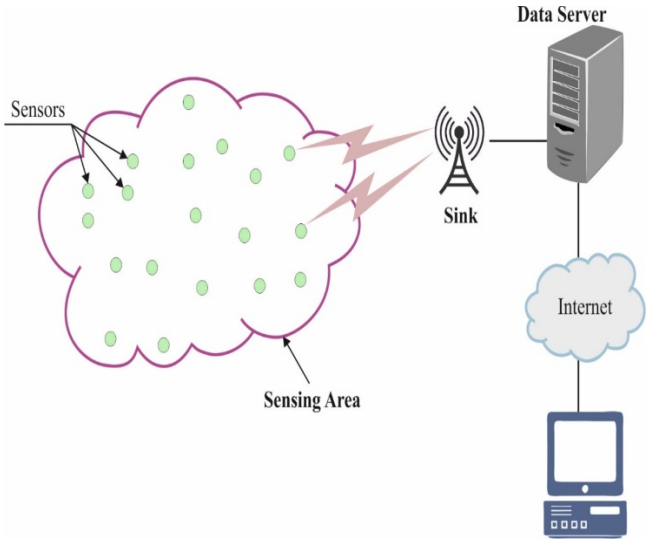


Fig. 1. Architecture of WSN.

Some of the classical (yet ineffective) routing protocols are direct transmission and minimum transmission energy types, which are suitable for preserving the energy of SN [2]. This problem is resolved with the help of clustering and cluster-related hierarchical routing protocol, such as low energy adaptive clustering hierarchy (LEACH) [2]. This is an effective model that ensures efficiency over previous protocols with the objective of energy preservation. Basically, it is computed for an identical system where the nodes have similar abilities, such as power, processing, and coverage sources. Consecutively, numerous protocols similar to LEACH have been presented to enhance the network lifetime significantly. Even though LEACH is a homogeneous protocol, the similarity of SN does not exist because of factors including manufacturing variations, diverse morphological aspects, and non-uniform physical terrain [3].

When the nodes have various configurations by means of classical factors, nodes are referred to as heterogeneous nodes and the network is framed as a heterogeneous network. Between the genres of heterogeneity, energy heterogeneity is highly applied because of its frequent dependence on non-replaceable batteries [4]. In the case of cluster-based routing, CH has been selected from all clusters by considering a few variables. When the CH collects from a cluster, then data aggregation is carried out to eliminate data redundancy. However, the CH experiences massive power exhaustion when compared to CM. Next, information gathered from CM is transmitted to BS using either single or multi-hop data transmission forwarded to users by internet sources [5, 6]. Developers have presented research principles on optimizing the CH election in homogeneous networks with the help of meta-heuristic approaches [7]. Thus, in heterogeneous WSN (HWSN), CH election is accomplished by using tactical amendments in the threshold-relied function used for the nodes.

The aim of the optimization module for CH election is to reduce power utilization. Further, optimized CH election for energy effective routing is assumed to be a non-deterministic polynomial-time hard (NP-hard) issue. Regardless of this, CH election is optimized using meta-heuristics that are embedded with major fitness functions [8]. At this point, the meta-heuristic model has been employed because of these features to convert the best solution. Thus, multitudinous optimization routing principles have been deployed to optimize CH selection and to ensure the scalability of a system that transmits data [9]. Security is another concurrent problem in HWSN, which can be tackled by using an intrusion detection system (IDS) [10]. This is mainly employed for tracking abnormal events that take place in the network. To attain effective performance in the IDS, machine learning (ML) and deep learning (DL) approaches have been found useful. DL is a modified form of ML, which is used to extract the features proficiently. Accordingly, the classical ANN is employed for handing nonlinear cases and DL methodologies are utilized for the extraction of features and to make decisions in a similar way to the human brain.

The contribution of this paper is defined as follows. An energy efficient cluster-based routing technique (EECRP-SID) with a secure intrusion detection system is introduced in HWSN. Initially, the type-II fuzzy logic clustering (T2FC) technique with three input parameters is applied for CH selection. These parameters are residual energy level (REL), distance to base station (DTBS), and node density (NDEN). In addition to CH selection, the salp swarm optimization (SSO) technique is employed to select optimal paths for inter cluster data transmission, which results in energy efficient HWSN. Finally, in cluster based WSN, a new IDS using long short-term memory (LSTM) is executed on the CHs to classify the existence of intruders in the network. The EECRP-SID method is implemented in MATLAB. The experimental outcomes demonstrated that the EECRP-SID model outperformed the compared methods in terms of distinct performance measures.

II. Related Works

Recently developed work is relevant to the newly deployed approach, as defined in the following. In [11,21], a hybridized grey wolf and crow search method depending on optimal CH selection (HGWCSOA-OCHS) was developed and used for enhancing network lifespan. This was achieved by reducing delays, distance between nodes, and energy utilization. The hybrid GWO and CSO approach in CH election maintains the trade-off between exploitation and exploration in a searching area. Moreover, a novel technique called PSO-based selection (PSOBS) for selecting optimal rendezvous points has been proposed [12]. Using PSO, the developed approach was used to find the best rendezvous points to accomplish remarkable network resources. Additionally, a weighted score was estimated for SN according to the data packets obtained from various sensors.

In Edla et al. [13], a clustering method was presented using shuffled complex evolution of PSO (SCE-PSO), which is an effectual FF using average cluster distance, CHs load, and number of loaded CHs in a system. The approach in [14] focused on CH selection using GA and KH methods for WSN. The key objective of this model was to enhance the lifespan of WSN routing with the BS node by designing an effectual routing approach on the basis of hybridized metaheuristic optimization models [15]. This originated as an initial framework with innovative and global searching capabilities of particle swarm optimization (PSO), diverse operator of differential approach, and pheromones of ant-colony optimization (ACO) technology to limit local searches and retain population diversity.

In Dattatraya and Rao [16], a new CH election process was developed to extend network duration and energy efficiency using fitness based glowworm swarm and the fruitfly algorithm (FGF) for WSN. In [17], an energy effective CH was selected based on the whale optimization algorithm (WOA), termed WOA-Clustering (WOA-C) was projected. Then, newly presented technology was used in the election of energy-aware CHs according to the FF, which applied the RE of a node and summation of the power of adjacent nodes.

III. Proposed Model

The proposed EECRP-SID technique involves three main phases: cluster construction, optimal path selection, and intrusion detection.

A. Clustering Methodology

The T2FC technique incorporates two stages: the CH selection and cluster formation phases [18, 22]. First, the BS node used the T2FC model to select appropriate CHs to accomplish even load distribution. Hence, the selected CHs are readily available for developing clusters and adjacent nodes. To select CH and cluster size, fuzzy logic and three input features (REL, DTBS, and NDEN) were employed. When compared with all other sources, energy is one of the most essential resources required by WSN. The process involved in T2FL is depicted in Fig. 2. Maximum power is consumed by CH (compared to CM) when computing certain operations such as aggregation, processing, and routing data. Here, REL is measured with the following function:

$$\mathrm{Er} = \mathrm{E0} - \mathrm{Ec}, \tag{1}$$

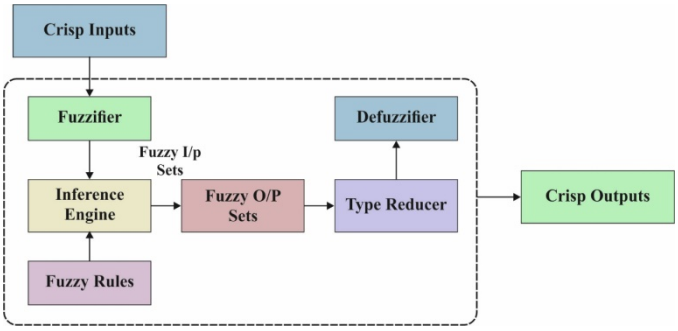


Fig. 2. Block Diagram of T2FL.

where E0 and Ec are the energy and power consumed by a node, respectively, and Er denotes the REL of the regular node.

The transmitted message utilizes power, which is directly proportional to the square of the distance over the candidate as well as the source node. Then, DTBS is evaluated as follows:

$$C = \frac{d_{avg}^2}{d_0^2} \tag{2}$$

Here, d_{avg} represents the higher distance between node and neighbours and d0 implies the broadcasting radius of a node.

The NDEN implies that closer neighbours develop a better score and a node with higher probability is considered as CH. The NDEN is measured using the following expression:

$$NDEN = \frac{|D_i - D_0|}{D_0} \tag{3}$$

Here, Di is the count of neighbours and D0 is the optimal number of neighbours. In this approach, T2FL is composed of the following four principles:

- Fuzzifier

The main aim of the fuzzifier is to convert the exact input as fuzzified measures. The input parameters with linguistic variables for CH selection is then transformed.

- Fuzzy Inference System

The performance of T1FL and T2FL is homogeneous. In this module, a set of 27 rules are used. A sample rule for CH election is depicted in Eq. (4).

$$Rule(i) IF x1\ is A_1(i) AND x2\ is A_2(i) AND x3\ is A_3(i)$$
$$THEN y1\ is B_1(i) AND y2\ is B_2(i) \tag{4}$$

Here, *i* refers to the i^{th} rule in fuzzy rule, while A_1, A_2, and A_3 are the equivalent fuzzy set of x_1, x_2, and x_3.

- Membership functions

Here, T2FL is defined by effective and poor membership functions (MFs). These functions are implied as T1FL MF. The interval between the two functions refers to the footprint of uncertainty (FOU) that describes a T2FL set. Consider the FOU is represented as *f*. If $f \in [0, 1]$, and $f \to 0$, then MF is implied as T1FL. If $f \to 0$ to 1, then T2FL comprises a wider range of FOU within [0, 1]. Therefore, the rule development in T2FL logic is similar to T1FL, as given in the following:

$$Type\ 2\ FL = Principal\ MF\ (Type\ 1\ FL) + FOU \tag{5}$$

- Type reducer/ Defuzzifier

The type-reducer produces a T1FL outcome, which is transformed as a mathematical result after the execution of defuzzifier. The node gathers PCH and cluster size, where it broadcasts a message to closer nodes, where the message is composed of Node-ID and the value of PCH. A node with a high possibility is treated as CH and sends CH_WON to adjacent nodes. A node accomplishes numerous CH_WONs from neighbouring nodes, then sends the CH_JOIN message and unifies closer CHs. Subsequently, to eliminate the premature death of CH, CH rotation is carried out. Further, when the RE of CH goes beyond the threshold measure, then CH rotation is performed. This mechanism is useful for eliminating premature death of CH and tends to enhance network lifetime.

B. Route Selection Algorithm

Developing a numerical approach that reflects the smart behaviour of a swarm is one of the fundamental steps in SI-reliant models for resolving the optimization issues. Numerical methods in addition to fish swarm, birds swarm, and ant swarm were employed in optimization issues. To model the salp chain, the individuals present in salps swarms were classified into leader and followers [19,23]. Initially, the leader is the important individual among the salps chain for computing the direction and foraging the route of a swarm and assisting the salp chain to identify food. The residual salps are considered as followers, who obey the leader to develop a chain-like structure. Therefore, while a numerical approach accelerates salps chain production, it is not suitable for resolving the optimization issues. Computing the global optimal value is one of the main aims in the optimization issues, where the global optimal score has been employed as food required by salps chain. Based on leader position, the salps chain is placed near to the food, which is expressed as follows:

$$X_j^1 = \begin{cases} F_j - c_1[(ub_j - lb_j)c_2 + lb_j]\ if\ c_3 < 0.5 \\ F_j + c_1[(ub_j - lb_j)c_2 + lb_j]\ if\ c_3 \geq 0.5 \end{cases} \tag{6}$$

where X_j^1 denotes the location of a leader in the j-th dimension; F_j mimics the place of food in the j-th dimension; ub_j defines the lower bound of the j-th dimension; and lb_j

refers the upper bound of the j-th dimension. These upper and lower bounds were employed to minimise excess searching space. Parameter c_2 denotes a random value from [0.1] that has been applied to control leader movement. Parameter c_3 means a random value from [0.1] that has been employed uniformly, regardless of whether the leader is moving towards or away from the food location. Parameter c_1 is depicted in Eq. (7), which is an extension factor applied for managing the global exploration and local exploitation.

$$C_1 = 2e^{-(4t/T)^2} \tag{7}$$

Here, t shows the recent count of iterations and T denotes the overall count of iterations. From Eq. (2), the adjustment factor c_1 would reduce gradually for every iteration. To make the followers obey the instructions of a leader to make a chain, Newton's law of motion was applied to upgrade the place of followers, defined as follows:

$$X_j^i = \frac{1}{2} a \cdot t^2 + v_0 \cdot t \tag{8}$$

where X_j^i is the place of the i-th follower in the j-th dimension if $i \geq 2; t$ implies time; v_0 mimics the first speed, and simulation of the follower's direction $a = v_{final}/v_0$. The speed of a follower is defined by $v = (x - x_0)/t$. Subsequently, the time variable of the optimization issues is depicted as an iteration count; thus, the iteration interval is from $t = 1$. The follower's basic speed is $v_0 = 0$. Eq (8) is improvised as follows:

$$X_j^i = \frac{1}{2}\left(X_j^i + X_j^{i-1}\right) \tag{9}$$

In the case of followers, a better position might be reached instead of using a recent solution (food). Then, the food is interchanged by an optimal location and upgraded by the leader to move in the direction of the food. The benefits of SSA are given in the following:

- It can be acquired from the numerical approach of SSA with simple structure and can be trained using parameters and operators, which makes the implementation easier.
- In optimization, SSA applies the best solution in the present iteration as food. Even the fitness values are applicable in a complete population that affects the quality of the food.
- Leaders are capable of exploring and moving closer, according to the position of the food. The followers have to move in a chain format, which demonstrates the model's simplicity. Under the consideration of benefits, the demerits are also defined in SSA.

The principle behind the best SSA is defined in the following:

Step 1. Allocate the algorithm variables: No. of iterations T, No. of ascidian populations N, test function dimension D.

Step 2. Allocate the salp population based on the maximum and minimum bounds, $t = 1$.

Step 3. Estimate the fitness value of a searching individual, and consider them with optimal fitness value in recent population as food F_j.

Step 4. Upgrade c_1 on the basis of Eq. (2) and produce random values c_2 and c_3.

Step 5. When $i = 1$, upgrade the leader's place on the basis of Eq.(1). When $i \geq 1$, upgrade the follower's place based on the Eq. (4). $t = t + 1$.

Step 6. Evaluate whether a method has attained a higher count of iterations. When the procedure is terminated, then the optimal value is provided back; else, repeat Step 3.

A flowchart of SSA is illustrated in Fig. 3. The task of finding an effectual path using SSA is explained in the following:

- Every CH sends the route information (such as node-ID, RE, and distance) from closer CHs. Thus, CH saves the details about a routing table.
- A route from CHs to BS is computed using a salp suited in the CHs occasionally. The fitness function is derived as

$$FF_{ij} = \frac{1}{d_{ij}} \tag{10}$$

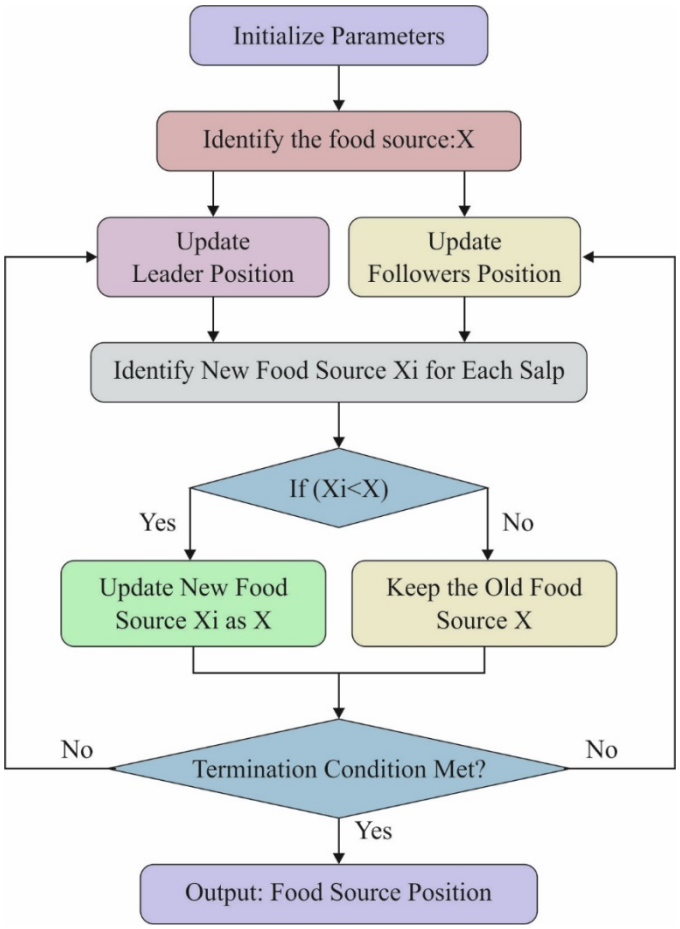


Fig. 3. Flowchart of SSA.

where d_{ij} denotes a distance between source node i to equivalent CH. The distance d_{ij} can be evaluated as a Euclidian distance, as in Eq. (11).

$$d_{ij} = \sqrt{(S(i).xd - s(j)xd)^2 - (S(i).yd - s(j)yd)^2} \quad (11)$$

The energy level of nodes is given by

$$E = \frac{E_0 - E_{residual}}{\sum_{k \in N} E_k} \quad (12)$$

where E_0 signifies an initial power and $E_{residual}$ refers to the remaining energy of a node. Terms α and β are parameters employed for regulation.

- During data transmission, a node with a higher probability is selected as a relay node from the CHs to BS.

C. LSTM based IDS

To determine the existence of intrusions in the network, the LSTM model is employed [20,24]. LSTM is a special extension of the recurrent neural network (RNN) structure, which is a DL model. It is composed of feedback connections used for eliminating long term dependencies, and is capable of computing the data points and a complete series of knowledge. For example, LSTM can be operated on un-segmented data, correlated recognition patterns, audio recognition, and abnormal prediction in network traffic or IDSs (intrusion detection systems). A typical LSTM unit is composed of four major sections: cell, input gate, output gate, and forget gate. Initially, cell memories over rare time intervals, and hence, 3 gates are standardized the data flow into and out of a cell.

LSTM systems are particularly suitable for classifying, computing, and developing predictions according to time-sequential data, because they suffer from scarce duration among significant events in a time series. Initially, the LSTM is deployed for affecting the diminishing gradient issues that identify classical RNN. The relative insensitivity to gap length is a combination of LSTM with RNNs, hidden Markov models (HMMs), and alternate series learning models in massive domains. An LSTM with four stages in a cell is illustrated in Fig. 4. In a cell state, the horizontal topmost line is implemented in the cell, which refers to the cell state. Then, the LSTM is not capable of eliminating the data from a cell state, which is carefully regularized by gates. A cell is composed of massive structures.

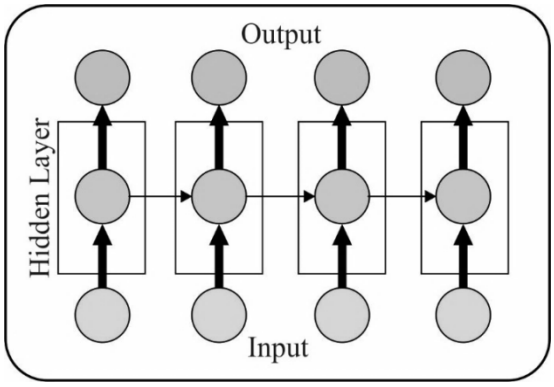


Fig. 4. Structure of LSTM Cell.

Step 1: The Initial step in LSTM is to find the unwanted decision made by a distant cell state. This is developed by a sigmoid layer (forget gate layer) as follows:

$$f_t = \sigma\left(W_f . [h_{t-1}, x_t] + b_f\right) \quad (13)$$

where h_{t-1} implies the result from the existing time stamp, $x_t = new$ input, and $b_f =$ bais.

Step 2: Next, a decision is saved in a cell state. This is composed of two parts. Initially, a sigmoid layer (input gate layer) that selects upgrading values and a $\tanh$ layer is applied to make a vector of novel contender measures, which is included in the state.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (14)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_c) \quad (15)$$

Step 3: The previous cell state is updated, C(t-l), as a novel cell state. Then, steps are selected for performing some events. Initially, the previous state is increased by the removal of images, then it is included in the cell state. This is referred to as novel candidate measures, which is scaled by a proportion for deciding a state update in all measures.

$$C_t = f_t * C_t + i_t * \tilde{C}_t \quad (16)$$

Step 4: This is the final stage, and the output is supported by a cell state that is an extended version. First, a sigmoid layer is implemented for selecting the portions of a cell state as a result. Then, the cell state is fixed by $\tanh$ and increases the final result of a sigmoid gate; thus, the consequent portions.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (17)$$

$$h_t = o_t * \tanh(C_t) \quad (18)$$

IV. Performance Validation

An analysis of the results of the EECRP-SID model was conducted in terms of energy efficiency with varying rounds, as shown in Fig. 5. The energy should deplete at a slower rate with an increased number of rounds. The figure shows that the MS-GAOC model exhibited worse outcomes and depleted all its energy at a faster rate compared to existing techniques. Next to that, the IABC model tried to provide better results over the MS-GAOC model by exhausting its energy at a slightly reduced rate. Moreover, the KHA algorithm accomplished somewhat satisfactory performance by attaining energy exhaustion at a moderate rate. Further, the MO-PSO algorithm reached near optimal results by attaining energy depletion at a minimum rate. However, the EECRP-SID model demonstrated superior energy efficiency by depleting the energy at a slower rate.

An average delay analysis of the EECRP-SID model was conducted and compared with the other methods under a varying node count, as illustrated in Fig. 6. As shown, the MO-PSO algorithm was found to have poor performance and resulted in a higher average delay, tending to increase significantly with an increase in the number of nodes. Next, the KHA algorithm required a somewhat less average delay compared to the MO-PSO model, but not compared to the other methods. The IABC model achieved reasonable results

with a moderate average delay, while the MS-GAOC model demonstrated competitive results by requiring a low average delay. However, the proposed EECRP-SID model only required a minimum amount of average delay.

Fig. 7 shows the PDR results analysis of the EECRP-SID model and existing techniques under a distinct node count. The figure demonstrates that the MO-PSO algorithm achieved an inferior outcome with the lowest amount of PDR. Subsequently, the MS-GAOC model surpassed the MO-PSO model by attaining slightly better PDR (but not compared to the other methods). The IABC model achieved an even higher PDR, whereas a near-optimal PDR value was achieved by the MS-GAOC model. However, the proposed EECRP-SID model proved to be superior in the clustering and routing processes by obtaining a maximum PDR.

In Fig. 8, the network lifetime analysis of the EECRP-SID model is presented and compared with existing techniques in terms of alive node count. From the figure, it can be noted that the MS-GAOC model attained a maximum number of dead nodes with an increasing number of nodes. This implies that it holds a lower number of alive nodes. Further, the IABC model achieved a somewhat better lifetime by offering a slightly higher number of alive nodes. The KHA model achieved a somewhat acceptable network lifetime with moderate alive node count. Although the MO-PSO algorithm exhibited near-optimal alive node count, superior network lifetime was demonstrated by the EECRP-SID model.

Next, the intrusion detection performance of the EECRP-SID model was evaluated using the KDD Cup 1999 dataset, in terms of different evaluation parameters (as shown in Fig. 9 and 10). The results indicate that the RF model resulted in a minimum sensitivity of 92.39%, specificity of 93.83%, F-score of 93.58%, and kappa of 85.99%. This was followed by the RBF Network, which achieved a somewhat higher sensitivity of 93.4%, specificity of 92.38%, F-score of 93.38%, and kappa of 85.79%. The RT model displayed a moderate outcome with a sensitivity of 95.68%, specificity of 95.39%, F-score of 95.84%, and kappa of 91.06%. The DT model exhibited an even better classifier outcome, with a sensitivity of 95.68%, specificity of 95.37%, F-score of 95.83%, and kappa of 91.03%. In addition, the LR achieved near optimal intrusion detection, with the highest sensitivity of 97.26%, specificity of 96.92%, F-score of 97.29%, and kappa of 94.19%. However, the proposed LSTM-IDS model achieved a higher sensitivity of 97.68%, specificity of 98.54%, F-score of 98.13%, and kappa of 96.21%.

Fig. 10 illustrates an accuracy analysis of the different IDSs on the applied test dataset. The resultant values meant that the CSA-PSO model demonstrated ineffective detection results, with the lowest accuracy of 75.51%. The GBT-IDS model achieved an accuracy of 84.25%, while the RBFNetwork and RF models achieved near identical accuracies of 92.93% and 93.04%, respectively. The FCM-IDS, DT, and RT models exhibited closer accuracy of 95.3%, 95.53%, and 95.55%, respectively. Although the existing CSA-IDS and LR models demonstrated competitive intrusion detection with accuracies of 96.88% and 97%, respectively, the presented LSTM-IDS model achieved the highest accuracy of 98.43%. These results verify the effectiveness of the proposed methods with respect to energy efficiency, lifetime, and intrusion detection.

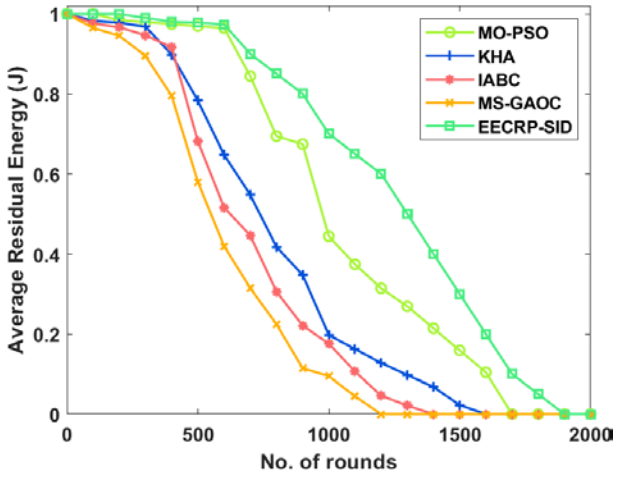


Fig. 5. Average Residual Energy Analysis of the EECRP-SID Model Compared with Existing Techniques.

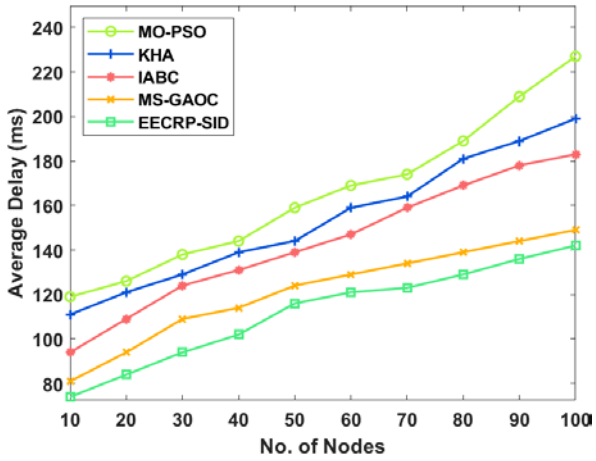


Fig. 6. Average Delay Analysis of the EECRP-SID Model Compared with Existing Techniques.

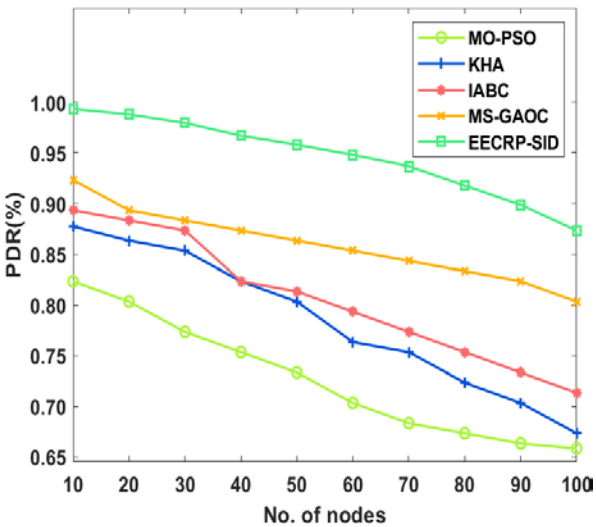


Fig. 7. PDR Analysis of the EECRP-SID Model Compared with Existing Techniques.

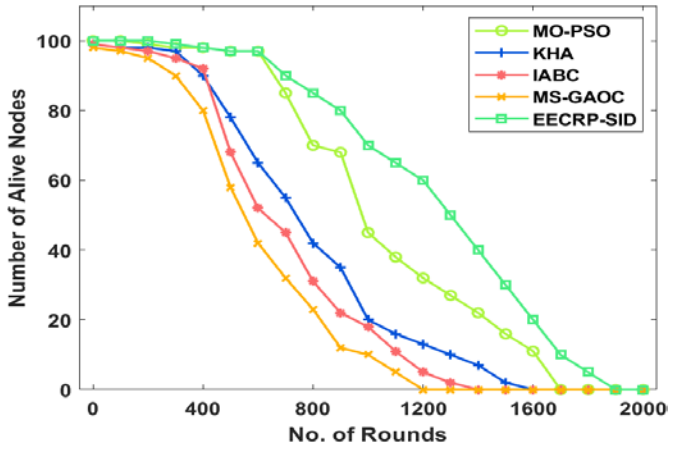


Fig. 8. Lifetime Analysis of the EECRP-SID Model Compared with Existing Techniques.

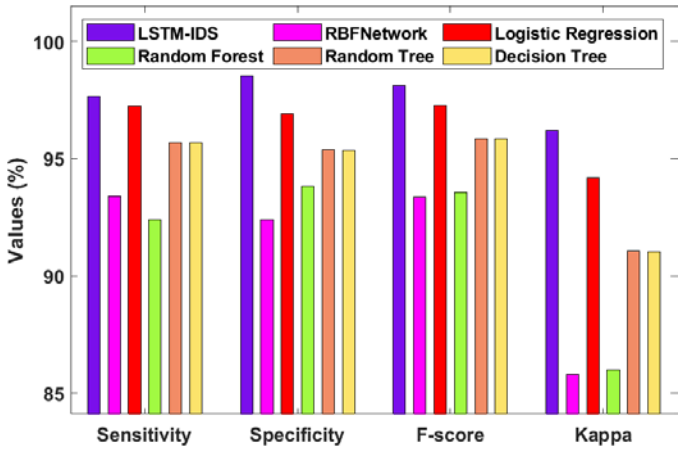


Fig. 9. Intrusion Detection Analysis of the LSTM-IDS Compared with Existing Techniques.

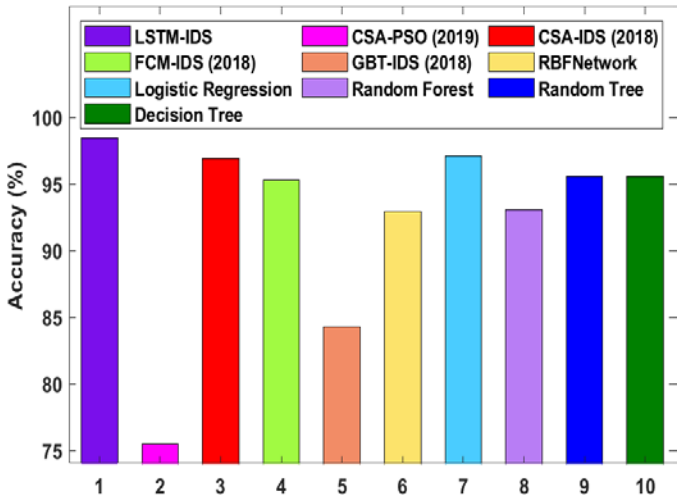


Fig. 10. Accuracy Analysis of the LSTM-IDS Compared with Existing Techniques.

V. Conclusion

This paper presents the development of a novel EECRP-SI model (EECRP-SID) for energy-efficient cluster-based routing protocol with secure IDS in HWSN. The proposed EECRP-SID technique involves three main phases: cluster construction, optimal path selection, and intrusion detection. First, the T2FC technique is applied to form clusters and elect proficient CHs. Next, an SSA based inter clustering routing task is employed. Finally, to determine the existence of intrusions in the network, the LSTM model is used. During intrusion detection in the HWSN process, the presented LSTM-IDS model achieved the highest accuracy of 98.43%. The experimental results analysis of the EECRP-SID model showcased superior results in terms of energy efficiency, lifetime, and intrusion detection. In the future, the performance of intrusion detection can be improved by the hybridization of metaheuristic algorithms for tuning the hyperparameters.

References

[1] Sharma, R., Vashisht, V., Singh, U., " EEFCM-DE: energy efficient clustering based on fuzzy C means and differential evolution algorithm in wireless sensor networks," IET Commun. 13 (8), 996–1007,2019b

[2] W.R. Heinzelman, A. Chandrakasan, H. Balakrishnan, Energy-efficient communication protocol for wireless microsensor networks, in: Proc. of 33rd Annu. Int. Conf. on Syst. Sci. Hawaii, IEEE, 2000.

[3] V. Mhatre, C. Rosenberg, Homogeneous vs heterogeneous clustered sensor networks: a comparative study, in: Int. Conf. on Commun., IEEE, pp. 3646–3651, 2004.

[4] M. Yarvis, N. Kushalnagar, H. Singh, A. Rangarajan, Y. Liu, S. Singh, Exploiting heterogeneity in sensor networks, in: Proc. of 24th Annu. Jt. Conf. on Comput. Commun. Soc., INFOCOM 2005, IEEE, pp. 878–890,2005

[5] J. Huang, D. Ruan, Y. Hong, Z. Zhao, H. Zheng, IMHRP: Improved multi-hop routing protocol for wireless sensor networks, in: J. Phys. Conf. Ser, IOP Publishing, p. 012054. 2017

[6] S. Kumar, P. Ranjan, R. Ramaswami, M.R. Tripathy, Resource efficient clustering and next hop knowledge based routing in multiple heterogeneous wireless sensor networks, Int. J. Grid High Perform. Comput. 9, 1–20, 2017

[7] A.A. Bara'a, E.A. Khalil, A new evolutionary based routing protocol for clustered heterogeneous wireless sensor networks, Appl. Soft Comput. 12, 1950–1957, 2012

[8] C.-W. Tsai, T.-P. Hong, G.-N. Shiu, Metaheuristics for the lifetime of WSN: A review, IEEE Sens. J. 16 , 2812–2831, 2016

[9] B.P. Deosarkar, N.S. Yadav, R.P. Yadav, Clusterhead selection in clustering algorithms for wireless sensor networks: A survey, in: Int. Conf. on Comput. Commun. Netw., ICCCn, IEEE, pp. 1–8, 2008.

[10] Butun, I., Morgera, S.D. and Sankar, R., A survey of intrusion detection systems in wireless sensor networks. IEEE communications surveys & tutorials, 16(1), pp.266-282, 2013

[11] Subramanian, P., Sahayaraj, J.M., Senthilkumar, S. and Alex, D.S., A Hybrid Grey Wolf and Crow Search Optimization Algorithm-Based Optimal Cluster Head Selection Scheme for Wireless Sensor Networks. Wireless Personal Communications, pp.1-21, 2020

[12] Tabibi, S. and Ghaffari, A., Energy-efficient routing mechanism for mobile sink in wireless sensor networks using particle swarm optimization algorithm. Wireless Personal Communications, 104(1), pp.199-216, 2019

[13] Edla, D.R., Kongara, M.C. and Cheruku, R., SCE-PSO based clustering approach for load balancing of gateways in wireless sensor networks. Wireless Networks, 25(3), pp.1067-1081, 2019.

[14] Karthick, P.T. and Palanisamy, C., Optimized cluster head selection using krill herd algorithm for wireless sensor network. Automatika, 60(3), pp.340-348, 2019

[15] Wang, H., Li, K. and Pedrycz, W., An Elite Hybrid Metaheuristic Optimization Algorithm for Maximizing Wireless Sensor Networks Lifetime With a Sink Node. IEEE Sensors Journal, 20(10), pp.5634-5649, 2020

[16] Dattatraya, K.N. and Rao, K.R., Hybrid based cluster head selection for maximizing network lifetime and energy efficiency in WSN. Journal of King Saud University-Computer and Information Sciences, 2019

[17] Jadhav, A.R. and Shankar, T., Whale optimization based energy-efficient cluster head selection algorithm for wireless sensor networks. arXiv preprint arXiv:1711.09389, 2017

[18] Nayak, P. and Vathasavai, B., Energy efficient clustering algorithm for multi-hop wireless sensor network using type-2 fuzzy logic. IEEE Sensors Journal, 17(14), pp.4492-4499, 2017

[19] Mirjalili, S., Gandomi, A.H., Mirjalili, S.Z., Saremi, S., Faris, H. and Mirjalili, S.M., 2017. Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. Advances in Engineering Software, 114, pp.163-191, 2017

[20] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H. and Xu, B., August. Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) , 207-212, 2016

[21] Alghamdi, T.A. Energy efficient protocol in wireless sensor network: optimized cluster head selection model. Telecommun Syst **74,** 331–345 (2020). https://doi.org/10.1007/s11235-020-00659-9

[22] Devaraj, A.F.S. Energy aware reliable route selection scheme with clustered RP model for wireless sensor networks to promote interaction between human and sensors. J Ambient Intell Human Comput (2020). https://doi.org/10.1007/s12652-020-02147-z

[23] B.K. Patle, Ganesh Babu L, Anish Pandey, D.R.K. Parhi, A. Jagadeesh, A review: On path planning strategies for navigation of mobile robot, Defence Technology,Volume 15, Issue 4,2019,Pages 582-606,ISSN 2214-9147

[24] Jesús Gonzalez, Wen Yu, Non-linear system modeling using LSTM neural networks, IFAC-PapersOnLine,Volume 51, Issue 13,2018,Pages 485-489, ISSN 2405-8963,

Moment Features based Violence Action Detection using Optical Flow

A F M Saifuddin Saif[1], Zainal Rasyid Mahayuddin[2]
Faculty of Science and Technology, American International University–Bangladesh, Dhaka, Bangladesh
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi, Selangor, Malaysia

***Abstract*—Instantaneous detection of violence is still an unsolved research problem although artificial intelligence lives its prosperous years. The severity of injury causes due to violence can be minimized by detecting violence in real time demands for effective violence detection. Various methods were previously proposed for violence detection which could not provide robust results due many challenges, i.e. noise, motion estimation, lack of appropriate feature selection, lack of effective classification approach, complex background and variations in illumination. This research proposes an efficient method for violence detection using moment features to use motion patterns to facilitate detection in each frame and provides smaller area as region of interest. This means probability for extraction of motion intensity is getting lost because of same colored object in the background is reduced and thus minimizes background complexity. After that, proposed method uses optical flow to calculate angles and linear distances in each frame. In this context, if there is any frame loss due to noise or illumination variation, proposed method uses Kalman filter to process that frame by illuminating noise. Finally, decision for violence is determined using random forest classifier from single feature vector by generating a set of probabilities for each class. Proposed research performed extensive experimentation where accuracy rate of 99.12% was achieved using frame rate of 35 fps which is higher comparing with previous research results. Experimental results reveal the effectiveness of the proposed methodology.**

***Keywords*—Violence detection; feature extraction; classification; optical flow**

I. Introduction

Surveillance applications have been used to monitor public and private areas where intelligent violence detection is still an unsolved research problem. Surveillance system with violence detection capability can be used to monitor various places, i.e. airports [1], football matches and protests [2], parks [3], stadium [4], internet video filtration [5], markets [6], Government offices [7]. Previous researchers proposed various methods to provide intelligent violence detection which could not provide satisfactory results due to many challenges, i.e. lack of efficient feature extractions [8, 9], lack of appropriate segmentation method [10], motion estimation [11, 12], variation of brightness or illumination [13], cluttered backgrounds [14], scene complexities [15], low level image [16], lack of efficient noise reduction method [17], occlusions [18, 19]. Proposed method by this research used motion intensity characteristics by extracting moment features to facilitate violence detection in each frame effectively.

Till now many researchers have done their researches regarding violence detection. Research in [20] used convolutional neural network (C3D) and CNN with long short-term memory (CNN-LSTM) through shallow neural network to learn high level spatial-temporal information from raw image data for violence detection. However, their proposed method was suitable for still images. Research in [21] used set of selectively distributed frames and spatio-temporal features using both space and time dimensions provided to fully connected neural framework to classify violence or non-violence action which required further validation in terms with complexity. Research in [22] distinguished physical violence by designing DT–SVM (Decision Tree–SVM) two-layer classifier. However, their overall methodology requires further improvement towards more complex scenes with both nearby and distant objects. In in [23] modeled crowd dynamics using temporal summaries of grey level co-occurrence matrix (GLCM) features. However, improvement towards adaptive selection of optimal parameters based on given data was required in their proposed methodology.

This research performs moment features extraction to facilitate violence detection as the basis of motion pattern. In this context, linear distances and angles are calculated using optical flow. Besides, if there is any frame loss due to noise or illumination variation, proposed method uses Kalman filter to consider that frame for further processing by illuminating noise. Overall contributions by this research are stated below:

- Proposed method uses moment features by implicating weighted average of pictorial intensities to use motion patterns for facilitating detection in each frame in lieu with reducing background complexity and provides smaller area as region of interest to reduce overall computation time per frame.
- After calculating linear distances and angles as the basis of optical flow, proposed method uses Kalman filter to rectify frame loss due to nose or illumination variation which plays significant role to optimally estimate distance and angles for higher accuracy.
- As part of overall proposed methodology, this research uses random forest classifier to classify single feature type in order avoid complication like using multiple types of features causes lower processing time per frame comparing with previous research methods.

Rest of this paper is organized as follows. Section 2 demonstrates comprehensive and critical reviews in the

existing research, Section 3 illustrates proposed methodology for violence detection, Section 4 depicts extensive experimental validation for the proposed method and finally Section 5 presents concluding remarks.

II. Background Study

Various methods aiming to solve violence detection problem has been proposed mentioned in Fig. 1.

Research in [20] applied two deep neural networks (DNNs), i.e. 3D-based convolutional neural network (C3D) and CNN with long short-term memory (CNN-LSTM) for learning high level spatial-temporal information from raw image data. They combined features map achieved from C3D and CNN-LSTM through designing a shallow neural network. However, combination of features map from C3D and CNN-LSTM is suited for objet detection from still images causes high complexity in their overall research. In this context, estimation of computation time was not considered in their research during validation. Research in [21] illustrated an end to end deep neural network for violence detection using surveillance cameras. They extracted set of selectively distributed frames from video in lieu with passing spatio-temporal features to a fully connected neural network in order to classify violence or non violence action. Although, they created spatio-temporal features by performing features extraction using both space and time dimensions through a custom build convolutional neural network and long short term memory LSTM recurrent neural network, validation against computation time or processing time per frame was ignored in their research. Research in [22] selected some prior features to distinguish physical violence from daily-life activities. They designed DT–SVM (Decision Tree–SVM) two-layer classifier, i.e. first layer was acted as decision tree for using benefits of previously selected features and second layer was SVM classifier which used features for classification. However, for nearby and distant objects under complex scenarios, their research could not provide satisfactory results. In addition, their proposed approach provided significant misclassification for the frames with significant changes in light and shadow. Research in [23] proposed real time descriptor to model crowd dynamics by encoding variations in crowd texture by implicating temporal summaries of grey level co-occurrence matrix (GLCM) features. They measured inter-frame uniformity and illustrated that violent behavior varies in a less uniform manner. In addition, they performed discrimination between abnormal and normal scenes by generating scene description. However, adaptive selection of optimal parameters based on given data requires further improvement in their research. Research in [24] extracted key features, i.e. speed, direction, centroid and dimensions where they used Linear SVM to classify input video as violent or non-violent. Their proposed method considered two feature vectors, i.e. Local Binary Pattern (LBP) and Violent Flows (ViF). As Local Binary Pattern (LBP) or Violent Flows (ViF) takes less time for calculation separately than applying these feature vectors together, for this reason combination of Local Binary Pattern (LBP) and Violent Flows (ViF) in their research did not provide significant direction for future improvement. Research in [25] tested Bag-of-Words framework for detecting fight or violence by constructing a versatile and accurate fight detector using local descriptors. Although, they achieved encouraging accuracy rate, computation cost for extracting local descriptor is prohibitive for practical applications, particularly in surveillance and media rating systems. Research in [26] proposed a method using extreme acceleration pattern estimated by Radon transform to the power spectrum of consecutive frames. Their method assumed that kinematic cues that represent violent motion and strokes can be used to detect fights. In addition, they hypothesized if motion is considered as sufficient characteristics for recognition, in that case their overall methodology requires significant additional computation in lieu with confusing the detector. However, global motion estimation in their research did not seem to improve results significantly. Their proposed method required further perfection by approximating the Radon transform, which was the most time-consuming stage. Research in [27] proposed Oriented VIolent Flows (OViF) for feature extraction to take optimum benefits of motion magnitude variations in statistical motion orientations. In addition, they implicated feature combination and multiclassifier combination strategies. However, their proposed method could detect violence only in crowded scenes. Research in [28] proposed a method where corner joints of pictures are detected using Shi-Tomasi corner detection algorithm. They used optical flow parameter which was calculated using Lucas-Kanade pyramid optical flow algorithm for violence detection. However, for discontinuous and fast motion their proposed method did not provide robust performance.

Proposed method by this research calculated distance and angles as the basis of optical flow. In addition, proposed method used Kalman filter to handle illumination variation in case of any frame loss due to noise. Besides, motion pattern based on moment features extraction is used to reduce background complexity in case of similar colored objects by implicating weighted average of pictorial intensities.

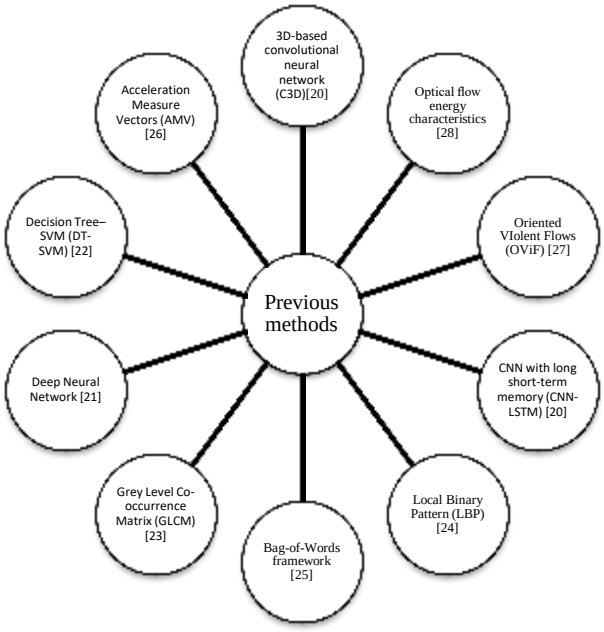


Fig. 1. Existing Methods in Previous Research for Violence Detection.

III. Proposed Methodology

This research follows nontracking based action recognition due to achieve computationally sound and effective violence action. Proposed research uses object motion to identify the event whether violent or non violent. This research uses moment features in lieu with optical flow to calculate linear distances and angles which works in pixel intensities of objects. Overall proposed method is depicted in Fig. 2.

A. Input Image and Preprocessing

Preprocessing is essential part to extract features efficiently to classify violence action. This step is the crucial as any inconsistency arising from it may lead to a misclassification of violence action. Proposed research uses monocular video camera using 35 fps frame rate for input video frames collection. Proposed research uses median filter [29, 30] to remove noise from the collected video frames. In this context, morphological processing such as resizing of the frame into 300x250 dimension, erosion and dilation are applied to ensure noise free frames to the next subsequent frames. In addition, this research also uses two frame differential approaches to find the difference between frames for finding the initial change between consecutive frames.

B. Moment Features Extraction

Proposed method extracted moment features from median filtered image. If m and n are the co-ordinates of median filtered frame $Z_f\ (m,n)$, raw moments of $Z_f\ (m,n)$ for order $(i+j)$ is defined as in (1).

$$R_{ij} = \sum\sum m^i y^j Z_f(m,n) \quad (1)$$

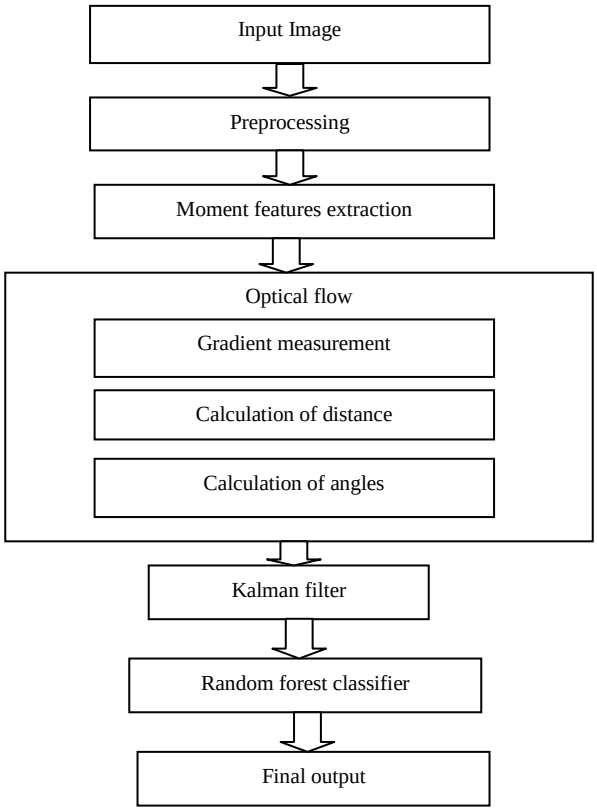


Fig. 2. Proposed Method for Violence Detection.

When considering $I_f\ (x,y)$ as 2D continuous function, (1) can be expressed as (2).

$$R_{ij} = \int\int m^i n^j Z_f(m,n) \quad (2)$$

Here, R_{ij} is denoted as raw moments to be used for calculating distances and angles as the basis of optical flow for further processing.

C. Calculation of Distance and Angle

Proposed method determines motion pattern among consecutive frames from the extracted moment features using two frame differential approach where angles and linear distances are calculated for each frame of input video. Parameters like pyramid scale, levels, window size, iteration are used to calculate motion. Pyramid scale of 0.5 is used as classical pyramid where each next layer is twice smaller than previous one. Value of levels 5 is used as mean number of pyramid layers including the initial frame. Proposed method uses larger window size of 20 to improve efficiency of the classification in terms with detecting motion. This research uses determined window around foreground object which limits the scope of motion segmentation to a smaller area. This means that the probability of the extraction of motion is getting lost because a similar colored object in the background is reduced in lieu with improving processing speed.

Proposed method calculates combination of distance and angles as the basis of optical flow for each frame of input video to classify whether action is violent or not. Distances are categorized for every pixel in the consecutive frame by 60° interval followed by summation of all distances. Proposed method calculates image gradients in horizontal and vertical direction in lieu with gradient along time. Proposed method uses optical flow to calculate angles and linear distance for each frame of input video. From the calculated angles, proposed method categorizes linear distances which generate oriented histogram of 6 bins and equally divided by 60 degree. Distance and angles are calculated as in (3) and (4).

$$K=\sqrt{dm^2 + dn^2} \quad (3)$$

$$L=\tan^{-1}\left(\frac{dn}{dm}\right) \quad (4)$$

Here, K is denoted as distance, angle is denoted by L, dm denotes distance change in horizontal direction and dn denotes distance change in vertical direction. Distance and angle is achieved for certain pixel position denoted as (m_1, n_1) of (i-1)th frame and same pixel position denoted as (m_2, ny_2) of ith frame. Pythagoras theorem [31] is used to calculate angles and linear distances by the proposed method where by angle L changes of degree is defined from (i-1)th frame to ith frame in every pixel. Linear distances of pixels between (i-1)th and frame nth frame are defined by K. Finally, proposed method uses Random forest classifier to classify whether the video scene contains violence or not.

D. Kalman Filter

Kalman filter is used to provide the best estimate of states in the presence of noise. Proposed method uses Kalman filter to optimally estimate distances and angles for higher accuracy rate. During distances and angles calculation, if there is any

frame loss due to noise or illumination variation then Kalman filter is used by the proposed method to process that frame by illuminating noise. Although, median filter was applied during preprocessing step, some frames can be often still noise may cause deviation in performance.

E. Classification

Finally, decision for violence is determined using random forest classifier from single feature vector by generating a set of probabilities for each class. In this context, probabilities are estimated using mean predicted class probabilities of the trees in the forest where class probability of a single tree is the fraction of samples of the same class in the tree. Class with highest probability is the one that is assigned to the frame as the "decision". In this context, ratio of the highest probability to the second highest probability is referred to as "confidence" of the decision. In this regard, proposed method by this research used adaptive threshold using (5) [32, 33]. Any decision with confidence more than threshold is considered as "violence" and others are "non violence".

$$T = V - \frac{V \times (log_2(I)+1)}{100} \quad (5)$$

Here, total number of frames is denoted as I, mean value of pictorial intensities is denoted as V in a video frame. Threshold value is denoted as T.

IV. Experimental Results and Discussion

A. Hardware and Software Set Up

C# programming language [34] is used to validate the proposed method by this research. Core i7 processor with 8 Gigabyte RAM is used for experimentation. Various performance metrics are used to validate the proposed methodology, i.e. accuracy rate [22, 35, 36], error [27, 37, 38], computation time [28, 39, 40] and frame rate [20, 41, 42, 43, 44].

B. Datasets

Proposed method is validated using Hockey Fight and Movies datasets [24, 25, 38, 26, 27]. Total 1000 videos are taken from Hockey fight dataset where 500 videos contain violent sequence and rest 500 videos contain non-violent video sequence. Besides, total 200 videos are taken from Movies datasets where 100 videos contain violent sequence and rest 100 videos contain non-violent video sequence. Whole datasets are divided into five sets for five cross validation. For Hockey Fight dataset and Movies datasets, resolution of frames is fixed to 300X250.

C. Experimental Results

Proposed method received accuracy rate of 99.12% for Movies dataset and 94.82% for Hockey Fight dataset shown in Table I. Accuracy rate achieved by the proposed method for Hockey Fight dataset is lower comparing with Movies dataset as there are some ambiguities in Hockey Fight dataset to distinguish that whether two players are playing or fighting with each other. Proposed method received error rate of 0.88% for Movies dataset and 5.18% for Hockey Fight dataset indicating that Hockey Fight dataset causes more error than Movies dataset. Proposed method required computation time of 0.0010 second for Movies dataset and 0.0020 sec second for Hockey Fight datasets indicating that for Hockey Fight dataset proposed method required more computation time due to bigger size of Hockey Fight dataset in lieu with containing more complex scenarios than Movies dataset.

D. Comparison with Previous Research Results

Research in [20] received accuracy rate of 63% using 3D-based convolutional neural network (C3D) and 61% using CNN with long short-term memory (CNN-LSTM) by using frame rate of 25 fps. Due to the additional step such as design of shallow neural network by combining features map obtained from C3D and CNN-LSTM can lead to increase computation complexity for the overall methodology. Research in [21] received accuracy rate of 94.5% using spatio-temporal features and passing them to a fully connected neural framework to classify the video to violence or non-violence action. Although, they performed feature extraction using both space and time dimensions to create spatio-temporal features through a custom build convolutional neural network, estimation of processing time per frame was ignored in their research. Research in [22] received accuracy rate of 97.6% by designing DT-SVM (Decision Tree SVM) using prior determined features to distinguish physical violence from daily-life activities. However, in case of complex scenes with both nearby and distant objects, their research could not provide satisfactory validation results. Research in [23], received accuracy rate of 90.5% using frame rate of 30 fps by implicating crowd dynamics with the use of encoding changes in crowd texture and temporal summaries of grey level co-occurrence matrix (GLCM) features. However, their research requires adaptive method for choosing optimal parameters based on given data. Research in [24] tested their proposed method in non-crowded and crowded scenarios to verify the effectiveness of Local Binary Pattern (LBP) features. They received accuracy rate of 89.1% with error rate of 10.9%. However, they did not validate their method based on computation time and frame rate. Research in [25] received approximate accuracy rate of 90% by using popular bag-of-words approach which can accurately recognize fight sequences. However, computational cost of extracting features in their research is not encouraging for practical applications. Research in [27] received accuracy of 88% using statistical motion orientation information. However, they received error rate of 12% as their proposed Oriented VIolent Flows (OViF) could detect violence in crowded scenarios only which demands more robust validation. Research in [28] received accuracy rate of 72% using histogram of the computed optical flow energy values. However, they received high error rate of 28% due to inability of their method to perform under discontinuous and fast motion. Research in [26] received accuracy of 98.9% using extreme acceleration patterns as the main feature where their required processing time was 0.0419 second. However, further perfection was required in their method by approximating Radon transform, which is the most time-consuming stage. Among all these previous methods stated above, proposed method by this research received higher accuracy rate of 99.12% with lower required computation time per frame of 0.0010 second and low error rate of 0.88% using 35 fps frame rate mentioned in Table II, Table III, Table IV and Table V.

TABLE I. ACCURACY, ERROR RATE AND COMPUTATION TIME FOR THE PROPOSED METHOD

Datasets	Accuracy	Error Rate	Computation Time	Frame Rate
Movies Dataset [25, 38, 26]	99.12%	0.88%	~ 0.0010 sec	35fps
Hockey Fight Dataset [24,25, 38, 26, 27]	94.82%	5.18%	~ 0.0020 sec	

TABLE II. COMPARISON WITH PREVIOUS METHODS BASED ON ACCURACY

Previous Methods	Accuracy
Proposed method	99.12%
C3D [20]	63%
CNN-LSTM [20]	61%
DNN [21]	94.5%
DT-SVM [22]	97.6%
GLCM [23]	90.5%
Local Binary Pattern (LBP) [24]	89.1%
Bag-of-Words framework [25]	90%
Oriented VIolent Flows (OViF) [27]	88%
Optical flow energy characteristics [28]	72%
Acceleration Measure Vectors (AMV) [26]	98.9%

TABLE III. COMPARISON WITH PREVIOUS METHODS BASED ON ERROR RATE

Methods	Error Rate
Proposed method	0.88%
Local Binary Pattern (LBP) [24]	10.9%
Oriented VIolent Flows (OViF) [27]	12%
Optical flow energy characteristics [28]	28%
Acceleration Measure Vectors (AMV) [26]	1.1%

TABLE IV. COMPARISON WITH PREVIOUS METHODS BASED ON COMPUTATION TIME

Previous Methods	Computation time
Proposed method	~ 0.0010 sec
Acceleration Measure Vectors (AMV) [26]	0.0419sec
Optical flow energy characteristics [28]	0.025 sec

TABLE V. COMPARISON WITH PREVIOUS METHODS BASED ON FRAME RATE

Previous Methods	Frame rate
Proposed method	35 fps
C3D [20]	25 fps
CNN-LSTM [20]	25 fps
GLCM [23]	30 fps

E. Analysis and Discussion

Research in [20] utilized 3D-based convolutional neural network (C3D) and CNN with long short-term memory (CNN-LSTM). For C3D and CNN-LSTM they achieved accuracy rate of 63% and 61% respectively using 25 fps frame rate. Although, their accuracy was not promising, usages of two deep neural networks (DNNs) were robust on learning high level spatial-temporal information from raw image data. They combined features maps obtained from C3D and CNN-LSTM networks by designing shallow neural network which acted as third scenario in their research demands for further validation to establish their overall research in terms with computational time. Proposed method by this research estimates approximate computation time of 0.0010 sec per frame in lieu with accuracy rate of 99.12% using moment features instead of combing other feature measurement indicates better validation performance than research in [20]. Research in [21] received accuracy rate of 94.5% by extracting set of selectively distributed frames of the video clip and passing spatio-temporal features to a fully connected neural architecture in order to classify the video as violence or non-violence action. However, due to usage of a custom build convolutional neural network and long short term memory LSTM recurrent neural network to process spatio-temporal features based on space and time dimensions for feature extractions, validation against computation time or processing time was totally ignored in their research. In this context, proposed method by this research uses moments feature to extract motion characteristics for classifying violent characteristics and estimates processing time per frame to validate the overall proposed methodology. Research in [22] received accuracy rate of 97.6% where they designed DT–SVM (Decision Tree–SVM) two-layer classifier, i.e. first layer was a decision tree to take benefits of prior determined features, second layer was SVM classifier to use features for classification. However, frames with significant variations in light and shadow caused misclassification in their research. In addition, their research requires further investigation in case of nearby and distant objects for complex scenarios. Proposed method by this research is validated using sufficient datasets comparing with research in [22] in lieu with that proposed method uses random forest classifier to classify single feature type in order to avoid complication like using multiple types of features causes lower processing time per frame comparing with previous research methods. Research in [23] received accuracy rate of 90.5% using 30 fps frame rate by introducing measurement of inter-frame uniformity in lieu with demonstrating violent behavior changes in a less uniform manner. Although, their proposed method performed discrimination between abnormal and normal scenes, in case of choosing optimal parameters based on given data initiates the need of adaptive method to be integrated with their overall proposed methodology which will surely demand for further validation. Proposed method by this research uses adaptive threshold during classification in lieu with extracting single feature type initially remedies the need of choosing additional parameters to achieve optimal performance like in research [23] causes gaining better performance in terms with accuracy and frame rate. Research in [24] achieved accuracy rate of 89.1% using Linear SVM to classify video as violent or non-

violent. They received error rate of 10.9% which indicates lower performance comparing with the proposed method by this research. In their research, Local Binary Pattern (LBP) or Violent Flows (ViF) takes less time for calculation than applying Local Binary Pattern (LBP) and Violent Flows (ViF) together. However, need of applying Local Binary Pattern (LBP) and Violent Flows (ViF) together instead of applying them separately did not provide any future direction for improvement in terms with performance by their research. In addition, their proposed method was not validated based computation time to indicate efficiency in terms with processing duration per frame. In this regard, proposed method by this research received higher accuracy rate of 99.12% and error rate of 0.88% using random forest classifier from single feature vector by generating a set of probabilities for each class indicates higher efficiency that research in [24]. Research in [25] achieved accuracy rate of 90% by constructing versatile and accurate fight detector using local features descriptors. Although, their accuracy rate was impressive, computational cost to construct local features vectors was impractical particularly in surveillance and media rating system. Proposed method by this research received accuracy rate of 99.12% by using moment features as means of the pixel intensity distribution to limit segmentation tasks in lieu with reducing computational complexity. Oriented VIolent Flows (OViF) by research in [27] received accuracy rate of 88% by taking full advantage of motion magnitude change information in statistical motion orientations. They received accuracy rate of 12% due to adaptation of features combination and multiclassifier strategies. In addition, their method was applicable only for crowded scenarios. In this context, proposed method by this research achieved accuracy rate of 99.12% with low error rate of 0.88% by measuring distance and angles in horizontal and vertical direction as the basis of optical flow strategy indicates better performance than research in [27]. Research in [26] received accuracy rate of 98.9% due to efficient estimation of extreme acceleration patterns as the main features by implicating Radon transform to the power spectrum of consecutive frames causes low error rate of 1.1% with required computation time of 0.0419second. Although, they hypothesized that motion was sufficient for recognition, global motion estimation experiments did not seem to improve results significantly. In addition, their proposed method needs further investigation to estimate relative importance of motion and appearance in formation for the recognition of violence or nonviolence actions. Proposed method by this research considers motion estimation using two frame differential approach and moments feature extraction where accuracy rate of 99.12% was achieved using 35 fps and error rate of 0.88% were received using 0.0010 second computation time per frame indicates better performance than in research [26] shown in Fig. 3, Fig. 4, Fig. 5 and Fig. 6.

Research in [28] received accuracy rate of 72% using Shi-Tomasi corner detection and histogram of the computed optical flow energy values. They received error rate of 28% using 0.025 second due to the usage of corner. Although, usage of corner features increases computation time in most of the research in computer vision domain, one of their significant achievement was that their method reduced time cost.

However, in case of discontinuous and fast motion, their method was not robust. Proposed method by this research showed better performance than in research [28] using moment features as the basis of uncertainty measurement of pictorial intensities to use motion pattern and later random forest classifier was used to classy violence and nonviolence behavior.

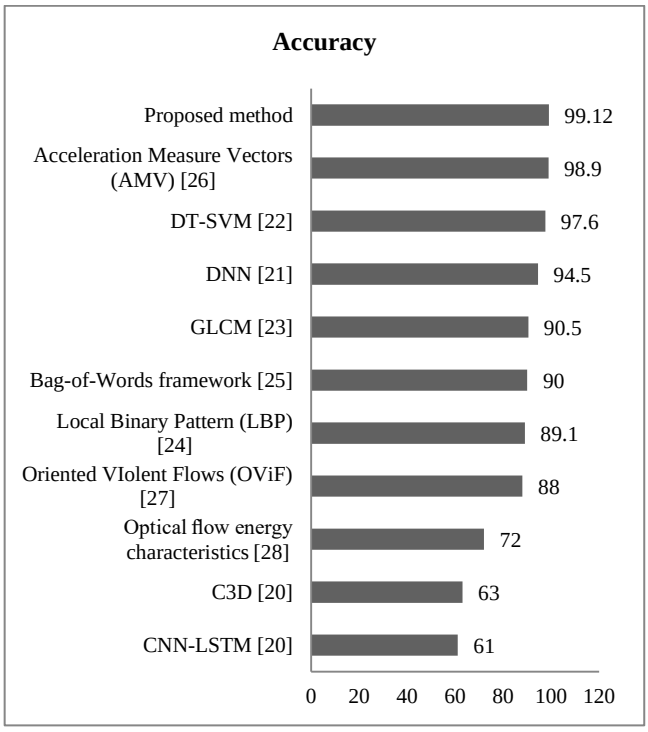


Fig. 3. Comparison among Proposed Method and Previous Research based on Accuracy.

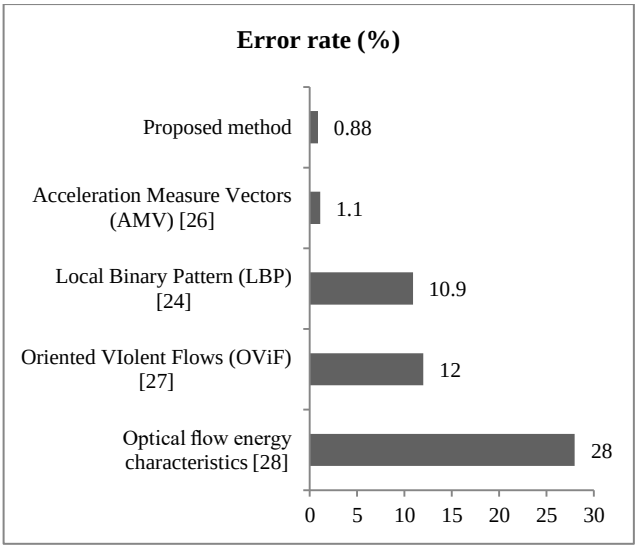


Fig. 4. Comparison among Proposed Method and Previous Research based on Error Rate.

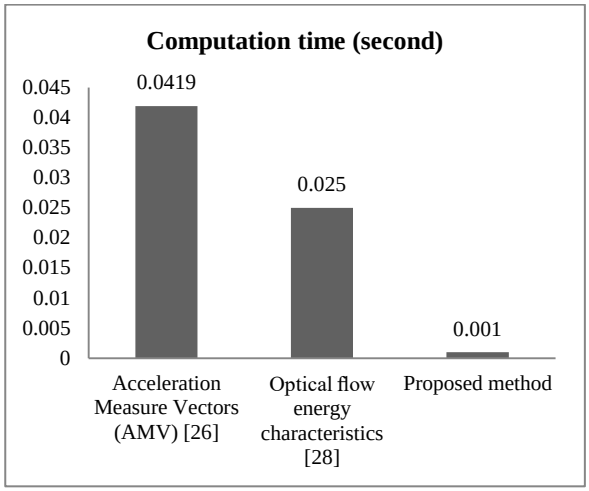


Fig. 5. Comparison among Proposed Method and Previous Research based on Computation Time Per Frame.

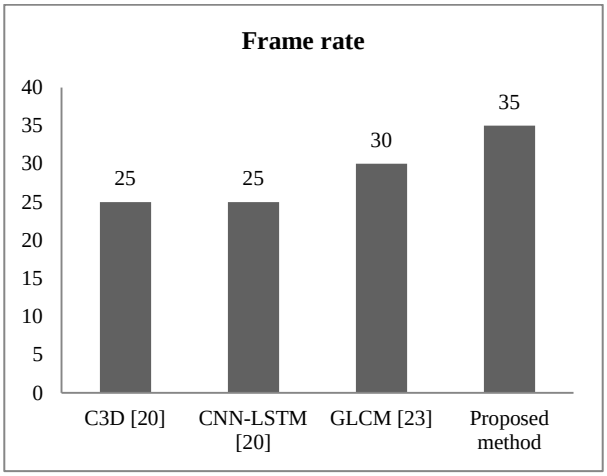


Fig. 6. Comparison among Proposed Method and Previous Research based on Frame Rate.

V. Conclusion

Proposed method used motion patterns by extracting moment features for uncertainty measurement of pictorial intensity distribution based on efficient scene interpretation. In case of similar colored object in the background, moment properties provide certain particular weighted average of pictorial intensities causes attractive interpretation which played vital role to minimize background complexity. After that, linear distances and angles are calculated for each frame as the basis of optical flow followed by Kalman filter to rectify frame loss due to noise or illumination variation which plays significant role to optimally estimate distance and angles for higher accuracy rate. Finally, proposed method used random forest classifier to classify single feature type in order to avoid complication like using multiple types of features causes lower processing time comparing with previous research methods. Experimental results for the proposed method reveal higher efficiency comparing with previous research results in terms with accuracy rate, computation time and frame rate. Proposed method achieved maximum accuracy of 99.12% using frame rate of 35 fps where required computation time per frame was 0.0010 sec. Performance of the proposed method reveals the potentiality to provide significant capability to surveillance applications for monitoring violence efficiently and reduce the impact of violence related injuries. In future, this research intends to be involved more complex activities, i.e. recognition of violence for distant objects and improvement of recognition performance for the misclassified samples.

Acknowledgment

The authors would like to thank Universiti Kebangsaan Malaysia for providing financial support under the Geran Galakan Penyelidikan research grant, GGP-2017-030.

References

[1] Jain and D. K. Vishwakarma, "State-of-the-arts Violence Detection using ConvNets," in 2020 International Conference on Communication and Signal Processing (ICCSP), 2020, pp. 0813-0817.

[2] K. Gkountakos, K. Ioannidis, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, "A Crowd Analysis Framework for Detecting Violence Scenes," in Proceedings of the 2020 International Conference on Multimedia Retrieval, 2020, pp. 276-280.

[3] R. Halder, "Discrete Wavelet Transform for CNN-BiLSTM-based Violence Detection."

[4] E. Fenil, G. Manogaran, G. Vivekananda, T. Thanjaivadivel, S. Jeeva, and A. Ahilan, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM," Computer Networks, vol. 151, pp. 191-200, 2019.

[5] J. Li, X. Jiang, T. Sun, and K. Xu, "Efficient violence detection using 3d convolutional neural networks," in 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2019, pp. 1-8.

[6] M. Ramzan, A. Abid, H. U. Khan, S. M. Awan, A. Ismail, M. Ahmed, M. Ilyas, and A. Mahmood, "A review on state-of-the-art violence detection techniques," IEEE Access, vol. 7, pp. 107560-107575, 2019.

[7] K. Singh, K. Y. Preethi, K. V. Sai, and C. N. Modi, "Designing an Efficient Framework for Violence Detection in Sensitive Areas using Computer Vision and Machine Learning Techniques," in 2018 Tenth International Conference on Advanced Computing (ICoAC), 2018, pp. 74-79.

[8] M. Sharma and R. Baghel, "Video Surveillance for Violence Detection Using Deep Learning," in Advances in Data Science and Management, ed: Springer, 2020, pp. 411-420.

[9] A. S. Saif, A. S. Prabuwono, and Z. R. Mahayuddin, "Moment feature based fast feature extraction algorithm for moving object detection using aerial images," PloS one, vol. 10, p. e0126212, 2015.

[10] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violence detection in surveillance video using low-level features," PloS one, vol. 13, p. e0203668, 2018.

[11] J. Ha, J. Park, H. Kim, H. Park, and J. Paik, "Violence detection for video surveillance system using irregular motion information," in 2018 International Conference on Electronics, Information, and Communication (ICEIC), 2018, pp. 1-3.

[12] J. Huang, G. Li, N. Li, R. Wang, and W. Wang, "A violence detection approach based on spatio-temporal hypergraph transition," in International Conference on Computer Analysis of Images and Patterns, 2017, pp. 218-229.

[13] A. Saif, A. S. Prabuwono, and Z. R. Mahayuddin, "Moving object detection using dynamic motion modelling from UAV aerial images," The Scientific World Journal, vol. 2014, 2014.

[14] S. Roshan, G. Srivathsan, K. Deepak, and S. Chandrakala, "Violence Detection in Automated Video Surveillance: Recent Trends and Comparative Studies," in The Cognitive Approach in Cloud Computing and Internet of Things Technologies for Surveillance Tracking Systems, ed: Elsevier, 2020, pp. 157-171.

[15] A. Traoré and M. A. Akhloufi, "2D Bidirectional Gated Recurrent Unit Convolutional Neural Networks for End-to-End Violence Detection in

Videos," in International Conference on Image Analysis and Recognition, 2020, pp. 152-160.

[16] T. Zhang, W. Jia, X. He, and J. Yang, "Discriminative dictionary learning with motion weber local descriptor for violence detection," IEEE transactions on circuits and systems for video technology, vol. 27, pp. 696-709, 2016.

[17] Q. Zhou, C. Wu, J. Xing, J. Li, Z. Yang, and Q. Yang, "Wi-Dog: monitoring school violence with commodity WiFi devices," in International Conference on Wireless Algorithms, Systems, and Applications, 2017, pp. 47-59.

[18] K. Deepak, L. Vignesh, and S. Chandrakala, "Autocorrelation of gradients based violence detection in surveillance videos," ICT Express, vol. 6, pp. 155-159, 2020.

[19] T. Zhang, W. Jia, B. Yang, J. Yang, X. He, and Z. Zheng, "MoWLD: a robust motion image descriptor for violence detection," Multimedia Tools and Applications, vol. 76, pp. 1419-1438, 2017.

[20] B. Peixoto, B. Lavi, J. P. P. Martin, S. Avila, Z. Dias, and A. Rocha, "Toward subjective violence detection in videos," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 8276-8280.

[21] M. M. Moaaz and E. H. Mohamed, "Violence Detection In Surveillance Videos Using Deep Learning," الحاسبات فى المعلوماتية الذشرة والمعلومات, vol. 2, pp. 1-6, 2020.

[22] L. Ye, L. Wang, H. Ferdinando, T. Seppänen, and E. Alasaarela, "A Video-Based DT–SVM School Violence Detecting Algorithm," Sensors, vol. 20, p. 2018, 2020.

[23] K. Lloyd, P. L. Rosin, D. Marshall, and S. C. Moore, "Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM)-based texture measures," Machine Vision and Applications, vol. 28, pp. 361-371, 2017.

[24] P. Vashistha, C. Bhatnagar, and M. A. Khan, "An architecture to identify violence in video surveillance system using ViF and LBP," in 2018 4th International Conference on Recent Advances in Information Technology (RAIT), 2018, pp. 1-6.

[25] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in International conference on Computer analysis of images and patterns, 2011, pp. 332-339.

[26] O. Deniz, I. Serrano, G. Bueno, and T.-K. Kim, "Fast violence detection in video," in 2014 international conference on computer vision theory and applications (VISAPP), 2014, pp. 478-485.

[27] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows," Image and vision computing, vol. 48, pp. 37-41, 2016.

[28] Z. Guo, F. Wu, H. Chen, J. Yuan, and C. Cai, "Pedestrian violence detection based on optical flow energy characteristics," in 2017 4th International Conference on Systems and Informatics (ICSAI), 2017, pp. 1261-1265.

[29] Z. R. Mahayuddin and A. F. M. S. Saif, "A Comprehensive Review Towards Segmentation and Detection of Cancer Cell and Tumor for Dynamic 3D Reconstruction," Asia-Pacific Journal of Information Technology and Multimedia, vol. 9, pp. 28-39, 2020.

[30] Z. R. Mahayuddin and A. S. Saif, "A Comprehensive Review Towards Appropriate Feature Selection for Moving Object Detection Using Aerial Images," in International Visual Informatics Conference, 2019, pp. 227-236.

[31] J. D. Sally, Roots to research: a vertical development of mathematical problems: American Mathematical Soc., 2007.

[32] A. S. Saif, A. S. Prabuwono, Z. R. Mahayuddin, and T. Mantoro, "Vision-based human face recognition using extended principal component analysis," International Journal of Mobile Computing and Multimedia Communications (IJMCMC), vol. 5, pp. 82-94, 2013.

[33] Z. R. Mahayuddin and A. S. Saif, "A Comparative Study Of Three Corner Feature Based Moving Object Detection Using Aerial Images," Malaysian Journal of Computer Science, pp. 25-33, 2019.

[34] A. S. Saif, A. S. Prabuwono, Z. R. Mahayuddin, and H. T. Himawan, "A review of machine vision based on moving objects: object detection from UAV aerial images," International Journal of Advancements in Computing Technology, vol. 5, p. 57, 2013.

[35] Z. R. Mahayuddin and A. F. M. S. Saif, "Efficient Hand Gesture Recognition Using Modified Extrusion Method based on Augmented Reality," TEST Engineering and Management, vol. 83, pp. 4020-4027, 2020.

[36] Z. R. Mahayuddin and A. F. M. S. Saif, "Augmented Reality Based Ar Alphabets Towards Improved Learning Process In Primary Education System," Journal Of Critical ReviewS, vol. 7, 2020.

[37] Z. R. Mahayuddin, A. S. Saif, and A. S. Prabuwono, "Efficiency measurement of various denoise techniques for moving object detection using aerial images," in 2015 International Conference on Electrical Engineering and Informatics (ICEEI), 2015, pp. 161-165.

[38] A. Saif and Z. R. Mahayuddin, "Moving Object Segmentation Using Various Features from Aerial Images: A Review," Advanced Science Letters, vol. 24, pp. 961-965, 2018.

[39] A. Saif, A. Prabuwono, and Z. Mahayuddin, "Adaptive long term motion pattern analysis for moving object detection using UAV aerial images," International Journal of Information System and Engineering, vol. 1, pp. 50-59, 2013.

[40] A. S. Saif, A. S. Prabuwono, and Z. R. Mahayuddin, "Adaptive motion pattern analysis for machine vision based moving detection from UAV aerial images," in International Visual Informatics Conference, 2013, pp. 104-114.

[41] A. S. Saif, A. S. Prabuwono, and Z. R. Mahayuddin, "Real time vision based object detection from UAV aerial images: a conceptual framework," in FIRA RoboWorld Congress, 2013, pp. 265-274.

[42] A. S. Saif, A. S. Prabuwono, and Z. R. Mahayuddin, "Motion analysis for moving object detection from UAV aerial images: A review," in 2014 International Conference on Informatics, Electronics & Vision (ICIEV), 2014, pp. 1-6.

[43] A. S. Saif and Z. R. Mahayuddin, "Vehicle Detection for Collision Avoidance Using Vision based Approach: A Constructive Review," Solid State Technology, pp. 2861-2869, 2020.

Object based Image Splicing Localization using Block Artificial Grids

P N R L Chandra Sekhar[1]
Department of Computer Science and Engineering
Gandhi Institute of Technology and Management
Visakhapatnam, AP, INDIA

T N Shankar[2]
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur, AP, INDIA

***Abstract*—People share pictures freely with their loved ones and others using smartphones or social networking sites. The news industry and the court of law use the pictures as evidence for their investigation. Simultaneously, user-friendly photo editing tools alter the content of pictures and make their validity questionable. Over two decades, research work is going on in image forensics to determine the picture's trustworthiness. This paper proposes an efficient statistical method based on Block Artificial Grids in double compressed images to identify regions attacked by image manipulation. In contrast to existing approaches, the proposed approach extracts the artefacts on individual objects instead of the entire image. A localized algorithm is proposed based on the cosine dissimilarity between objects and exploit the tampered object with maximum dissimilarity among objects. The experimental results reveals that the proposed method is superior over other current methods.**

***Keywords*—Image forensics; splicing localization; block artificial grids; object segmentation; double compression**

I. Introduction

Now-a-days, people freely share their ideas, pictures, and comments on social networking sites. The usage of images grows enormously in different ways, such as the Government initiative towards digitizing all areas, evidence in the court of law, journalism, science, and forensics discovery [1]. Simultaneously, the widely available image editing tools induced interest in making the images or videos manipulate with ease that cannot trace out to human vision. Copy-move, splicing, resampling, cloning are few manipulation attacks to tamper images. A manipulated image significantly impacts the trustworthiness when used for evidence [2] [3]. It brings a significant challenge in image forensics to discover the original one from manipulated at the same time establish its authenticity and locate the tampered region [4].

Digital Image Forensics from Multimedia security aims at designing powerful techniques to detect manipulation attacks on images [5]. Active methods like watermarking, authentic code embedded in the original image, and verifying its authenticity. In contrast, passive methods like tampering detection do not require any external clue to assess the image's authenticity. Different tampering techniques in the literature assume that images taken from different camera models or different processing operations introduce inherent patterns into tampered image [6] [7][8] [9]. Furthermore, it assumes that these underlying patterns consistent throughout the original image, and when any manipulation attacks it, there will be inconsistency in those patterns. These inconsistency statistics can thus be used as forensic features to identify image tampering [10] [11].

In the image splicing tampering, a part of the source image is copied and pasted into the donor image. Some post-processing techniques will apply to the tampered region to make the attack invisible and difficult to trace to the human eye [12]. This challenge attracted many researchers to find various techniques for detecting image splicing. Many of these techniques extract image features and use classification to reveal for forgery, and they achieve even high success rates [13][14]. However, it is worth locating the tampered region for many real-time purposes to gain confidence. However, image splicing localization brings many more challenges as it requires pixel-level analysis rather than image-level analysis [15] [16].

The images captured by digital cameras store in the Joint Photographic Experts Group (JPEG) format. Lossy compression is used in the JPEG format and is responsible for the proliferation of images on websites and social networking sites. The image divides into 8 x 8 non-overlapping blocks in JPEG compression, and the discrete cosine transform (DCT) is evaluated for each block and then quantified using a regular quantization matrix. When any splicing attack manipulates the image, it leads to discontinuities, and these statistical traces use to exploit tampering attacks, such as JPEG quantization artefacts and JPEG grid alignment discontinuities [17] [18].

A. Related Work

The tampered blocks will undergo single compression when there is a splicing attack, while the remaining blocks will have double compression(DQ). In [19], the authors created periodic DCT patterns and evaluated each block of the image concerning its conformance of the model. Any block whose probability distribution distinguishes from the original classifies as blocks manipulated by a tampering attack. A similar approach found in [20] where the authors assume that the distribution of JPEG coefficients changes with the number of recompressions and proposes training a set of support vector machines (SVM) for the first digit artefacts and estimated the probability distribution of each block as a single or double compressed thereby exposed the splicing attack.

In [21] comparing the discontinuities using the quality factor adopted in the tampered region with the principle that a JPEG ghosts - a local spatial minimum- will correspond to the tampering attack. The limitation of the method is; it works only if the tampered region has a lower quality factor than the

rest of the image. An alternative to the DQ discontinuities, in [22], the authors created a model on the entire image DCT coefficient distributions using the degree of quantization. The inconsistencies became indicative of the tampering attack. The difference between this method and the DCT-based is that the output is not probabilistic, making the technique relatively difficult to interpret although efficient.

In [17], tampering detection and localization uses the probability distribution of its DCT coefficients. Three features that can truly distinguish tampered regions from original ones are used and obtain accurate localization results. But, the refining of the probability map in post-processing influences localization results. To overcome it, [23] used a mixture model based on normalized grey level co-occurrence matrix (NGLCM) and obtained more accurate localization with the prior knowledge of both tampered and original regions. To get this, they used conditional probabilities of tampered regions and original regions of DCT blocks in first, second, and third-order statistics.

In recent works, deep-learning techniques applied for tampering detection and localized region. These methods learn the relevant features automatically from the network [24]. In [25] extracted the histograms of DCT coefficients from the input image and designed a one-dimensional convolutional neural network (CNN) with DCT coefficients as input to identify tampered regions by distinguishing single and double-compressed areas. In [26], proposed a two-layer CNN, in which the stacked auto-encoder model learns the elaborate features for the individual patch of the spliced image and uses contextual information to make the localization accurately. These methods provide block-based accuracy.

For obtaining pixel-level accuracy, [27] proposed a fully convolutional network (FCN) to locate spliced regions. FCN is a particular type of CNN, which replaces the fully connected layers with the convolutional layers having a 1x1 kernel. It distinguishes each pixel as spliced or original. The authors used three FCNs to deal with different scales of image contents, but these methods have drawbacks that they lose or smooths detailed structures and ignore small objects. To improve this effect, in [24] used a region proposal network (RPN), which is a kind of FCN and can be trained end-to-end specifically for detection. Using FCN and RPN, the authors achieved better results than FCN methods as well as other conventional methods. The computational complexity of deep-learning techniques is high.

In [28] proposes localization architecture that uses resampling features to capture artefacts. The Long short-term memory (LSTM), followed by an encoder network, is designed to differentiate tampered regions from the original. The decoder network learns features to localize the tampered region. The final soft-max layer learns the network parameters through the back-propagation algorithm from ground truth masks. The model is capable of localizing at the pixel level with high precision.

Although the deep learning-based techniques improve accuracy, they require training on large labelled databases, and the computational complexity is very high. The networks extract high-level visual features and neglect low-level features, which can be sources for forensic cues. In this paper, we move towards proposing a statistical-based forensic technique that can localize the tampered region from a single image in the presence of double compression. Unlike other techniques that produce probability maps from 8x8 DCT coefficients, we proposed an adequate statistical model that characterizes the fingerprints of block artificial grids (BAG) and works for any compression with any quality factor in the spatial domain.

B. Our Contribution

Over the years, various splicing localization techniques proposed in the literature. Still, there is scope for robustness and effectiveness to improve as splicing is complex. In this regard, we are offering the following contributions to our proposed work.

i) We propose object-based segmentation, and the features extracted from the individual objects and for each object, we estimate the variance of the BAG noise

ii) Instead of probability maps, we proposed a statistical-based localization algorithm based on pair-wise dissimilarity among objects to classify the suspicious object from the original ones.

The rest of the paper organizes as follows: Section II described JPEG fingerprints from block artificial grids to speed up computation time. Section III outlines the proposed statistical method to expose and localize the splicing attack. The experimental and evaluation results present in Section IV, and finally, the paper concluded in Section V.

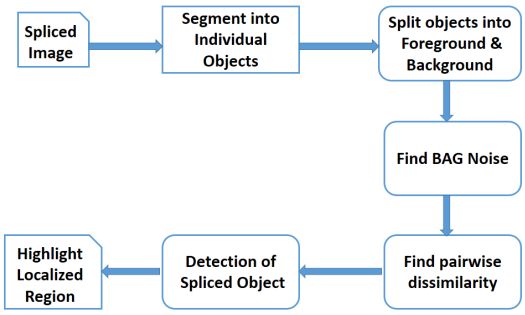


Fig. 1. The Proposed Frame Work

II. Proposed Method

The primary goal is to localize the tampered region in the spliced image. As shown in Fig. 1, the proposed method is in three levels: object-based image segmentation to extract individual objects from the spliced image and estimate each object's variance using block-artificial grids and the proposed localization algorithm on pair-wise dissimilarity among objects to expose tampered region.

A. Object Segmentation

Object Detection is a complicated computer vision problem to detect and classify objects from an individual image or videos. In many existing popular object detection frameworks, Mask R-CNN [29] is a frequently used one developed by Facebook research. It is an extension of Faster R-CNN that

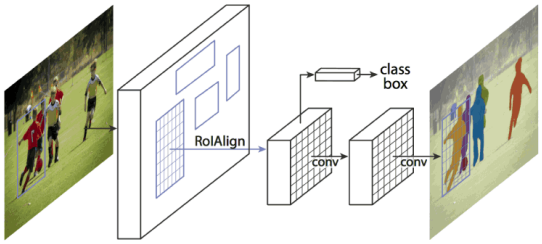


Fig. 2. Mask R-CNN Frame Work adopted from [29]

estimates the object's mask and human pose. It overcomes the COCO suite challenge by segmentation of instances, detecting bounding-box objects, and individual key points.

Using the Mask R-CNN framework, as shown in Fig. 2, performed object detection and segmentation [30] for the given spliced image extracted individual masks of all objects. Then for each mask, find its object from the input image along with the bounding box area. The object corresponds to the mask considered a foreground object, and the remaining part in the bounding box region is the background object.

B. Block Artificial Grids

The lossy JPEG compressed image leaves horizontal and vertical breaks in the image and is commonly refers to as Block Artificial Grids (BAG). The image's BAGs are roughly at the border of a 8 x 8 block with a periodicity of 8 at both horizontal and vertical edges. When any manipulation attack alters the image, the BAGs appear within the block instead of at borders. Thus this JPEG fingerprint is used in image forensics [31].

While compress the image using a digital camera, it introduces noise such as natural noise, BAG noise due to the JPEG compression factor. The artificial grid lines in a 8 X 8 block are feeble than the border edges. In [31], the authors extracted weak horizontal and vertical lines of a grayscale image with a periodicity of 8 separately to enhance these weak lines, and then combined them is referred to as BAGs.

In this paper, we focus on extracting BAGs in colour images. Since the luminance component in the JPEG standard is 8 x 8 blocks, we used only the luminance component rather than C_b and C_r of components of the YC_bC_r image.

The second-order difference of an image regards as weak horizontal edges of an image. For the given image $I(m,n)$, the absolute second-order difference $d(m,n)$ is obtained by

$$d(m,n) = |2I(m,n) - I(m+1,n) - I(m-1,n)| \quad (1)$$

A median filter is applied to enhance the weak edges and remove the interference coming from strong image edges. To further reduce the edge influence as in [31] ignored differentials greater than an experimental threshold. Then the enlarged horizontal edges are accumulated for every two subsequent blocks as:

$$e(m,n) = \Sigma_{i=n-16}^{16} d(m,i) \quad (2)$$

Then to equalize the amplitudes throughout the resultant image, a local median is reduced from each element.

$$e_r(m,n) = e(m,n) - median[\{e(i,n)|m-16 \leq i \leq m+16\}] \quad (3)$$

Thus, the weak horizontal edge image w_h obtained by applying the periodical median filter as:

$$w_h(m,n) = median[\{e_r(i,n)|i = m-16, m-8, m, m+8, m+16\}] \quad (4)$$

where $w_h(m,n)$ are elements of extracted horizontal BAG lines. The five elements in Eq. 4, with spacing eight used in the median filter, makes the strong BAGs and weak BAGs smooth, and rest are removed. As more elements used in the median filter, BAGs can extract in a better way.

The vertical BAGs w_v are also similarly extracted.

$$w_v(m,n) = median[\{e_r(m,i)|i = n-16, n-8, n, n+8, n+16\}] \quad (5)$$

The final BAG obtained by combining Eq. 4 and 5 as

$$w_b(m,n) = w_h(m,n) + w_v(m,n) \quad (6)$$

Eq. 6 gives BAGs for the original image. In the tampered image, the BAGs appear at some abnormal position, such as the block center. So, for a fixed 8 x 8 block $w_m n$, these abnormal BAGs can be obtained as [31].

$$\begin{aligned} w_m n = & \, Max\{\Sigma_{i=2}^{7} w_b(i,n)|2 \leq n \leq 7\} \\ & - Min\{\Sigma_{i=2}^{7} w_b(i,n)|n = 1,8\} \\ & + Max\{\Sigma_{i=2}^{7} w_b(m,i)|2 \leq m \leq 7\} \\ & - Min\{\Sigma_{i=2}^{7} w_b(m,i)|m = 1,8 \end{aligned} \quad (7)$$

C. Localization of Splicing Region

Mask R-CNN object detection framework [30] is used to detect individual masks from the spliced image. For each mask, first split into the foreground and background objects and extracted the BAGs, as discussed in Section II-B.

To expose discrepancies in BAGs of individual objects, we find BAG noise from Eq. 7 as:

$$\mu = \frac{1}{R}\Sigma w_m n(i,j)\sigma = \frac{1}{R}\Sigma(w_m n(i,j) - \mu)^2 \quad (8)$$

μ is mean, σ is variance, and R represents the no of BAG features in $w_m n$.

After BAG noise obtained for each object, pair-wise dissimilarity among objects evaluated as follows:

For each pair of the distinct foreground or back-ground objects, let the BAG noise be S_1 and S_2. Then the cosine dissimilarity between the objects defined as:

$$L_D = 1 - \frac{C(S_1, S_2) + 1.0}{2} \tag{9}$$

where

$$C(S_1, S_2) = \frac{S_1^T . S_2}{\|S_1\| . \|S_2\|} \tag{10}$$

$C(S_1, S_2)$ is the cosine angle between two BAG noises. The metric L_D gives values in the range [0,1]. Where the values near to 0 represent similar BAG noise levels of both objects, and near to 1 represents different levels.

Algorithm 1 Algorithm for identifying probable tampered object from Dissimilarity Matrix

Input: Estimated noise levels of N Individual objects of Spliced Image
Output: Tampered Object *Find Dissimilarity matrix*

for $i = 1$ to N **do**
  for $j = 1$ to $i - 1$ **do**
    $DM(i, j) = L_D(S_i, S_j)$
  end for
end for
Find the pair having maximum dissimilarity
for each column in DM **do**
  $[COLMAX_j, COLIDX_j] = max(DM_j)$
end for
$[cmax, cidx] = max(COLMAX)$
for each row in DM **do**
  $[ROWMAX_i, ROWIDX_i] = max(DM_i)$
end for
$[rmax, ridx] = max(ROWMAX)$
$DM(rmax, cmax)$ *has maximum dissimilarity*
Now find which object has maximum dissimilarity
$RROW = ridx, count = 0$
for each $COLIDX$ **do**
  if $(RROW = COLIDX_j)$ **then**
    $count = count + 1$
  end if
end for
retrun the tampered object
if $(count \geq 0)$ **then**
  $T_P = RROW$
end if
return T_P

The probable tampered object with maximum dissimilarity with other objects is exposed from the dissimilarity matrix using the proposed localization algorithm 1.

III. Experimental and Performance Analysis

This section evaluates the proposed method on two datasets and compares its performance with contemporary techniques.

Typically, CASIA dataset [32] is a widely used evaluation dataset for JPEG image splicing forgery detection, and it consists of 7491 authentic and 5123 spliced images with JPEG, TIFF, and BMP types of images. We randomly selected 1000 tampered images of animals, persons, birds, vehicles with the size 384 x 256 and segmented the objects using the Mask R-CNN framework. The proposed method is tested on those chosen tampered images of the CASIA dataset for localizing spliced regions.

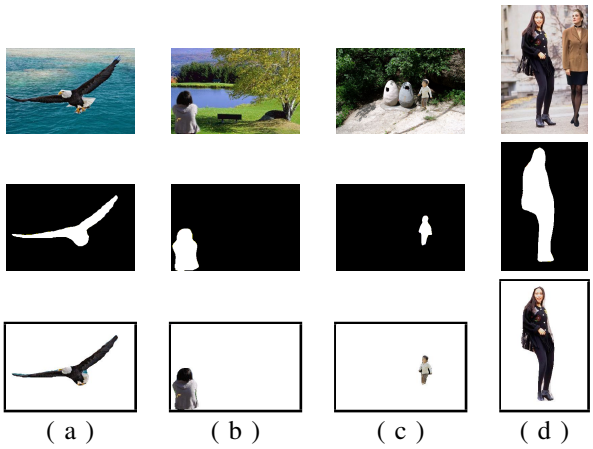


Fig. 3. Visual Evaluation of Proposed Method on CASIA Dataset

The qualitative evaluation of splicing images on the CASIA dataset shows in Fig. 3. The first row consists of randomly chosen four images, and the respective ground truth masks given in the second row. The proposed method results are in the last row, where the spliced region is highlighted, and the remaining area is marked as white. From the results, the proposed method's superiority is very clearly evident to localize the spliced region.

To increase the proposed method's robustness, we have evaluated our approach on the Image Manipulation Dataset (IMD) [33]. The dataset contains a 48 pixel high-resolution JPEG compressed images with size 3264 x 2448 with different quality factors ranging from 20% to 100%. The images were cropped to 2048 x 1536 to reduce the computational complexity and spliced each other and obtain 600 spliced images. Then the proposed method was assessed on those images.

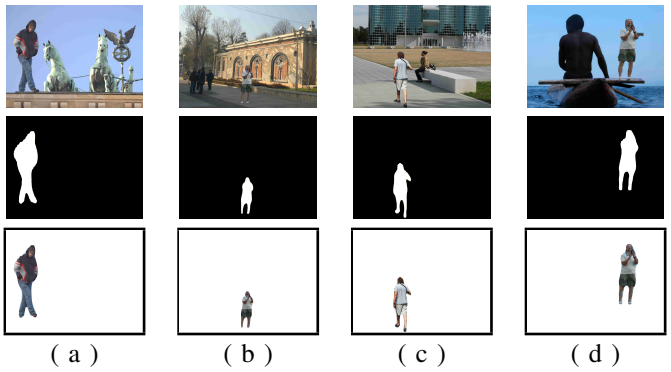


Fig. 4. Visual Evaluation of Proposed Method on the High Resolutioned Images from [33]

The evaluation results on our customized IMD spliced dataset obtained from [33], shown in Fig. 4. The first row contains randomly chosen four sample images from the dataset. The ground truth masks are in the second row, and the

proposed method results are in the third row. From the results, the proposed method works well on high-resolution images.

A. Localization Accuracy

The accuracy of splicing localization evaluates based on pixel-level F-measure. Two metrics, True Positive Rate (TPR), measure the rate of pixels that are indeed detected as spliced, and False Positive Rate (FPR), a measure of the rate of pixels that are falsely detected as spliced, are used to evaluate F-measure.

$$TPR = \frac{TP}{TP + FN} * 100 FPR = \frac{FP}{FP + TN} * 100 \quad (11)$$

Where TP is True Positive, FP is False Positive, TN is True Negative, and FN is False Negative. It expects to have high TPR and low FPR in the results. From these metrics, the F-measure defines as follows:

$$F = 2 * \frac{TPR * FPR}{TPR + FPR} \quad (12)$$

We evaluated average TPR and FPR and F-measure for all the selected images from the CASIA dataset and compared them with [23] and [24] to analyze the performance of the proposed method.

The method of [23] is based on a normalized gray level co-occurrence matrix on 8x8 DCT coefficients and, using the Bayesian posterior probability map, localized the tampering objects. Whereas, the method [24] uses a deep learning method based on Fully Convolutional Networks (FCN) with Region Proposal Network (RPN) to localize the tampered region. To evaluate the superiority of the proposed method, we compared our results with conventional and deep learning methods.

Table I contains the Comparative results of the proposed method with [23] and [24] methods on both datasets based on average F-measure. FCN methods [24] prove to have superior performance than the conventional statistical-based methods [23]. From the results, it is evident that BAG noise on individual objects in the proposed method enables us to have much superior performance than [23].

The method is robust when it has a stable performance even after applying some post-processing operations on the spliced image. To evaluate the proposed method's robustness, we applied JPEG compression with different quality factors, Gaussian blur, and added Gaussian noise to all the spliced images and tested.

For JPEG compression, eight different quality factors ranging from 20 to 90 are considered. For Gaussian blur, Gaussian smoothing kernel with standard deviation $\sigma = 1.0$ is used, and for Gaussian noise, the variance of 0.03 and 0.05 are considered.

The evaluation results on IM Dataset has been shown in Table II. As the quality factor (QF) in JPEG compression decreases and additional post-processing operations included, the FCN and NGLCM methods decrease in their average F-measure values. In contrast, the proposed method has superior as well as stable performance even in such situations.

The IM dataset images are very high-resolution, and we try to downscale the quality factor to the lowest level 20. Fig. 5 is a graph showing the proposed method's performance with other existing methods. Both FCN+RPN and NGLCM methods decreased their average F-measure as the JPEG compression quality factory is reduced towards 20. The proposed method outperforms and gives stable performance even when the quality factor reduces because the BAGs are affected only in those objects than the rest of the image.

B. Computational Complexity

The effectiveness of any method depends on its average computation time spent is minimal to get the desired result. In the proposed method, after segmenting the individual objects, we obtain BAG features from each object instead of the whole image, thereby saving a lot of computation time. For localization, also we used a simple statistical method instead of unsupervised learning techniques. Table III gives the average running time spent by each method. Among the methods, the proposed method takes less time than other methods.

IV. Conclusion

This paper is proposed an efficient method for splicing localization based on block artificial grids in a double compressed JPEG image. When a JPEG image spliced with another image's object, the block artificial grids move from 8x8 gridlines to its centre. Taking this clue, we exposed splicing forgery through object segmentation. The method is straightforward, effective than other conventional methods that use JPEG fingerprints. The proposed method also robust even when the quality factor is low in high-resolution JPEG compression. The method fails on low-resolution images, and we considered it as our future work.

References

[1] A. M. Qureshi and M. Deriche, "A review on copy-move image forgery detection techniques," in *IEEE 11th International Multi-Conference on Systems*, 2014, pp. 1–5.

[2] J. A. Redi, W. Taktak, and J. Dugelay, "Digital image forensics: a booklet for beginners," *Multimed Tools and Applications*, vol. 51, pp. 133–162, 2011.

[3] P. More, T. N. Shankar, and P. Borse, "Storage covert channel concealment in tcp field," *International Journal of Control Theory and Applications*, vol. 10(1), pp. 1–7, 2017.

[4] H. Farid, "Image forgery detection a survey," *IEEE Signal Processing Magazine*, vol. 26(2), pp. 16–25, 2009.

[5] G. K. Birajdar and V. H. Mankar, "Digital image forgery detection using passive techniques," *Digital Investigation: The International Journal of Digital Forensics & Incident Response*, vol. 10(3), pp. 226–245, 2013.

[6] P. Kakar, N. Sudha, and S. W., "Exposing digital image forgeries by detecting discrepancies in motion blur," *IEEE Transactions on Multimedia*, vol. 13, pp. 443–452, 2011.

[7] T. N. Shankar and K. Spurthy, "Intrusion detection system using frequent item set in manet," *Journal of Adv Research in Dynamical and Control Systems*, vol. 10(1), pp. 356–362, 2018.

[8] T. N. Shankar, R. K. Senapati, P. M. K. Prasad, and G. Swain, "Volumetric medical image compression using 3d listless embedded block partitioning," 2016, vol. 1(2100), pp. 1–16.

[9] N. T. Babu, "Fpga implementation of hybrid system using timing attack resistant cryptographic technique," *Advances And Applications In Mathematical Sciences*, vol. 17(1), pp. 271–292, 2017.

TABLE I. COMPARATIVE RESULTS ON CASIA AND IMD DATASETS USING AVERAGE F-MEASURE

Method	CASIA 2.0	IMD
FCN+RPN	0.7388	0.6234
NGLCM	0.6524	0.5572
Proposed	0.7852	.0692

TABLE II. COMPARATIVE RESULTS FOR ROBUSTNESS ON IM DATASET USING AVERAGE F-MEASURE

Method	(JPEG Compression)		(Gaussian Blur)	(Gaussian Noise)	
	QF=50	QF=70	$\sigma = 1.0$	variance=0.03	variance=0.05
FCN+RPN	0.4365	0.6158	0.6187	0.6132	0.6092
NGLCM	0.3934	0.4323	0.5412	0.5389	0.5395
Proposed	0.6418	0.6596	0.7520	0.7514	0.7520

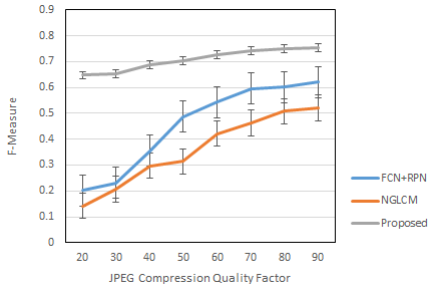


Fig. 5. Comparative Results of JPEG Quality Factory with F-Measure

TABLE III. AVERAGE RUNNING TIME

Method	FCN+RPN	NGLCM	Proposed
(Average Running Time in sec)	97.3	78.9	16.8

[10] P. N. R. L. C. Sekhar and T. N. Shankar, "Review on image splicing forgery detection," *International Journal of Computer Science and Information Security*, vol. 14(11), pp. 471–475, 2016.

[11] K. Spurthy, T. N. Shankar, and R. K. Senapati, "Improving authentication of an iris recognition system by digital signature via elliptic curve cryptosystem," 2016.

[12] K. Bahrami, A. C. Kot, and L. Li, "Blurred image splicing localization by exposing blur type inconsistency'," *IEEE Trans. Inf. Forensics Security*, vol. 10(5), pp. 999–1009, 2015.

[13] Y. Zhang, C. Zhao, Y. Pi, and L. S., "Revealing image splicing forgery using local binary patterns of dct coefficients," in *Liang Q. et al. (eds) Communications, Signal Processing, and Systems. Lecture Notes in Electrical Engineering*. Springer, 2012, vol. 202), pp. 181–189.

[14] K. Spurthy and T. N. Shankar, "An efficient cluster-based approach to thwart wormhole attack in adhoc networks," *International Journal of Advanced Computer Science and Applications*, vol. 11(9), pp. 312–316, 2020.

[15] Z. He, W. Lu, W. Sun, and J. Huang, "Digital image splicing detection based on markov features in dct and dwt domain," *IEEE Transactions on Pattern Recognition*, vol. 45(12), pp. 4292–4299, 2012.

[16] K. L. P. Rao, K. R. Rao, and K. R. R. M. Rao, "Adaptive energy efficient decentralized hierarchical dynamic cluster based routing protocol in wsn," *Advances And Applications In Mathematical Sciences*, vol. 17(1), pp. 185–191, 2017.

[17] T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of jpeg artifacts," *IEEE Transactions on Information Forensics and Security*, vol. 7(3), pp. 1003–1017, 2012.

[18] M. Zampoglou, S. Papadopoulos, and Y. Kompatsiaris, "Large-scale evaluation of splicing localization algorithms for web images," *Multimedia Tools and Applications*, vol. 76(4), pp. 4801–4834, 2017.

[19] Z. Lin, J. He, X. Tang, and T. Ck, "Fast, automatic and fine-grained tampered jpeg image detection via dct coefficient analysis," *Pattern Recognition*, vol. 42, pp. 2492–2501, 2009.

[20] I. Amerini, R. Becarelli, R. Caldelli, and A. Del Mastio, "Splicing forgeries localization through the use of first digit features," *IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 143–148, 2014.

[21] H. Farid, "Exposing digital forgeries from jpeg ghosts," *IEEE Trans. Inf. Forensics Security*, vol. 4(1), pp. 154–160, 2009.

[22] T. Bianchi, A. De Rosa, and A. Piva, "Improved dct coefficient analysis for forgery localization in jpeg images," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 2444–2447.

[23] F. Xue, W. Lu, Z. Ye, and H. Liu, "Jpeg image tampering localization based on normalized gray level co-occurrence matrix," *Multimedia Tools and Applications*, vol. 78, pp. 9895–9918, 2019.

[24] B. Chen, X. Qi, Y. Wang, Y. Zheng, H. J. Shim, and Y. Shi, "An improved splicing localization method by fully convolutional networks," *IEEE Access*, vol. 6, pp. 69 472–69 480, 2018.

[25] Q. Wang and R. Zhang, "Double jpeg compression forensics based on a convolutional neural network," *EURASIP Journal of Information Security*, vol. 2016(1), pp. 23–30, 2016.

[26] Y. Zhang, J. Goh, L. L. Win, and V. Thing, "Image region forgery detection: A deep learning approach," *Proceedings of the Singapore Cyber-Security Conference (SG-CRC)*, vol. 14, pp. 1–11, 2016.

[27] B. Liu and C. M. Pun, "Locating splicing forgery by fully convolutional networks and conditional random field," *Signal Process., Image Communication*, vol. 66, pp. 103–112, 2018.

[28] J. H. Bappy, C. Simons, L. Nataraj, B. S. Manjunath, and A. K. Roy Chowdhury, "Hybrid lstm and encoder-decoder architecture for detection of image forgeries," *IEEE Transactions on Image Processing*, vol. 28(7), pp. 3286–3300, 2018.

[29] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[30] W. Abdulla, *Mask r-CNN for object detection and instance segmentation on Keras and TensorFlow*. RCNN, 2017. [Online]. Available: https://github.com/matterport/Mask

[31] W. Li, Y. Yuan, and N. Yu, "Passive detection of doctored jpeg image via block artifact grid extraction," *Signal Processing*, vol. 89, pp. 1821–1829, 2009.

[32] J. Dong, W. Wang, and T. Tan, "Casia image tampering detection evaluation database," in *IEEE China Summit and International Conference on Signal and Information Processing, Beijing*, 2013, pp. 422–426.

[33] V. Christlein, C. Riess, J. Jordan, C. Riess, and E. Angelopoulou, "An evaluation of popular copy-move forgery detection approaches," *IEEE Transactions on Information Forensics and Security*, vol. 7(6), pp. 1841–1854, 2012.

Multi-Channel Muscle Armband Implementation: Electronic Circuit Validation and Considerations towards Medical Device Regulation Assessment

Martha Rocio Gonzales Loli[1], Elsa Regina Vigo Ayasta[2], Leyla Agueda Cavero Soto[3], Jose Albites-Sanabria[4]
Universidad Nacional Federico Villarreal, Lima, Peru[1, 2, 3]
Universidad Cientifica del Sur, Lima, Peru[4]

***Abstract*—Multi-channel muscle arrays are commonly used as sensors in bionic prosthetic devices offering an innovative solution to recover motion in transradial amputees. This study presents preliminary assessments towards validation of a muscle armband for usage in transradial users. Analog and digital components were designed based on medical agencies' recommendations to assess future compliance with Latin American medical device regulations. The study follows two approaches, an exploratory and pre-experimental design. Design was validated and confronted among research literature and medical device regulations. For validation, a pre-experimental design was guided by a quantitative paradigm. Muscle signal was assessed before and after the condition circuit for up to four muscle signals in real time. The present study considers both the conditioning muscle signal circuit and the embedded logic implementation to record signals from the designed muscle armband. Results show that the proposed device allows to record noninvasive signals with a frequency from 20-500 Hz.**

Keywords*—Component; muscle armband; surface electromyography; medical device regulation; transradial users

I. Introduction

According to the US Department of Health and Human Services, 200 to 600 people lose a limb every day by accident or as consequence of a disease [1]. According to Dr. Luis Farro, in Peru, amputations by accident are most prevalent on adults and young adults with a prevalence of superior amputations [2]. Main causes of hand amputations are accidents with fireworks and occupational accidents, within which the fireworks industry manufacturing has the highest percentage.

The main population segment affected by hand amputations are laborers or informal workers who usually receive a minimum wage. This type of serious accident, such as an amputation, has an important psychological impact in patients and their family. Losing the ability to perform activities of daily living, losing their autonomy, losing their role in the society, being unable to perform his trade, plunges amputees into a great frustration and depression. The injuries in this type of amputation are quite different depending on of each case. Hence, customization is an important aspect to consider in the design, as it must be perfectly adapted to the injury user specific. Available commercial solutions must comply with international and local medical devices regulations.

Currently, the Peruvian and Latin American market for limb amputations offer different types of solutions such as mechanical, electrical, myoelectric and cosmetics devices.

Mechanical prosthesis can be manufactured locally; however, due to traditional manufacturing processes, the cost of these prostheses, although not as high as myoelectric ones, is still too much for the available budget of patients. Just the device terminal (hand) of the mechanical prosthesis costs over two thousand dollars and manufacturing the prosthetic socket to suit the type of amputation of the patient is a separate cost. These prostheses are offered for forearm amputations and wrist disarticulation. The National Rehabilitation Institute (INR) offers the manufacturing service of prosthetic sockets for the two types of amputation mentioned above and the costs are on average 300 to 500 dollars depending on your insurance program.

Nevertheless, mechanical prosthetic devices are very limited in their functionality offering as much as two or three different hand movements, e.g. open and close hands to grasp items. Companies that offer myoelectric or bionic prostheses are very limited and usually imported from European countries. These devices have high costs that make it very inaccessible for frequent amputation patients with tag prices over fifty thousand dollars. An important part of the bionic prosthetic devices are sensors placed on remain parts of the patient's body, usually muscle sensors.

Analysis of surface muscle activity or electromyographic signals (sEMG) is widely used in conditions that involve diagnosis, treatment, and rehabilitation of various motor disabilities [3].

Muscle sensors measure the electrical activity generated by contracting or relaxing one or more muscle groups. A muscle sensor contains conductive elements, also called electrodes, which capture these signals. EMG activity can be recorded both non-invasively (electrodes placed on the skin) and invasively, with electrodes implanted inside nerves or around muscle fibers.

Muscle signals can be recorded in monopolar or bipolar form. Monopolar configurations refer to the use of an electrode disposed over the muscle to be measured and a reference electrode preferably located away from the muscle or in an inert area. Bipolar configurations, in contrast, are comprised of two electrodes and a reference. Usually the

electrodes are arranged on the muscle of interest at 1-2 cm each to avoid capturing interference from other muscle groups, also knowns as crosswalk. The advantage of the latter configuration lies on the reduction of common noise between both electrodes [4].

Superficial muscle signals produce voltages that vary between 0-10 mV and contain a useful range of information between 20-500 Hz. It is recommended that the signals obtained by the electrodes (especially dry electrodes) be arranged with an analog muscle conditioning circuit. This is because a weak signal that travels large distances is susceptible to the inherent noise through the cable in which they travel. Electrodes containing a preprocessed circuit of the captured signal encapsulated in conjunction with the electrode are known as active electrodes.

When recording muscle signals, usually more than one muscle group needs to be analyzed simultaneously. This impose a particular challenge for portable systems. Components of the embedded system must be carefully chosen; microcontrollers need to be chosen carefully to suffice the requirements usually based on low power consumption while offering high and efficient processing demands.

II. Methods

The present study follows two approaches, an exploratory and pre-experimental design.

In the first phase, components and designs were validated and confronted among research literature and medical device regulations.

The pre-experimental design was guided by a quantitative paradigm. Muscle signal was assessed before and after the condition circuit for up to four muscle signals in real time. Muscle armband was based on dry electrodes in bipolar configuration following methodology developed in [5].

To validate the device. A BK Precision 4040B voltage signal generator was used to vary the input of the signal conditioning circuit. Sinusoidal signals were used with frequencies from 10 to 1000 Hz. Selected frequencies were derived after first phase analysis and discussion.

A fined machine state algorithm guided the study protocol to record muscle signals from two non-amputee users. Users were asked to perform four different gestures. Training and test data was recorded on repeated instructions. Each set of instructions lasted at least 180 seconds. A general session took about 30 minutes, the participant could stop at any requested time.

III. Results

Literature of over fifty research articles was revised from Scopus and Web of Science database. Literature was complemented with European and Latin American analysis of international and local medical device regulations. Table I shows main findings for device design considerations.

The European project Surface EMG for non-invasive assessment of muscles (SENIAM) proposed recommendations towards effective surface muscle signal recordings. This study takes into account these recommendations for electrode distance and positioning [6]. In addition, several studies have applied a similar configuration as proposed in Table I with successful outcomes, numbers of sensors placed on forearm vary from 2 to 8 in different studies [7,8]. The proposed system is capable of up to 8 muscle sensors recording in real time.

An embedded system was designed cable of recording up to four muscle sensors in real time. Table II shows basic requirements of the microcontroller based on input estimation and signal and classification processing load.

The above described processor characteristics will be responsible for digital signal conversions of up to four muscle signals simultaneously, signal processing of different muscle features, and muscle decoding of different hand gestures.

The signal conditioning circuit was separated into five stages: differential amplification (instrumental), high pass filter, low pass filter (1st stage), customized amplification, low pass filter (2nd stage). Logic implementation was based on four main states: reading, training, classification, emergency stop flag.

Characteristics of the recorded signal were given by the nature of a muscle signal, which has a main spectra band between 20-500 Hz [9,10]. In order to avoid aliasing, a 2 kHz sampling frequency was selected.

Data was acquired through eight ADC ports of the processors and saved in an array with the following structure:

[CANAL 1];[CANAL 2];[CANAL 3] . . . . [CANAL 8]; [Timestamp]

Main information was comprised of each of the muscle signals and a timestamp based on clock frequency of the processor. Fig. 1 shows how data could be read once digitized by the processor.

TABLE I. Design Considerations

Categories	**Observation**
Electrode type	Dry electrodes, stainless steel 316
Configuration	Bipolar to address common noise
Electrode distance	1-2 cm to avoid crosswalk
Protection	Isolation circuit to prevent reverse current
Power	Low power consumption, high precision components -
Number of sensors	4 to 8 muscle sensors around the arm

TABLE II. Micro-Controller Design

Features	**Details**
ADC resolution	12 bits
ADC channels	4
Voltage	1.8 V a 3.3 V
Communication	I2C or UART
Timers	2 timers
Clock frequency	Over 100 MHz

```
1;0;0;44;2688;1993;2463;1013;10205
4;0;7;49;2690;1854;2434;543;10206
4;0;25;58;2688;1806;2410;542;10207
4;4;61;100;2683;1838;2423;634;10208
2;36;109;152;2684;1874;2431;761;10209
3;211;352;400;2688;1960;2438;1113;10210
4;467;709;802;2690;2116;2494;1766;10211
6;603;927;1079;2693;2250;2519;2325;10212
5;717;1113;1341;2688;2364;2534;2756;10213
5;901;1374;1722;2688;2486;2558;2773;10214
```

Fig. 1. EMG Data Record Test.

Several tests were performed in order to validate acquired data. In order to prevent packet lost, header and check-sum methods were implemented. Fig. 2 presents the method implementation on c-code.

UART protocol communication was chosen for data transmission. The data array showed above summed a total of 22 bytes. A sampled frequency of 2KHz meant information must be transferred in less than 500 milliseconds. Given those requirements, baud rate was chosen by mathematical calculation. For instance, a baud rate of 115200 Bauds, would allow a transference of 22 Bytes in 1528 ms; this would not met the 2kHz sampled required frequency. In contrast 460800 Bauds, would allow packet transference in 382 ms which suffice system demands.

The ARM Cortex M4 offers direct memory access (DMA) for peripheral ports which offers time efficiency and does prevent the system from blocking execution while reading signals [11].

Signal classification, feature extraction and decoding algorithms were implemented in a previous study. The same study also provides information to validate the logic implementation behind the system.

Fig. 3 shows ARM Cortex M4 implementation to record fours EMG signals simultaneously from the arm of a healthy user.

```
void setTrama(void)
{
        trama[0]='S';
        trama[1]='T';
        checksum=getChecksum(valuesSIM_ADC);
        trama[2]=checksum&0xFF;
        trama[3]=checksum>>8;
        for(int i = 0;i<8;i++)
        {
                trama[2*i+4]=valuesSIM_ADC[i]&0xFF;
                trama[2*i+5]=valuesSIM_ADC[i]>>8;
        }
        trama[20]=23;
        trama[21]=98;
}
uint16_t getChecksum(uint16_t values[])
{
        uint16_t rslt = values[0];
        for(int i=1;i<8;i++)
        {
                rslt=rslt^values[i];
        }
        return rslt;
}
```

Fig. 2. C-Method Implementation.

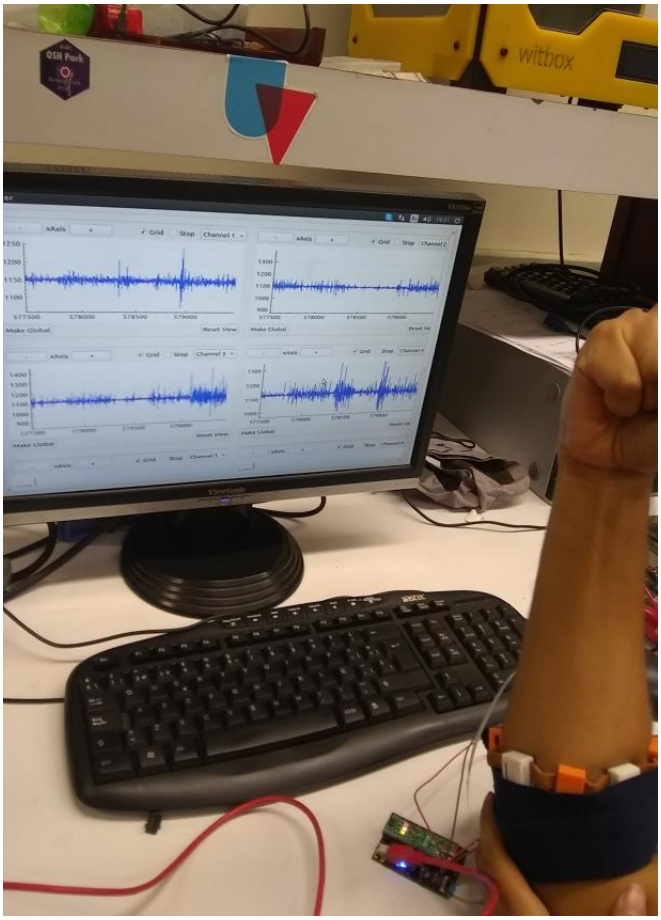


Fig. 3. Real Time Muscle Signal Recording.

Output of the instrumental amplifier and its variability with different power supplies are measured, several tests are performed controlling the offset to the circuit input.

Fig. 4 shows basic logic functionality of the system. The implemented machine state algorithms handle each of the states through several on/off flags and functions calls on each state.

After the start command, parameters and weights from the model are trained and cross validated off-line in a computer. Once the system is trained, continuous classification is performed. An emergency stop is comprised of an external input that can interrupt and stop the system at any given time.

The first test is performed using a signal with randomly chosen characteristics of 356Hz, 6.5mV, Senoidal, input offset: 0 + -1mV. The offset is varied in 5 mV intervals.

It can be seen that the offset of the signal is amplified by the instrumental amplifier in proportion to its gain (11), in general offsets are suppressed largely by a high pass filter; however, this filter leaves a small offset of up to 24mV approximately. Which is amplified in the following stages where this offset becomes very noticeable.

Results obtained in Table III are comparable to other studies [12,13,14] and thus reaffirm the potential viability of the system.

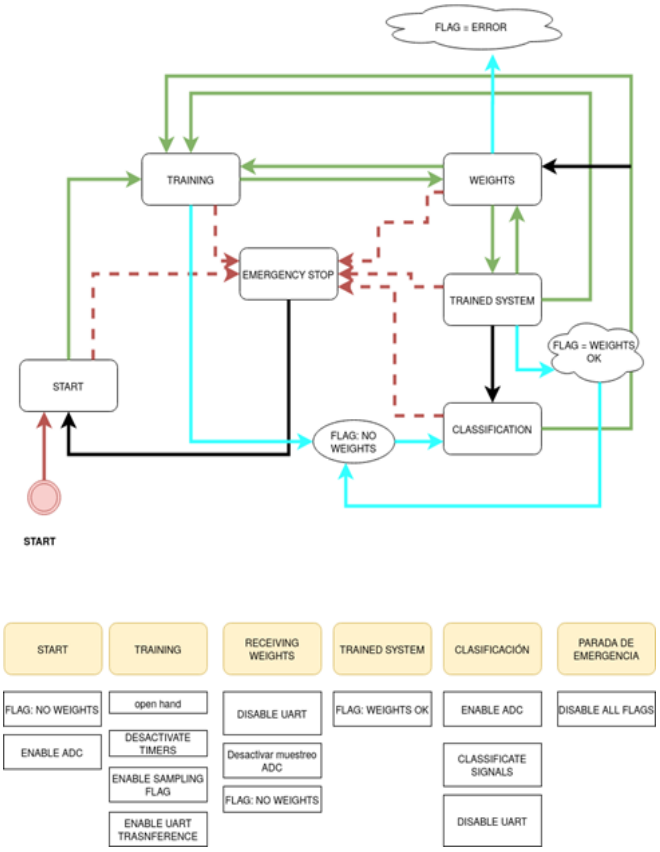


Fig. 4. Finite Machine State Implementation.

TABLE III. ANALOG TEST VALIDATION

Input (Hz)	A	B	C	D
1 Hz	2 mV	2 mV	20 mV	20 mV
5 Hz	7 mV	8 mV	100 mV	200 mV
40 Hz	52 mV	90 mV	3 V	4.5 V
100 Hz	88 mV	100 mV	4 V	5 V

According to the medical and healthcare regulation authorities, muscle sensors can be classified as class I. Device system must not only comply with technical aspects such as circuit isolation and electrode medical material grade. Usage protocols and instructions are critical for device safety. Guides should include corrective and preventive actions to diminish harms in patients' health. Product development must also take into consideration some of the following regulations according to 21CFR:

- Registration: 807.
- Medical device listing: 807.
- Premarket Approval: 804.
- Investigational device exemption: 812.
- Quality system regulation: 820.
- Medical device reporting: 803.

IV. CONCLUSIONS

The present study considers a muscle processing circuit and the embedded logic implementation to record signals from an electromyographic sensor in compliance with local and international medical device regulations.

The present work seeks to contribute to the state of the art of EMG sensors from a technical point of view. The final goal in any sensor lies on its easiness of use and massive scale while meeting the minimum requirements for its use in both clinical and non-clinical setups.

One of the motor disabilities with the highest rate of people affected is the amputation of one of the members caused by disease or trauma [15]. The proposed solution for users who suffer some type of amputation is the replacement of the limb by an artificial limb also known as a prosthesis.

The development of prostheses in recent years has focused on improving the functionality and appearance of the artificial limb. Today, there are different prosthetic alternatives for users. From purely cosmetic alternatives, mechanical prostheses and recently bionic prostheses usually controlled by muscle or myoelectric signals. Different companies sell this type of prosthesis (BeBionics, iLimb, Touch Bionics, Vincent) with prices starting at 60 thousand dollars [16].

The field of prosthetic development is constantly growing, both at the research level and at the commercial level. Within the Peruvian Background, the biomechanics department of the National Institute of Rehabilitation develops mechanical prostheses with a cost of 7 thousand soles.

There are also initiatives that promise to reduce prosthetic costs through additive manufacturing. Within the framework of bionic prostheses, there are research projects aimed at the development of robotic prostheses in Peru by groups from different universities. Multidisciplinary and transdisciplinary skills are necessary in order to achieve results and analysis coming from specialists from different areas such as medicine, engineering, law, business among other areas [17,18]. In addition, research in these field is in compliance with sustainable development goals for 2030, especially in Peru most research is being made in sustainable water treatments [19]; however, disabilities and unemployment must be also taken into account.

However, the future work of these projects as well as the costs landed towards a commercial proposal are not fully defined until the time of writing this work. One of the important components remains the muscle sensor to be used in these prostheses, accompanied by classification algorithms that allow interpreting the movements of the person so that he can open or close the hand just by thinking about it [20].

The design of advanced and adaptive classification algorithms based on convoluted networks and non-linear filters using different signal characteristics can improve the signal [21] and its study is open to potential students interested in the use of electromyographic sensors for study and application of these in the improvement of the quality of life of people with motor disabilities. Selection and

implementation of the processor logic and architecture must go hand-by-hand in order to meet all above described criteria.

ACKNOWLEDGMENT

The authors would like to acknowledge medical, technical, and law experts who contributed with their expertise in the different aspects of this project.

REFERENCES

[1] Ümenapf, G., & Morbach, S., “Amputation statistics—how to interpret them?” Deutsches Ärzteblatt International, 114(8), 128, 2017.

[2] Farro, L., Tapia, R., Bautista, L., Montalvo, R., & Iriarte, H., “Características clínicas y demográficas del paciente amputado” Revista Médica Herediana, 23(4), 240-243, 2012.

[3] Albites Sanabria, J. L., “Braingate-Enabled Intracortical Control of Commercial Tablet Computers by Individuals with Tetraplegia” Doctoral dissertation, Brown University, 2017

[4] Beck, T. W., Housh, T. J., Cramer, J. T., Malek, M. H., Mielke, M., Hendrix, R., & Weir, J. P., “A comparison of monopolar and bipolar recording techniques for examining the patterns of responses for electromyographic amplitude and mean power frequency versus isometric torque for the vastus lateralis muscle” Journal of neuroscience methods, 166(2), 159-167, 2007

[5] Albites-Sanabria, J., “Analysis and design of an in-house low cost dry EMG sensor for bionic transradial prosthesis” International Journal of Scientific and Technology Research, 9(4), pp. 2919-2921, 2020.

[6] Stegeman, D., and H. Hermens. "Standards for surface electromyography: The European project Surface EMG for non-invasive assessment of muscles (SENIAM).", 2007.

[7] Takala, Esa-Pekka, and Risto Toivonen. "Placement of forearm surface EMG electrodes in the assessment of hand loading in manual tasks." Ergonomics, 56(7), 1159-1166, 2013

[8] Reinvee, Märt, and Mati Pääsuke. "Overview of contemporary low-cost sEMG hardware for applications in human factors and ergonomics." Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Vol. 60. No. 1. Sage CA: Los Angeles, CA: SAGE Publications, 2016.

[9] Shahzad, Waseem, et al. "Enhanced Performance for Multi-Forearm Movement Decoding Using Hybrid IMU–sEMG Interface." Frontiers in neurorobotics 13, 2019.

[10] Tang, Xueyan, et al. "Hand motion classification using a multi-channel surface electromyography sensor." Sensors, 12(2), 1130-1147, 2012.

[11] Alvarado-Diaz Witman, Meneses-Claudio Brian, Fiorella Flores-Medina, Patricia Condori, Natalia I. Vargas-Cuentas and Avid Roman-Gonzalez, “Acquisition and Classification System of EMG Signals for Interpreting the Alphabet of the Sign Language” International Journal of Advanced Computer Science and Applications(IJACSA), 10(8), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100868

[12] Heywood, Sophie, et al. "Low-cost electromyography–Validation against a commercial system using both manual and automated activation timing thresholds." Journal of Electromyography and Kinesiology, 42, 74-80, 2018.

[13] Pancholi, Sidharth, and Amit M. Joshi. "Electromyography-based hand gesture recognition system for upper limb amputees." IEEE Sensors Letters, 3(3), 1-4, 2019.

[14] Mastinu, Enzo, Max Ortiz-Catalan, and Bo Håkansson. "Analog front-ends comparison in the way of a portable, low-power and low-cost EMG controller based on pattern recognition." 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2015.

[15] Shambulinga M and G. Sadashivappa, “Supervised Hyperspectral Image Classification using SVM and Linear Discriminant Analysis” International Journal of Advanced Computer Science and Applications(IJACSA), 11(10), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0111050

[16] Hofmann, M., Harris, J., Hudson, SE, & Mankoff, J., “Helping hands: Requirements for a prototyping methodology for upper-limb prosthetics users” In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 1769-1780). ACM, 2016.

[17] Ayala, C. E., Garro, L. L., Sanabria, F. M., Aldana, J. J., Colina, F. J., & Albites, J. L., “Competencias interculturales en el proceso de formación en investigación en una universidad privada de Lima, Perú” Revista Espacios, 40(44), 2019.

[18] Simbron-Espejo, S. , Sanabria-Boudri, F. , Colina-Ysea, F. , Albites-Sanabria, J., "University faculty research training and performance: A case from peru" Universal Journal of Educational Research, 8(11), pp. 5053-5060, 2020.

[19] Rios-Garay, J., Cisneros-Pariona, L., Albites-Sanabria, J., “Analysis and efficiency of an affordable in-house anaerobic digester for water treatment” International Journal of Emerging Trends in Engineering Research, 8(9), pp. 6012-6016, 2020.

[20] Xu, W., Sun, Y., & Li, C., “Research on EMG Signal Extraction Based on Biomimetic Hand” In 7th International Conference on Management, Education, Information and Control (MEICI 2017). Atlantis Press, 2017.

[21] Altin, C., & Er, O., “Comparison of different time and frequency domain feature extraction methods on elbow gesture's EMG” European journal of interdisciplinary studies, 2 (3), 35-44, 2016.

A Novel Machine Learning based Model for COVID-19 Prediction

Tamer Sh. Mazen
Lecturer Dept. Management Information System
Modern Academy of Computer Science and Management, Cairo, Egypt

***Abstract*—Since end of 2019, the World Health Organization (WHO) provided the name COVID-19 for the disease caused by the novel coronavirus. Coronavirus is a family of viruses that are named according to the spiky crown existed on the outer surface of the virus. The novel coronavirus, also known as SARS-CoV-2, which is a contagious respiratory virus that first reported in Wuhan, China. According to the rapid and sudden spread for COVID-19, it attracts most of the scientists and researchers all over the world. Researchers in the data science field are trying to analyze the worldwide infection cases day-by-day to gain a complete statistical view of the current situation. In this paper, a novel approach to predict the daily infection records for COVID-19 is presented. The model is applied for Egypt as well as the highest 10 ranked countries based on the number of cases and rate of change. The proposed model is implemented based on supervised Machine-Learning Regression algorithms. The dataset used for prediction was issued by WHO starting from 22 Jan 2020.**

***Keywords*—Coronavirus; COVID-19; coronavirus in Egypt; supervised machine learning; regression models**

I. Introduction

Since the end of 2019, the outbreak of COVID-19 began in Wuhan, China. The new virus is a form of Coronaviruses, that affects the respiratory system such as the SARS virus. COVID-19 consists of a protein membrane with a diameter of 50-200-200 nm, inside which the DNA of the RNA virus is enveloped, which forms the spinal bumps on the surface of the virus and gives it a distinctive coronary shape [1], [2]. "Fig. 1", shows the internal structure of SARS-COVID virus [3].

Rational decisions are the goal that governments seek so as to address the COVID-19 epidemic. The prediction process is one of the most important tools needed to face that problem. The prediction models are used to predict the number of new daily confirmed cases, recoveries, and deaths. The prediction of newly confirmed cases helps the governments to update their precautionary procedures as well as getting ready by the needed hospitals equipment and the human preparation.

Nowadays, the main target is to find a cure for the killing virus as well as to predict its spread rate. Many researches in the data science field were found to study the statistical situation of the virus.

Furqan Rustam et al. [4] tested a set of different Machine Learning based models in order to predict the number of future COVID-19 patients. The used models are linear regression, least-absolute shrinkage and selection operator, support vector machines, and exponential smoothing. Results showed that the exponential smoothing based model provided the most accurate prediction results, while the support vector machines provided the worst results as compared to the four selected models.

Nanning Zheng et al. [5] proposed a new hybrid Artificial Intelligence (AI) based model using Natural Language Processing (NLP), and the Long-Short Term Memory (LSTM) network, in addition to the Improved-Susceptible Infected (ISI) model, presented so as to predict the future cases in China. The proposed model could predict to a high degree the actual epidemic-cases.

In [6] Li Yan et al. proposed a machine learning based model using 3 clinical parameters to detect the new death rate of current patients. The accuracy of the proposed model is more than 90%.

Renato R. Silva et al. [7] used a Bayesian based methodology in order to detect the peak of the outbreak in one of the Brazilian countries (Goias) based on the number of confirmed cases. They found that, the peak will be reached between 7 to 10 weeks from the beginning of the crisis supposing that, there will be no change in the governmental control during the upcoming period.

In this paper, a novel prediction model that predicts the number of new confirmed cases is presented. The proposed model uses a set of statistical based techniques in a supervised machine learning process. The model is tested on Egypt as well as the top 10 ranked countries for COVID-19 till end of September 2020. The results of the proposed model are compared against the Bayesian Ridge regression model.

The next sections of the paper will be as follows. In Section 2, the distribution of COVID-19 all over the world is presented. Section 3, shows COVID-19 in Egypt. The proposed model is explained in Section 4 followed by the experimental results in Section 5. Finally, the conclusion is presented in Section 6.

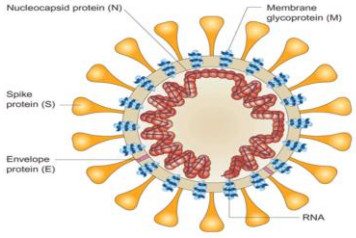


Fig. 1. SARS-COVID Structure.

II. State of the Art

As a subset of Data Science, Artificial Intelligence (AI) and Machine Learning (ML) are playing a major role in the analysis and visualization of the COVID-19 crisis. These predictions will provide a help to the healthcare systems and government institutions to speed up investigations about the virus rapid and terrible spread.

"Fig. 2" illustrates the distribution for COVID -19 case all over the world. During the interval starting from 22 January till end of September 2020, the number of confirmed cases, deaths, and recovery cases are (38,917,803), (1,098,254), and (26,885,286) consequently.

"Table I" and "Fig. 3", show the number of confirmed cases, deaths and recovery cases over the world grouped by continents. As seen, Europe is the most affected continent followed by Asia, America, Africa and finally Australia.

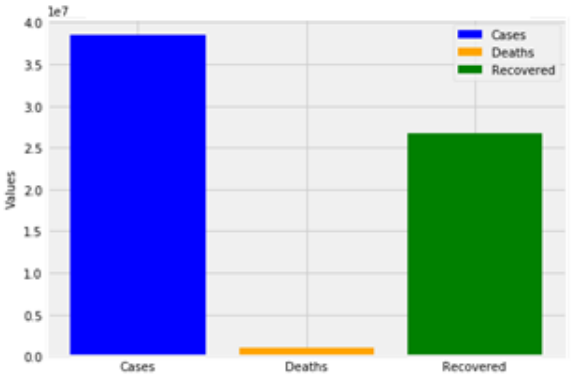


Fig. 2. COVID-19 Distribution All Over the World.

TABLE I. COVID-19 CONFIRMED, DEATH AND RECOVERED CASE OVER THE WORLD TILL 30 SEPTEMBER 2020

Continent	Confirmed	Deaths	Recovered
Africa	10450293	38407	1285804
America	4672236	384875	8993415
Asia	13986946	214996	10341346
Europe	8584922	238143	2867398
Australia	748801	938	27438

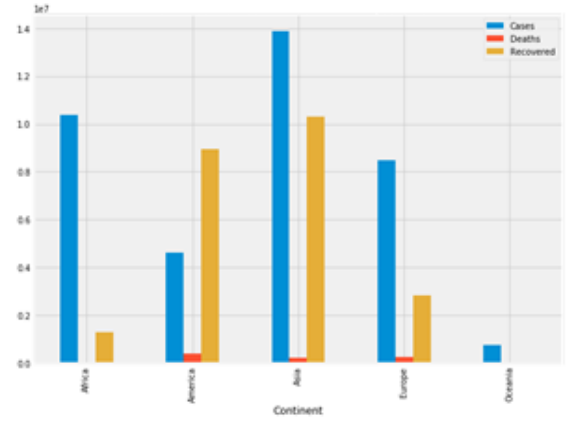


Fig. 3. COVID-19 Distribution over the World/Continent.

"Table II" and "Fig. 4", show the number of confirmed cases, deaths, and recovery cases for the mostly infected 10 countries classified by WHO on 30 September. "Table III" and "Fig. 5", show the rate of change percentage of the top 10 countries.

TABLE II. TOP 10 COUNTRIES BASED ON COVID-19 CONFIRMED CASES

Country	Confirmed	Deaths	Recovered
USA	7916099	216872	3155794
India	7307097	111266	6383441
Brazil	5140863	151747	4526393
Russia	1332824	23069	1035141
Argentina	931967	24921	751146
Colombia	930159	28306	816667
Spain	908056	33413	150376
Peru	853974	33419	753959
Mexico	829396	84898	703489
France	820376	33058	106374

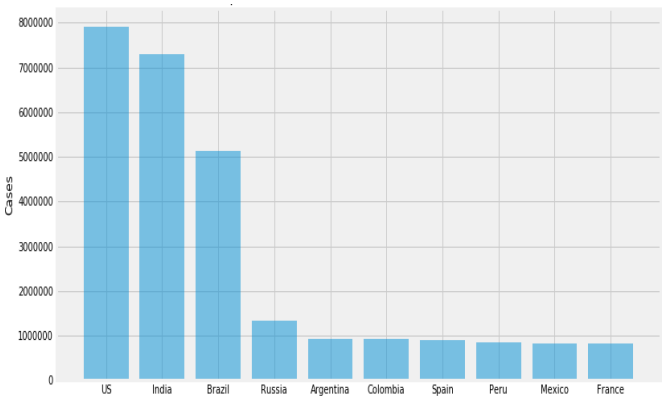


Fig. 4. Top 10 Countries on Deaths Cases in COVID-19.

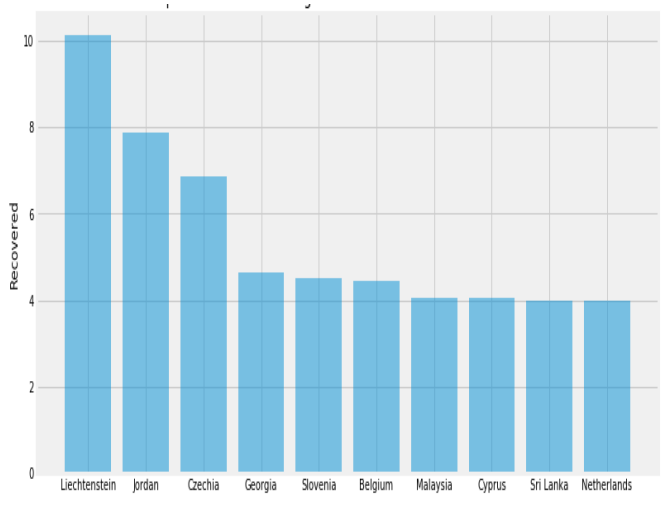


Fig. 5. Top 10 COVID-19 Countries based on Rate of Change.

TABLE III. TOP 10 COVID-19 COUNTRIES BASED ON RATE OF CHANGE

Country/Region	Rate of change %
Liechtenstein	10.14
Jordan	7.88
Czechia	6.86
Georgia	4.64
Slovenia	4.52
Belgium	4.44
Malaysia	4.07
Cyprus	4.05
Sri Lanka	4.00
Netherlands	3.98

III. COVID-19 IN EGYPT

In this paper, the rank of Egypt based on the number of confirmed cases and the rate of change is calculated. It is found that, Egypt's rank is 43 around the world based on the number of infections. while, its rank based on the change in rate is 143. These calculations are performed using the WHO dataset till end of September 2020.

"Table IV" and "Fig. 6" show COVID-19 cases in Egypt till 30 September 2020.

There are different regression models found in the literature to predict the number of new confirmed COVID-19 cases such as Support Vector Machines (SVM) [8], Linear regression [9], binomial regression [10], and Bayesian Ridge regression [11].

During our experiments, it was found that the Bayesian Ridge regression model could provide the most accurate prediction results as compared to the other techniques.

TABLE IV. COVID-19 CONFIRMED, DEATH AND RECOVERY CASES IN EGYPT

Country	*Confirmed*	*Deaths*	*Recovered*
Egypt	*104915*	*6077*	*97920*

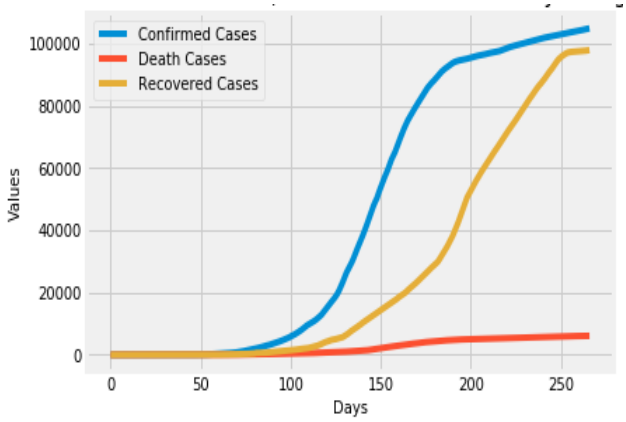


Fig. 6. COVID-19 Cases in Egypt.

IV. PROPOSED PREDICTION MODEL

The main idea of this paper is to build a hybrid model based on mathematical and statistical approaches in a machine learning based environment. The model is performed using rate of change, geometric mean and standard deviation [12] [13] [14] [15].

As seen in "Fig. 7", the data set consisting of the number of confirmed COVID-19 cases [16] , number of deaths [17], and the number of recovery cases [18] is collected. Then, the data preprocessing is performed and the data is split into 80% training and 20% testing. The steps of the proposed model are explained in details with an example on Egypt.

"Table V" shows the sample of data used as an example where, X represents the days starting from the 110th to the 119th day from the start of the epidemic while Y represents the number of confirmed cases on each day (each of the numbers below is multiplied by 10^3).

A. Steps of the Proposed Model

Step 1: Splitting the data set

Dataset was split into a training set and testing set, 80%, and 20%, respectively.

Step 2: Calculating the number of newly confirmed

Where new cases are calculated as,

$$New_{Case[i]} = y[i] - y[i-1] \quad (1)$$

Where, $y[i]$ is the number of confirmed cases at day i.

Step 3: Calculating the Rate of Change (RoC):

The Rate of Change (RoC) for the newly confirmed cases is calculated by the next formula [19].

"Table VI" shows the RoC calculation for the sample example.

$$RoC = \frac{New_{Case[i]}}{y[i-1]} \quad (2)$$

Step 4: Calculating the Geometric mean (GM)

It is a n^{th} root for the RoC for n days [20] [21].

$$GM = \left(\prod_{i=1}^{n}(RoC^2)\right)^{1/n} \quad (3)$$

Here, in this example the GM = 0.149956677.

TABLE V. SAMPLE OF COLLECTED DATA ABOUT EGYPT

X	*110*	*111*	*112*	*113*	*114*	*115*	*116*	*117*	*118*	*119*
Y	*9.4*	*9.7*	*10*	*10.4*	*10.8*	*11.2*	*11.7*	*12.2*	*12.7*	*13.4*

Step 5: Calculating the standard deviation,

Standard Deviation "STD" is calculating the extent of deviation of values from the average value:

$$STD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2} \quad (4)$$

Where, n is the number of days.

Here, in this example the STD = 0.1334

Step 5: Calculating the newly expected cases and boundaries

The new expected cases based on the proposed model are calculated using the following formula:

$$Exp_{Case} = present_{newcase} * 1 + GM \quad (5)$$

$$Exp_{Lower} = Exp_{Case} - (Exp_{Case} - STD) \quad (6)$$

$$Exp_{Upper} = Exp_{Case} + (Exp_{Case} - STD) \quad (7)$$

"Table VII" represent results of steps 4, 5 and 6 calculating GM, SD and boundaries.

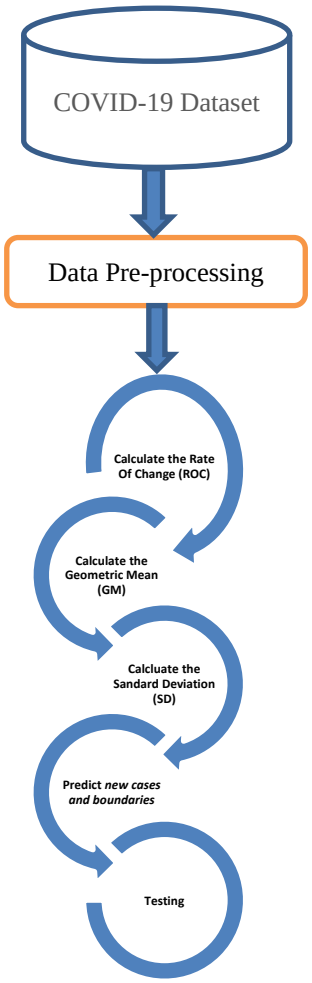


Fig. 7. The Proposed Model Diagram.

TABLE VI. THE ROC CALCULATION

X	110	111	112	113	114	115	116	117	118	119
Y	9.4	9.7	10	10.4	10.8	11.2	11.7	12.2	12.7	13.4
New Case	-	346	347	338	398	399	491	510	535	720
RoC			0.03	0.02	0.17	0.03	0.23	0.08	0.04	0.34

TABLE VII. THE EXPECTED NEW CASES AND BOUNDARIES

X	110	111	112	113	114	115	116	117	118	119
Y	9.4	9.7	10	10.4	10.8	11.2	11.7	12.2	12.7	13.4
New Case	-	346	347	338	398	399	491	510	535	720
RoC			0.03	0.02	0.17	0.03	0.23	0.08	0.04	0.34
Case			348	334	418	415	529	546	571	588
Lower					327	372	376	467	488	516
Upper					341	465	543	591	604	627

B. Testing the Model Accuracy

The proposed model accuracy is tested using both the Mean Square Error (MSE) [22] [23] and the correlation (R) between the expected values and the real values [24] [25].

$$MSE = \frac{\sum_{i=1}^{n}(E^2)}{n} \quad (8)$$

Where E is the difference between expected and real data.

Here, in the example the $MSE = \frac{33016}{7} = 4716.571429$.

$$r = corr(Exp_{Case}, New_{Case}) \quad (9)$$

Here, in the example $r = 0.880848$

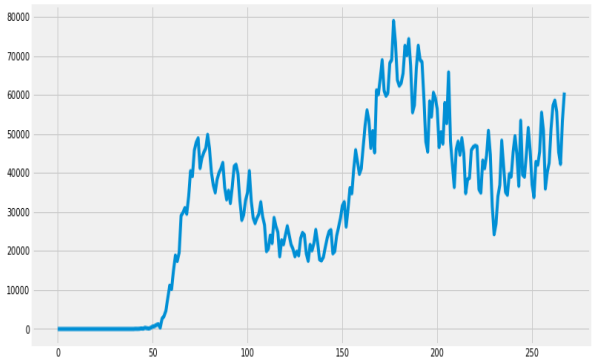
There is a strong relation between results that indicate to model has highest accuracy.

V. EXPERIMENTAL RESULTS

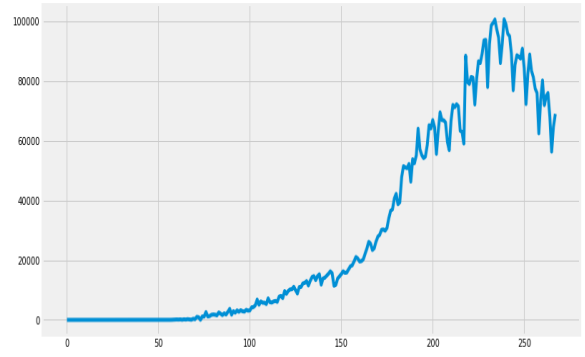
The proposed model is compared against the Bayesian Ridge regression model, as it was most accurate model for COVID-19 predictions amongst the other state of the art techniques.

"Fig. 8" illustrates the daily cases that predicted by proposed model for the highest rated for COVID-19 till end of September 2020. These 10 highest countries are USA, India, Brazil, Russia, Argentina, Colombia, Spain, Peru, Mexico, and France.

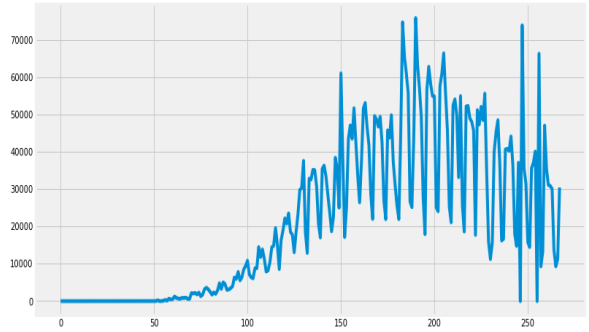
"Fig. 9" illustrates the comparison between the proposed model versus the Bayesian Ridge model applied for Egypt as well as the 10 highest rated countries for COVID-19 till end of September 2020. The red lines in the figure represent the daily prediction results while the blue lines represent the real values. As seen, the proposed model is more accurate than its counterpart over all the countries.



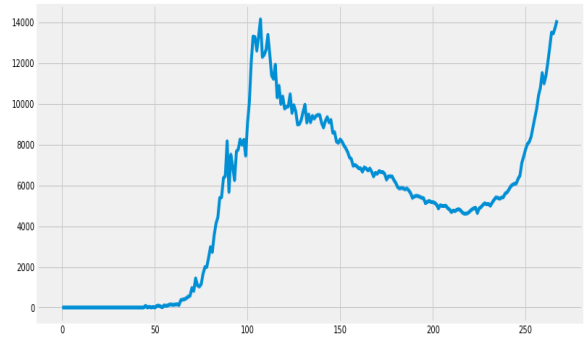
(a) USA



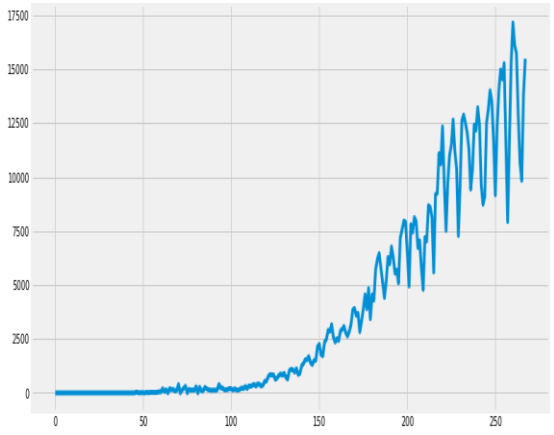
(b) India



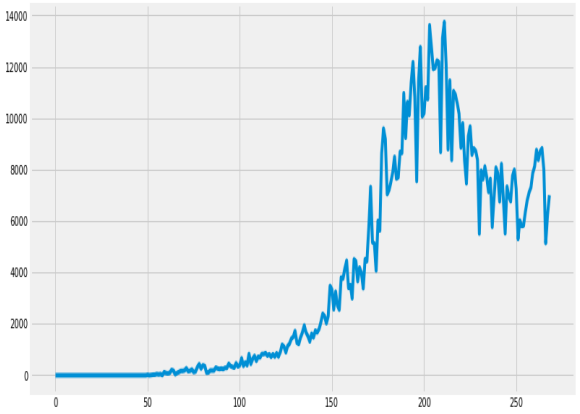
(c) Brazil



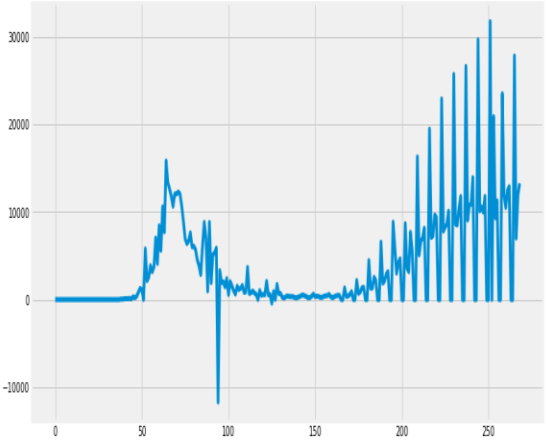
(d) Russia



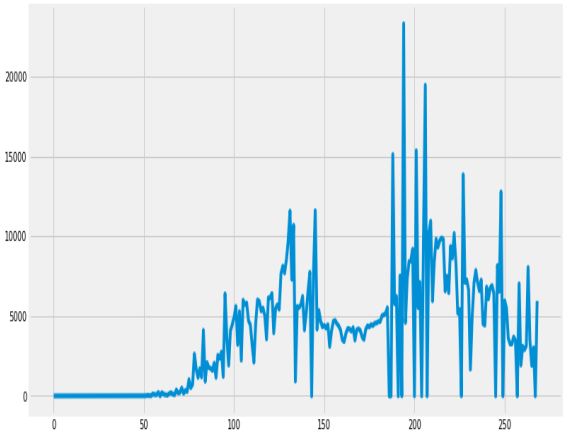
(e) Argentina



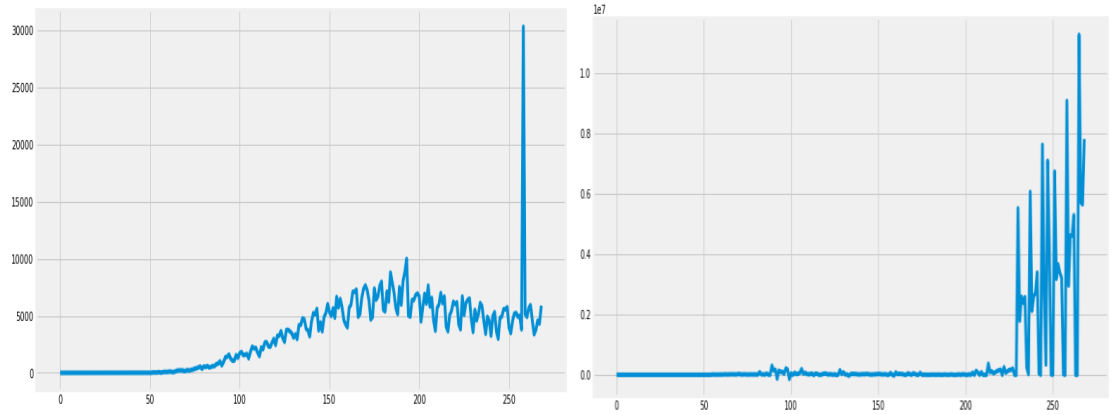
(f) Colombia



(g) Spain

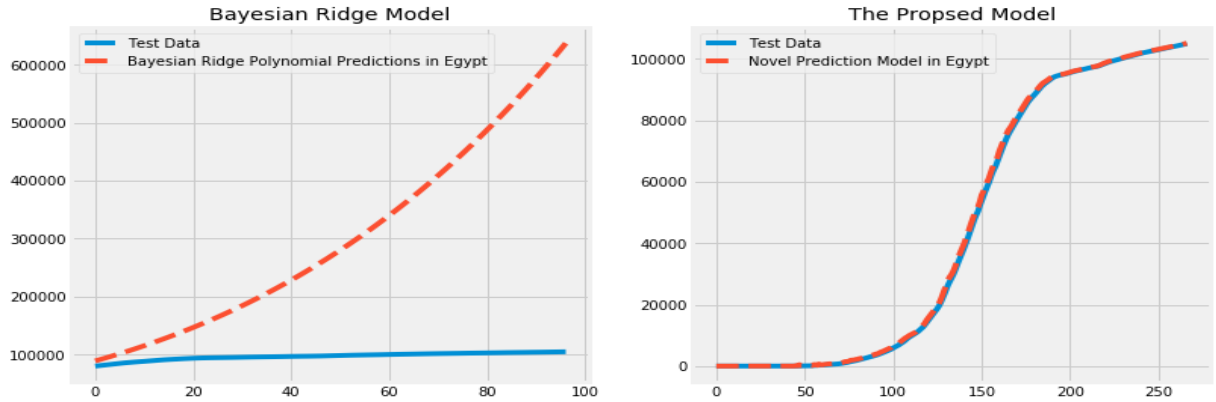


(H) Peru

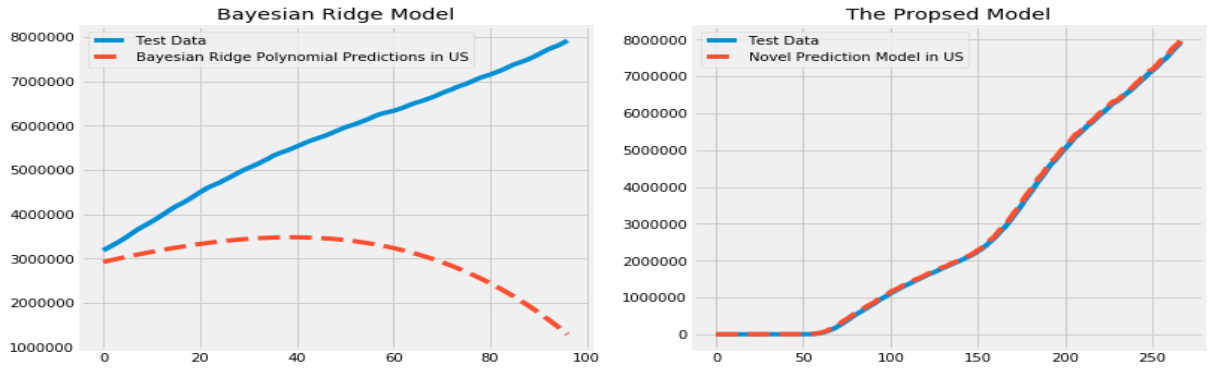


(i) Mexico (j) France

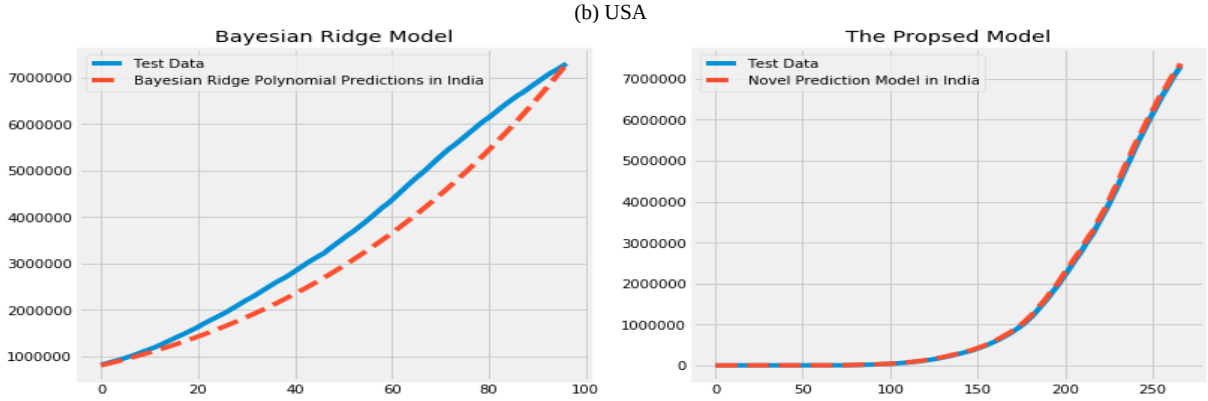
Fig. 8. Daily Prediction Cases used Model.



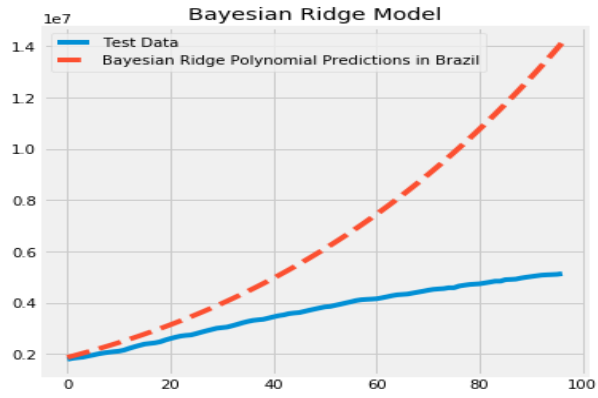
(a) Egypt

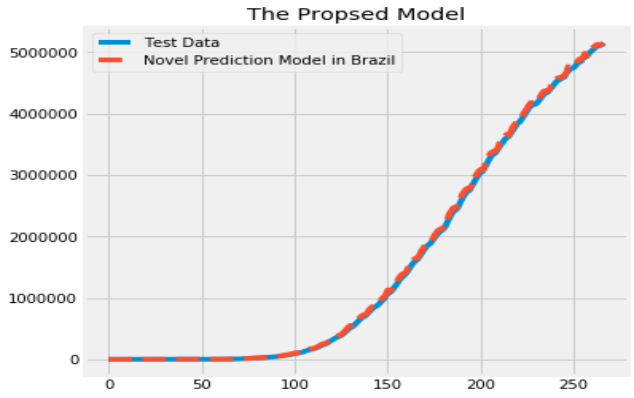


(b) USA

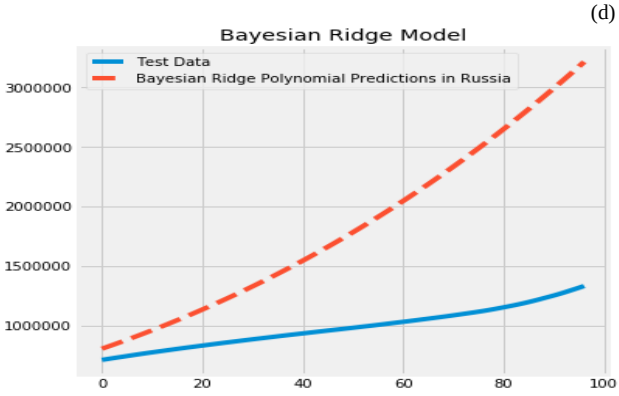


(c) India

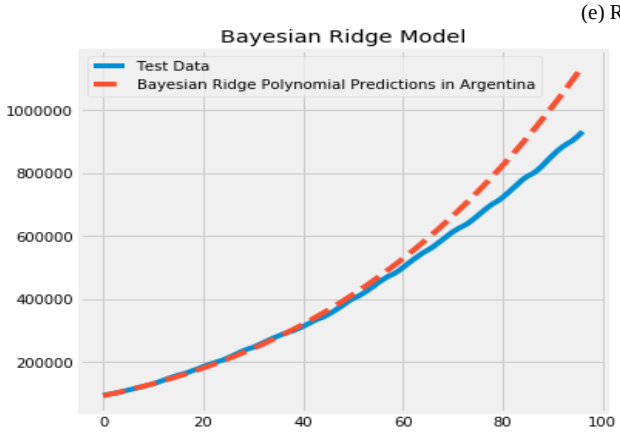


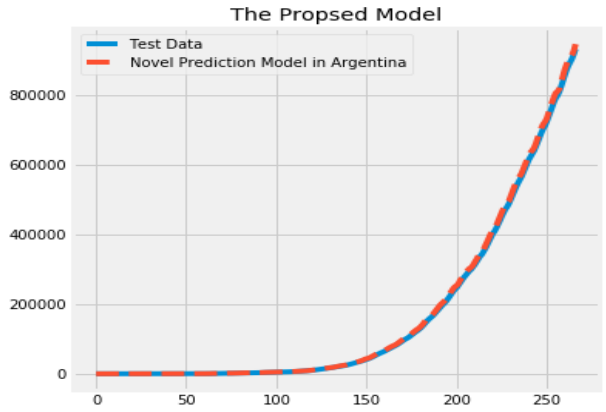


(d) Brazil

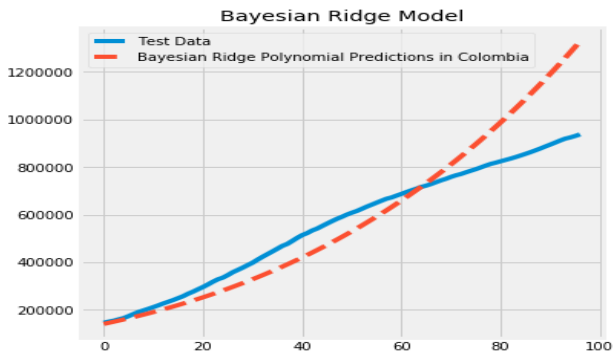


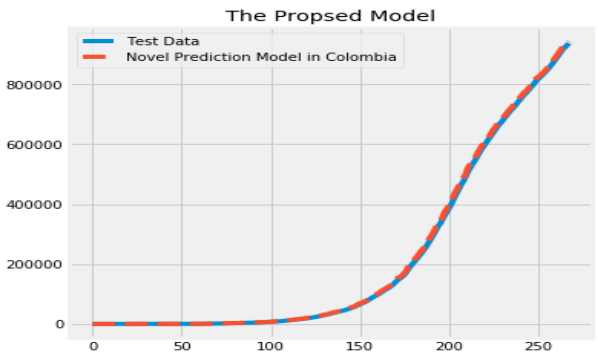
(e) Russia



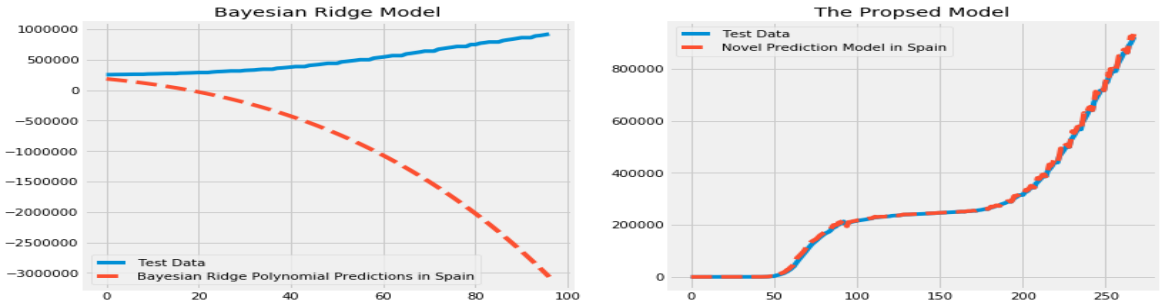


(f) Argentina

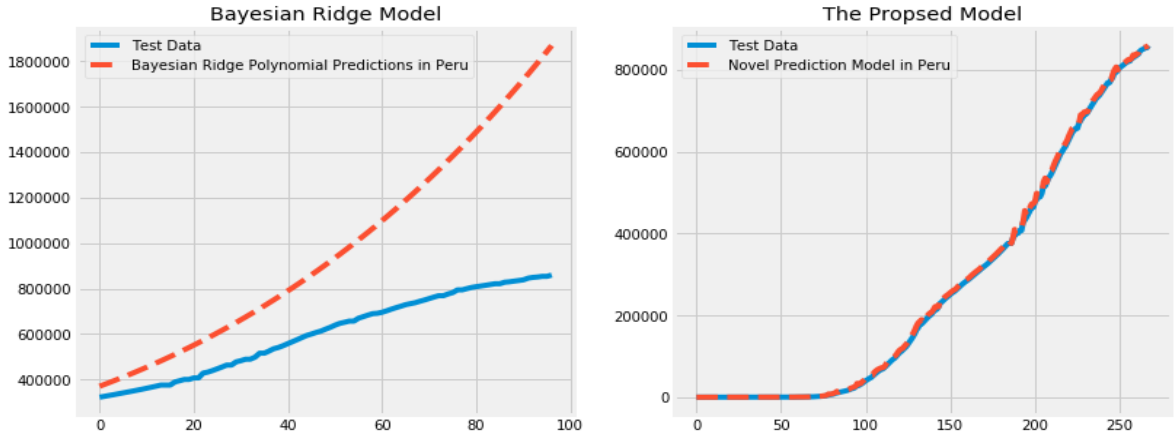




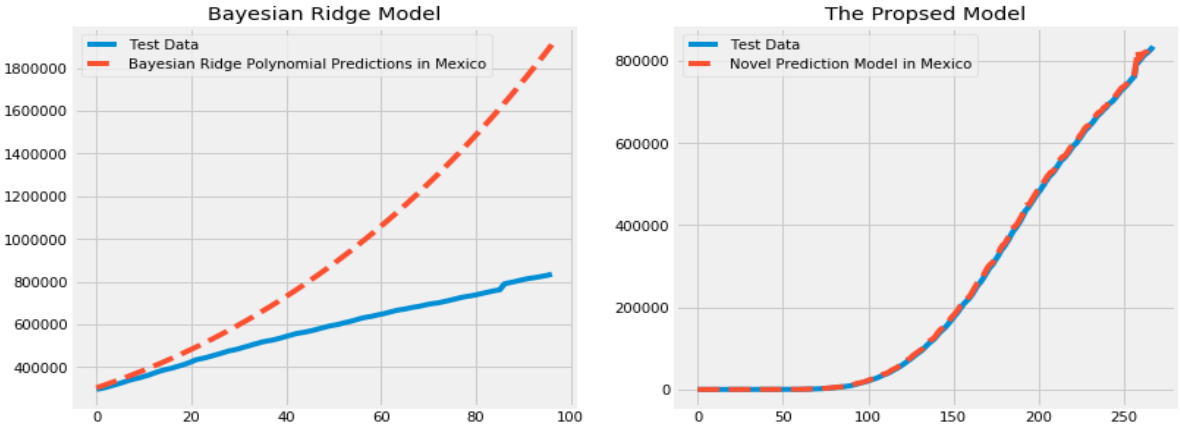
(g) Colombia



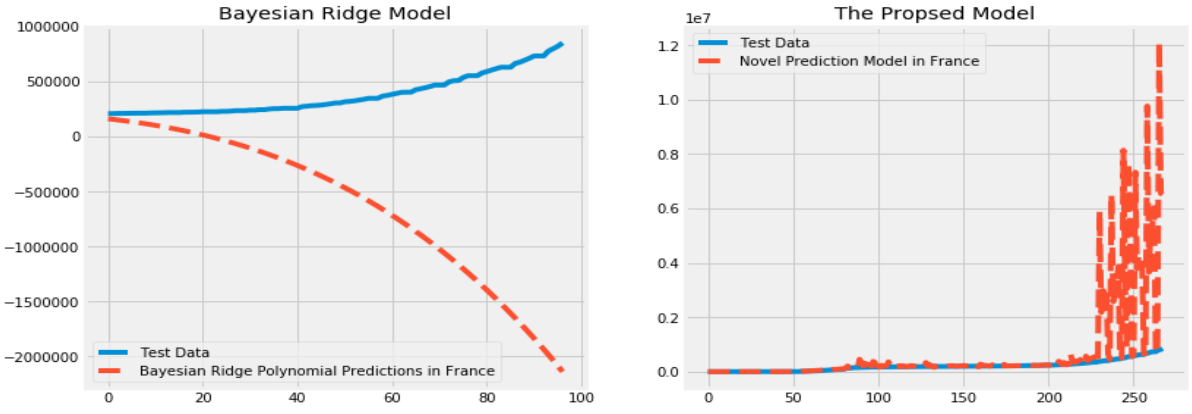
(h) Spain



(i) Peru



(j) Mexico



(k) France

Fig. 9. Bayesian Ridge Model vs the Proposed Model Daily Predictions for Egypt as well as the Highly Rated 10 Countries.

The sharp peaks found for the prediction results corresponding to France were according to the sudden jump in the number of cases during that period.

The Mean Square Error (MSE) for the two models are presented in "Table VIII" as seen, the MSE for the proposed model over all the testing countries is less than that of Bayesian Ridge model.

TABLE VIII. MODELS EVALUATION COMPARISON MSE

Country	Bayesian Ridge	Proposed Model
	MSE	MSE
Egypt	13.57	11.44
US	29.19	17.03
India	22.74	16.21
Brazil	26.16	20.23
Russia	26.29	15.55
Argentina	19.19	14.73
Colombia	26.18	13.77
Spain	18.52	14.89
Peru	22.76	14.18
Mexico	26.86	15.64
France	26.00	19.29

VI. CONCLUSION

COVID-19 or corona virus pandemic is the danger that threaten both peoples and governments all over the world. Many researches tried to predict the number of newly infected cases, deaths, and recoveries. In this paper, a new hybrid-machine learning based model is proposed so as to predict the newly expected infections. The model is tested on Egypt as well as the 10 highly rated COVID-19 countries till end of September 2020. The proposed model is compared against one of the most accurate prediction models found in the literature i.e. Bayesian Ridge model. Results showed the powerful of the proposed model as compared to its counterpart all over the countries under study.

REFERENCES

[1] Yan Gao et al., "Structure of the RNA-dependent RNA polymerase from COVID-19 virus," Science, vol. 368, no. 6492, pp. 779-782, 15 May 2020.

[2] S. Anastasopoulou and M. Athanasia, "The biology of SARS-CoV-2 and the ensuing COVID-19," ACHAIKI IATRIKI, vol. 39, no. 1, p. :29–35, 2020.

[3] Stadler K, Masignani V, Eickmann M, et al., "SARS--beginning to understand a new virus," Nat. Rev. Microbiol., vol. 1, no. 3, p. 209–218, 2003.

[4] F. Rustam, A. MEHMOOD, A. RESHI , S. ULLAH, B.-W. ON4, W. ASLAM and G. S. CHOI , "COVID-19 Future Forecasting Using Supervised Machine Learning Models," IEEE Access , vol. 8, pp. 101489 - 101499, 25 May 2020.

[5] N. Zheng, S. Du, J. Wang, H. Zhang, W. Cui, T. Yang, B. Lou, Y. Chi, H. Long, M. Ma, Q. Yuan and S. Zhang, "Predicting COVID-19 in China Using Hybrid AI Model," IEEE Trans Cybern., 8 May 2020.

[6] Yan, L., Zhang, H., Goncalves, J. et al., "An interpretable mortality prediction model for COVID-19 patients," Nature Machine Intelligence, vol. 2, pp. 283-288, 2020.

[7] R. Silva, . W. D. Velasco, d.-S. W. Marques and C. A. G. Tibiica, "A Bayesian analysis of the total number of casesof the COVID 19 when only a few data isavailable. A case study in the state of Goias, Brazil," ,10.1101/2020.04.19.20071852, 2020.

[8] V. Jakkula, "Tutorial on support vector machine (svm)," School of EECS,Washington State University, 2006.

[9] Y. S. I. a. P. A. Mokhade, "Use of Linear Regression in Machine Learning for Ranking," IJSRD - International Journal for Scientific Research & Development, vol. 1, no. 5, 2013.

[10] S. Sperandei, "Understanding logistic regression analysis," Biochemia medica, vol. 24., pp. 12-8, 2014.

[11] W. a. Bruna, "Bayesian Linear Regression," 2019, 2019.

[12] D. J. S. ,. T. A. W. ,. J. D. C. ,. J. J. C. David R. Anderson, Quantitative Methods for Business 13 edition, Cengage Learning, 2015.

[13] C. S. M. S. Heumann, Introduction to Statistics and Data Analysis, Springer, 2016.

[14] W. K. H. O. Okhrin, Basic Elements of Computational Statistics, Springer, 2017.

[15] J. A. S. Betty R. Kirkwood, Essential Medical Statistics, Blackwell Science, 2003.

[16] WHO, "Confirmed Case , https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv ,"

[17] WHO, "Death, https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv".

[18] WHO, "Recoverd, https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv".

[19] D. Kirchman, "Calculating microbial growth rates from data on production and standing stocks," Marine Ecology-progress Series , 2002.

[20] J. &. L. Y. Lawson, "The Geometric Mean, Matrices, Metrics, and More," The American Mathematical Monthly, 2001.

[21] W. Y. a. M. Warshaure, "Arithmeyic and Gemometric Mean," Menemeni Matematik(descoviring mathematics), vol. 27, no. 2, pp. 17-22, 2002.

[22] T. &. D. Chai, "Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature," Geoscientific Model Development, vol. 7, 2014.

[23] K. M. Cort J. Willmott, "Advantages of the mean absolute error (MAE) overthe root mean square error (RMSE) in assessingaverage model performance," Climate Research Clim Re, vol. 30, pp. 79-82, 2005.

[24] P. &. B. C. &. S. L. Schober, "Correlation Coefficients: Appropriate Use and Interpretation," Anesthesia & Analgesia, 2018.

[25] A. D. Bland JM, Correlation, regression, and repeated, BMJ, 1994.

COVID-19 Transmission Risks Assessment using Agent-Based Weighted Clustering Approach

P.Vidya Sagar[1], T. Pavan Kumar[2], G. Krishna Chaitanya[3], Moparthi Nageswara Rao[4]
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India

***Abstract*—Coronavirus is a pandemic disease spreading from human-to-human rapidly all over the world. This virus is origin from common cold to severe disease such as MERS-CoV and SARS-CoV. Initially it was identified in China, December 2019. The main aim of this research is used to identify the COVID-19 transmission risks assessment from human-to-human within a cluster. The agent-based weighted clustering approach is used to identify the corona virus infected people rapidly within a cluster. In the weighted clustered approach, the normal agents are consisted as susceptible node and the corona virus infected people are considered as malicious node. The Cluster Head (CH) is elected based upon some weighting factors and the trust value is evaluated for all the agents within the cluster. The cluster head were periodically transfers the malicious node information to all other nodes within the cluster. Finally, the agent-based weighted clustering machine learning model approach is used to identify the number of corona virus infected people within the cluster.**

***Keywords*—COVID-19; machine learning; weighted clustering; malicious node; susceptible node; head; trust**

I. Introduction

A corona virus was initially recognized in human lungs in 2012. The novel corona virus is not same as Severe Acute Respiratory Syndrome (SARS) in 2003. However, similar the SARS virus, the novel corona virus is most related to those originate in bats. In 2012 the novel corona virus is happened in two clustered regions like Jordan and Saudi Arabia. Now World Health Organization (WHO) announced corona virus disease 2019 (COVID-19) is a pandemic. A pandemic defines spreading the disease wide range of area and affecting exceptionally high proportion of the population. This novel corona virus was named Corona virus Disease 2019 (COVID-19) by WHO in February 2020. The virus is referred to as SARS-CoV-2 and the associated disease is COVID-19. As of 10September 2020, over 28,050,253 cases have been identified globally in 188 countries with a total of over 908,434 fatalities. Also 20,117,616 were recovered. The primary symptoms for Corona virus are mild fever, Fatigue, Aching muscles, Breathing problem, Dry cough along with less typical symptoms of Headache, Diarrhea, Phlegm buildup and Hemoptysis [24]. The person is having all above symptoms then the person is affected with COVID-19 virus. The virus gets into human lungs and it affect lung functionality with the impact increases up to 14 days. The corona virus can transferred through droplets with different particle size. Respiratory droplet particle sizes are >5-10μm and droplet nuclei particle size is <5μm. The respiratory droplets are spreading easily through direct contact compare to droplet nuclei. The droplet transmission occurs within 1m direct contact with COVID-19 infected people [25].

The shape of the Corona virus is shown in Fig. 1. The gray surface is a spherical envelope that surrounds the nucleus of the virus, containing genetic material. Orange bits are a "membrane proteins," or M proteins, the most abundant structural protein in the virus and one that gives it form, says Eckert. These and other proteins vary from one type of virus to another and can be used to help understand or identify one virus from another. Yellow bits are envelope proteins (E proteins), the smallest of the structural proteins. They "play an important role either in regulating virus replication — such as virus entry — assembly and release," research. Red spikes: These clumps of proteins (called S proteins) are "what the virus uses to gain entry into and attach to the cell," says Eckert. They also create the effect of a halo, or corona, around the virus.

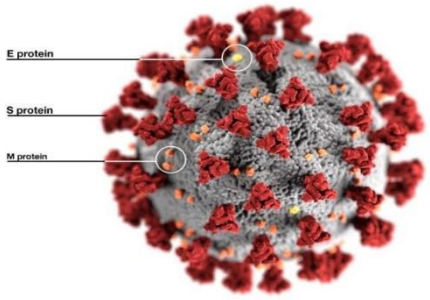


Fig. 1. The Shape of COVID-19 Virus.

II. Literature Review

To reduce the spreading of corona virus with the help of recognizing cases and clusters, patient isolation, contact tracing, and community transmission prevention. The small number of clusters which identify 22 probable primaries COVID-19 cases rapidly [1]. The agent-based model is used to identify the contact rates between agents and structure of in-person contact network [2]. The agent based fine-grained computational simulation model is used to identify the COVID-19 cases from children to adult. The simulation model compares numerous interference policies isolation, air travel, home quarantine and social distancing [3]. The COVID-ABS, a new SEIR agent-based model is used to simulate the pandemic situation from agent-agent contact, business, and government. The COVID-ABS model was implemented in

python programming language [4]. The INFEKTA agent-based model combines individual agent virus spread to complex network with Euclidean space is measured within a city [5].

The REINA agent-based model which is used to simulate different categories of plan action on timeline. The result combines that utilizing complete testing, contact tracing and targeted isolation measures [6]. The cluster based mathematical model to forecast the rough trail of COVID-19. The COVID-19 spread is analyzed for three countries like Italy, United States of America and India the results show that the spread of each country is high accuracy [7]. The agent-based model considered few parameters like social distance restrictions, business opening, quarantine, control approaches on the infection progression. The result shows that the social distancing restricts the business activity participation [8]. The agent-based model applies for cluster planning and each agent is allocated for three key attributes like intelligence, talkativeness, and credibility. The problem-solving ability is better for small groups compared to large groups [9]. The fully computable model is utilized for two-state model which is used to identify the healthy or infectious people permitting for in environment simulation and risk assessment [10]. The two crucial methods are utilized for spread of an infected disease. They are agent-based model and equation-based model. The result shows that the equation-based model gives better performance compared to agent-based model [11].

The agent-based model is a feasible and powerful modeling tool for both biology and mathematics classrooms [12]. The Covasim (COVID-19 agent based simulator) is an open source model to include demographic information on age structure and population size, social distance, schools, workplaces, hygiene measures etc. to apply and inspect virus dynamics and policy decisions in European countries [13]. The simulation model could help the individuals to take better decision during COVID-19 pandemic situation. The simulation decision maker provides better decision-making results for individuals [14]. The COVID-19 infected people spread their virus within their family and the person mobility infection goes to other families to form a new cluster [15]. The virus infection outbreak is influenced by many factors like immunity level, population density and age structure of the population. These factors are considered to evaluate COVID-19 risk assessment using agent-based model [16]. The individuals are considered as agent that move, become infected and spread the virus to others. The simulation model is used to restrict the agent movement and mandatory to wear a mask on the spread of COVId-19 [17]. The RT-PCR testing is used to diagnosis the COVID-19 virus rapidly. In cluster-based approach the COVID-19 positive cases are identified using RT-PCR testing in Singapore [18,19]. The age based social contact virus spread is assessed using simulation model. The result shows that mid-elder age people affected more compared to child or young people [20,21]. The Bats-Hosts-Reservoir-People transmission network model is used to identify the human infection. Reservoir-People (RP) transmission network model is used to assess the transmissibility of the SARS-CoV-2 [21,22]. The mathematical model is used to assess the transmission risk of COVID-19 in various facilities. The agent-based simulation model which is used to take better decision for prevention of COVID-19 [22,23]. The synergic deep model is used to learn and predict various metrics like duration of days, discharge disposition, and inpatient expense for total hip arthroplasty [24,25]. The attribute-based health record protection algorithm is used to protect healthcare service information access like control confidentiality, credibility, and secrecy [26,27].

III. Methodology

A. Transmission Mode of COVID-19 Virus

CoVID-19 virus spread can be classified into two categories: close contact and Airborne. Lung infections can be spread through droplets of dissimilar sizes: while the droplet particles are >5-10 μm in diameter they are stated to as respiratory droplets, and when they are <5μm in diameter, they are stated to as droplet nuclei. The COVID-19 virus spread primarily spread between publics through respiratory droplets and contact routes.

The droplet spread occurs between a person is in close contact within 1m with someone having respiratory symptoms. The transmission also occurs through infected person surrounding environment. So, the virus spread happened in two ways: direct and indirect contact. In direct contact the infected people contact other people directly and in indirect contact the people touching with virus surfaces indirectly. The direct and indirect contacts are shown in Fig. 2 and 3.

Airborne transmission is dissimilar from droplet spread as it refers to the occurrence of bacteria within droplet nuclei, which are usually measured to be particles <5μm in diameter, can persist in the air for long periods of time and be spread to others over distances larger than 1 m.

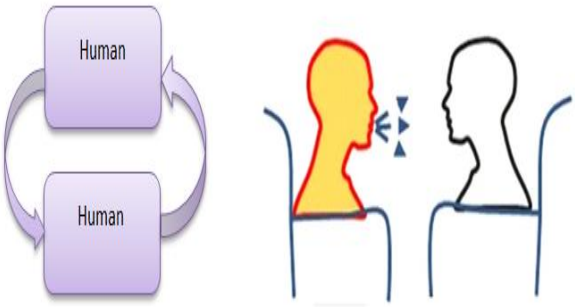


Fig. 2. Direct Contact or Close Contact.

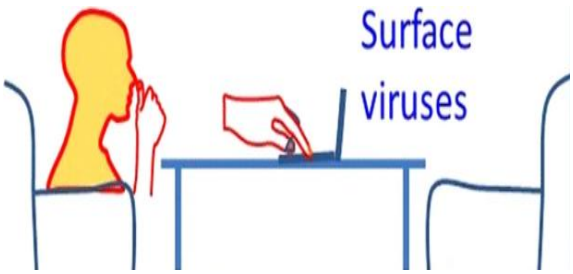


Fig. 3. Indirect Contact.

B. Agent based Weighted Clustering Approach

The agent based weighted clustering proposed model is used to find the transmission risk within a cluster. Every individual are considered as agent to perform predefined operations. The agents are interacting with the environment in multidimensional space. The agents can freely move from one cluster to another cluster freely. Here clusters are considered as city. Every cluster have cluster head which is elected based upon the agent weighting parameters like immunity, age group, and mobility. The cluster head is used to monitor the agents within the cluster. Within a cluster the normal people are considered as normal node and corona virus infected people are considered as malicious node. If any malicious node is found within a cluster, then the malicious node is removed from the cluster immediately.

The set of A agents [a_1,….a_A.] Each agent a_i, (i belongs to1…. A) is selected randomly. Each agent a_i is applied for set of rules to change its position, state or relationship with agents. In the proposed model two different types of agents A(k)= {a_1(k),…$a_{A(K)}$} and B(k)= {b_1(k),…b_B(K)} are defined. The agent A and B change their position in each iteration k of the simulation. The agent A signifies the susceptible elements in the environment and B signifies the infected individuals. The aagents A and B behaviors are characterized by two rules to simulate COVID-19 transmission. They are Rule I and Rule II. The Rule I defines the agent a_i infected or not. The Rule II defines mobility of the agent is identified.

Rule I: The random number r is generated between 0 and 1. If the value of r is less than or equal to threshold value, then the agent a_i is considered to be infected, otherwise it is not affected. The value of R can represent the radius of the cluster range. Here value of R=1mt. When the agent a_i is recognized as infected, a_i is deleted from A and added as new agent b_{new} within the infected agents B. Fig. 4 illustrates the operation process of Rule I [28,29].

In Fig. 4 set of 8(A) susceptible elements and 2 infected agents A= a_1,…, $a_8$and B= b1,b2. In figure a_3 and a_5 maintain close relationship between b_1 and b_2. Assume the probability of infection for a_3 and a_5 are 0.2 and 0.9, respectively. The probability of a_5 infected value is high and $a_3$is value is low compared to predefined threshold value 0.5. This virus can be identified with the help of following formula:

$$COVID-19\ Diagnosis = \begin{cases} COVID-19+Veif\ (1).5 < T \\ COCID-19-veif\ 0.5 \geq T \end{cases} \quad (1)$$

Where, T is threshold value. The agent infected value is less than threshold value then it considered as COVID-19 positive otherwise it is negative [30,31].

Rule II: In this rule mobility between the agents A and B determined. This is illustrated in Fig. 5.

In Fig. 5, set of 2(A) susceptible agents and 1(B) infected agent A= a_1, a_2 and B= b1. In fig the mobility probability of a_1= 0.8, a_2=0.1 and b_1=0.1. Due to low mobility probability of a_2 is very low. So, it is not affected. But the contact and mobility probability of a_1 and b_1 is high. So the infected probability chance is high for both a_1 and b_1. In figure, a_1 mobility is considered for a_1(k) to a_1(k+1) and b1 mobility consisted for b_1(k) to b_1(k+1). The a_2 mobility is same position [32,33].

The following weighting parameters are considered for electing a cluster head within a cluster. They are mobility of the agent, immunity level of the agent and age group of the agent.

Cluster Head=W1 (MA) + W2 (IA) + W3 (AA) (2)

Where W1 (MA) = Weighing factor of Mobility of the Agent

W2 (IA) = Weighting factor of the Immunity level of the agent

W3 (AA) = Weighting factor of the Age group of the agent

The agent has low mobility, high immunity level and the middle age group, then the agent considered as cluster head for within a cluster. The cluster head is a health inspector agent to monitor the other agent behaviors periodically. If any malicious agent node is found within a cluster, then the node the node is removed from the cluster immediately. The trust value between two agents can be represented as T_{xy}. The trust value between two agents can be calculated as below equation.

T_{xy}= LM+WM+MSD (3)

Where T_{AB}=Trust value between agent A to B

LM = Low Mobility

WM = Wearing Mask

MSD = Maintain Social Distance

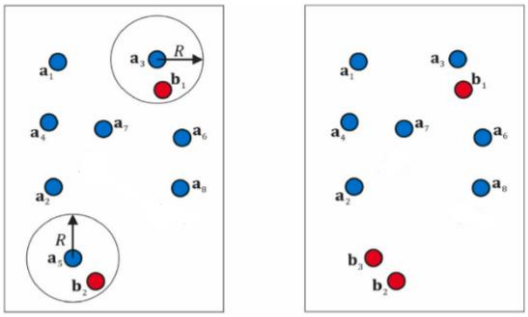


Fig. 4. Operation of Rule I (a) Initial Configuration and (b) Final Configuration.

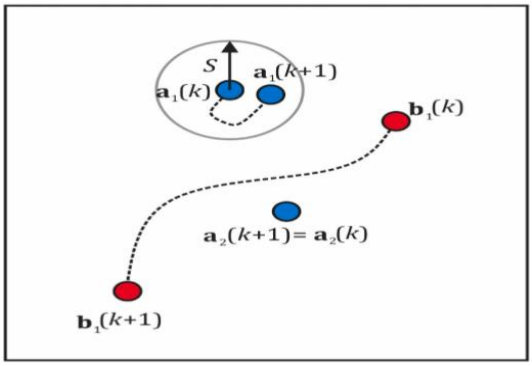


Fig. 5. Operation of Rule II.

The above three parameters are satisfied between agents A and B, then the agent A trust B. The virus infection can be identified within a cluster with the help of Aarogya Setu App in Fig. 6. In Fig. 6(a) agents is installed Aarogya Setu App using Bluetooth to identify the COVID+ agent. In Fig. 5(b) identify the location of COVID+ agent and send alerts message to all other agents within a cluster. If the agent is not having Bluetooth feature mobile phone, then the trust value will be calculated for that particular agent.

In Fig. 7, the cluster head (health official) to monitor all other agents periodically within a cluster. If any agent is identified COVID+, then the infected agent removed from the cluster immediately and the infected agent history is recorded in the application server. The COVID+ agent alert information is passed to all other agents immediately within cluster. The agent based weighted clustering algorithm works as follows:

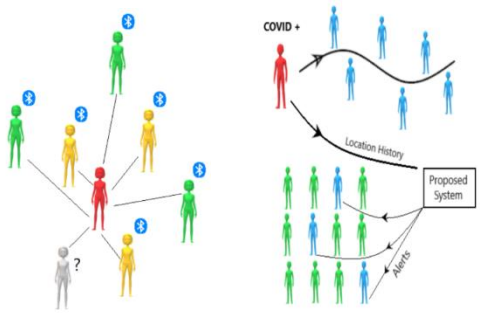
Agent based weighted clustering algorithm

1. Initialize the number of agents and clusters

   A=susceptible Agent, B= Infected Agent
2. Elect a cluster head with weighting parameters

   Cluster Head (CH) = W1 (MA) + W2 (IA) + W3 (AA)
3. Find the trust value between two agents A and B

   T_{xy}= LM+WM+MSD
4. Install Aarogya Setu App for all the agents within a cluster
5. If the agent is not having Bluetooth features mobile then
6. Calculate the trust value of that particular agent
7. Apply the operation of Rule I
8. Check the agent infection result is COVID +ve or COVID –ve
9. If COVID +ve then
10. Remove the infected agent from the A group and added into B group
11. Else
12. Continue with A group
13. Apply the operation of Rule II: Mobility
14. Check the agent infection result is COVID +ve or COVID –ve
15. If COVID +ve then
16. Remove the infected agent from the A group and added into B group
17. Else
18. Continue with A group
19. Stop



(a) (b)

Fig. 6. (a) Agent using Aarogya Setu App using Bluetooth (b) COVID + agent Location Discovery.

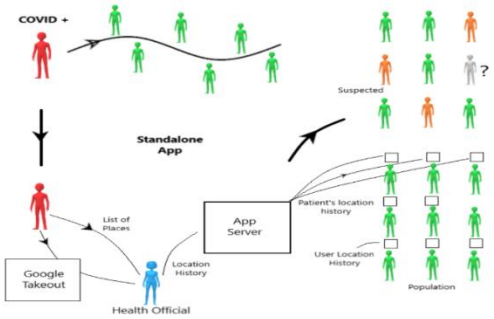


Fig. 7. Cluster Head (Health Official) to Monitor All other Agents within a Cluster.

IV. Experimental Results

The simulations are conducted for 400 (A) susceptible agents and 2 infected individual (B). The 300x300 dimension is utilized for environment simulation. Fig. 8 shows the agents contact results in dissimilar iteration of the simulation process.

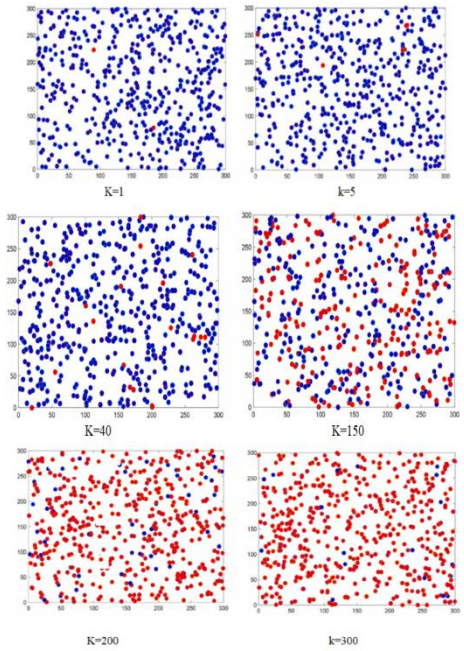


Fig. 8. Results in different Iteration of the Simulation Process. Blue Circle Represents Susceptible Agents and Red Circle Represents Infected Agents [34, 35].

The number of infected agents with iterations is illustrated in Fig. 9.

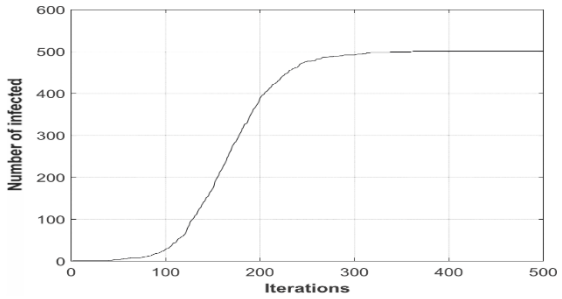


Fig. 9. Progression of the Simulation in Terms of the Number of Infected Agents [36].

V. Conclusion

The agent based weighted clustering approach to evaluate the COVID-19 transmission risk in environment has been presented. The cluster head is elected based upon the weighting parameters. The Aarogya Setu App is utilized for identify the location of COVID-19 positive agents. The behavior of every individual is characterized by set of rules and trust calculation between agents. The simulations are conducted for different iterations to identify the COVID-19 transmissions risk are evaluated. In future this, work will be implemented in global level transmission risk.

References

[1] Yuki Furuse et al, "Clusters of Coronavirus Disease in Communities, Japan, January–April 2020", Emerging Infectious Diseases, 2020,Vol:26, No:9, pp- 2176-2179.

[2] Christopher Wolfram, "An Agent-Based Model of COVID-19", Complex Systems, https://doi.org/10.25088/ComplexSystems.29.1.87.

[3] Sheryl L. Chang et al, "Modelling transmission and control of the COVID-19 pandemic in Australia", arXiv:2003.10218v3 [q-bio.PE] 3 May 2020.

[4] Petronio C. L. Silva et al, "COVID-ABS: An Agent-Based Model of COVID-19 Epidemic to Simulate Health and Economic Effects of Social Distancing Interventions", arXiv:2006.10532v2 [cs.AI] 8 Jul 2020.

[5] Jonatan Gomezet al, "INFEKTA: A General Agent-based Model for Transmission of Infectious Diseases: Studying the COVID-19 Propagation in Bogot´a - Colombia", https://doi.org/10.1101/2020. 04.06.20056119, 2020.

[6] Jouni T. Tuomisto et al, "An agent-based epidemic model REINA for COVID-19 to identify destructive policies", https://doi.org/10.1101/2020.04.09.20047498, 2020.

[7] R. Ravinder et al, "An Adaptive, Interacting, Cluster-Based Model Accurately Predicts the Transmission Dynamics of COVID-19",https://doi.org/10.1101/2020.04.21.20074211, 2020.

[8] Ali Najmi et al, "Determination of COVID-19 parameters for an agent-based model: Easing or tightening control strategies", https://doi.org/10.1101/2020.06.20.20135186, 2020.

[9] Shun Cao et al, "An Agent-Based Model of Leader Emergence and Leadership Perception within a Collective", Complexity, https://doi.org/10.1155/2020/6857891,2020.

[10] Renaud Di Francesco, "Agent Based Model for Covid 19 Transmission: -field approach based on context of interaction", Computers in Biology and Medicine, 2020.

[11] Elizabeth Hunte et al,"A Comparison of Agent-Based Models and Equation Based Models for InfectiousDisease Epidemiology",DOI:10. 21427/rtq2-hs52.

[12] Erin N. Bodine et al, "Agent-BasedModelingandSimulationin Mathematicsand BiologyEducation", Mathematical Biology, 2020, https://doi.org/10.1007/s11538-020-00778-z.

[13] Cliff C. Kerr et al, "Covasim: an agent-based model of COVID-19 dynamics and interventions", doi: https://doi.org/10.1101/2020 .05.10.20097469.

[14] Christine S. M. Currie et al, "How simulation modeling can help reduce the impact of COVID-19", Journal of Simulation, 2020, Vol:14, Issue:2, pp- 83-97.

[15] Pengcheng Zhao et al, "A Comparison of Infection Venues of COVID-19 Case Clusters in Northeast China", International Journal of Environmental Research and Public Health, 2020, doi:10.3390/ijerph17113955.

[16] Elizabeth Hunter et al, "An open-data-driven agent-based model to simulate infectious disease outbreaks", PLOS ONE, 2020, https://doi.org/10.1371/journal. pone.0208775.

[17] Anass Bouchnita and Aissam Jebrane, "A Multi-Scale Model Quantifies The Impact of Limited Movement of the Population and Mandatory Wearing of Face Masks in Containing the COVID-19Epidemic In Morocco", Mathematical Modeling of Natural Phenomena, 2020, https://doi.org/10.1051/mmnp/2020016.

[18] Sarah Ee Fang Yong et al, "Connecting clusters of COVID-19: an epidemiological and serological investigation", The Lancet Infectious Diseases, Vol:20, Issue:7, PP: 809-815, 2020.

[19] Yang Liu et al, "What are the underlying transmission patterns of COVID-19 Outbreak? An age-specific social contact Characterization", EClinicalMedicine,https://doi.org/10.1016/j.eclinm.2020.100354.

[20] Tian-Mu Chen et al, "A mathematical model for simulating the phase-based transmissibility of a novel coronavirus", Infectious Diseases of Poverty, 2020, https://doi.org/10.1186/s40249-020-00640-3.

[21] Erik Cuevas, "An agent-based model to evaluate the COVID-19 transmission risks in facilities", Computers in Biology and Medicine, 2020, DOI: https://doi.org/10.1016/j.compbiomed.2020.103827.

[22] Sundar Prakash Balaji Muthusamy et al., "Synergic deep learning based preoperative metric prediction and patient oriented payment model for total hip arthroplasty", Journal of Ambient Intelligence and Humanized Computing, 2020, https://doi.org/10.1007/s12652-020-02266-7.

[23] AzathMubarakali, M. Ashwin, Dinesh Mavaluru, A. Dinesh Kumar, "Design an attribute based health record protection algorithm for healthcare services in cloud environment", 2020, DOI: 10.1007/s11042-019-7494-7.

[24] "World Health Organization Q&A on coronaviruses(COVID-19)", https://www.who.int/news-room/q-a-detail/q-a-coronaviruses, Accessed in April, 2020.

[25] "Modes of transmission of virus causing COVID-19: implications for IPC precaution recommendations",https://www.who.int/news-room/commentaries/detail/modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution-recommendations, Accessed in March, 2020.

[26] Chaitanya, G. Krishna, et al. "A Survey on Twitter Sentimental Analysis with Machine Learning Techniques." International Journal of Engineering & Technology 7.2.32 (2018): 462-465.

[27] Gogineni Krishna Chaitanya and Krovi.Raja Sekhar, "A Human Gait Recognition Against Information Theft in Smartphone using Residual Convolutional Neural Network" International Journal of Advanced Computer Science and Applications(IJACSA), 11(5), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0110544

[28] Bezawada, A., Marella, S.T. and Gunasekhar, T., 2018. A Systematic Analysis of Load Balancing in Cloud Computing. International Journal of Simulation--Systems, Science & Technology, 19(6).

[29] N. S. Dey and T. Gunasekhar, "A Comprehensive Survey of Load Balancing Strategies Using Hadoop Queue Scheduling and Virtual Machine Migration," in IEEE Access, vol. 7, pp. 92259-92284, 2019, doi: 10.1109/ACCESS.2019.2927076.

[30] Suresh, Ganzi, et al. "Processing & Characterization of LENSTM Deposited Co-Cr-W Alloy for Bio-Medical Applications." International Journal of Pharmaceutical Research (IJPR) Volume 10.1 (2018).

[31] Madhav, Boddapati TP, Yalavarthi Usha Devi, and Tirunagari Anilkumar. "Defected ground structured compact MIMO antenna with low mutual coupling for automotive communications." Microwave and Optical Technology Letters 61.3 (2019): 794-800.

[32] Shekar, S. Chandra, et al. "Wavelet Based Protection Scheme On Renewable Energy Integrated Multi-Terminal Transmission System." International Journal of Pure and Applied Mathematics 120.6 (2018): 721-736.

[33] Manikanta, B. Tejo, V. Ranga Rao, and M. Achyutha Kumar Reddy. "Performance of wood ash blended reinforced concrete beams under acid (HCl), base (NaOH) and salt (NaCl) curing conditions." International Journal of Engineering & Technology 7.3 (2018): 1045-1048.

[34] Ravuvari, A. K., Yechuri, S., Chaitanya, C., & Rajesh, C. (2018). Improved light efficiency in Si solar cells by coating mesoporous TiO2 and cu-modified mesoporous TiO2. Solar RRL, 2(12), 1800214.

[35] Gogineni Krishna Chaitanya and Krovi Raja Sekhar, "GAIT based Behavioral Authentication using Hybrid Swarm based Feed Forward Neural Network" International Journal of Advanced Computer Science and Applications(IJACSA), 11(9), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0110939

[36] Gogineni Krishna Chaitanya and Krovi.Raja Sekhar(in press), "Knowledge-Based Gait Behavioural Authentication Through A Machine Learning Approach" International Journal of Biomedical Engineering and Technology(in press).

AUTHORS' PROFILE

Dr P.Vidya Sagar is an Indian academician who is serving as an Associate Professor in the Department of Computer Science& Engineering in KL University Vijayawada, Andhra Pradesh, India. He got the Ph.D.(Computer Science & Technology) from Sri Krishnadevaya University, Andhra Pradesh, India, in 2016. M.Tech. (Computer Science & Engineering) from Acharya Nagarjuna University, Andhra Pradesh, India,2010. The major domain/specialization of doctorate is Software Engineering application with Deep Learning; Image processing, Data Mining and Networking. I had around 10 yrs of IT industrial experience with major MNC's & currently acting as reviewer/editorial member if international journals and organize member for international conferences.

Dr Pavan Kumar is an Indian academician who is serving as Professor in the Department of Computer Science& Engineering in KL University Vijayawada, Andhra Pradesh, India. He got the Ph.D. (Computer Science &Engineering) from Acharya Nagarjuna University, Andhra Pradesh, India, in 2016. The major domain/specialization of doctorate is in Computer Networks. He is currently acting as reviewer/editorial member if international journals and organize member for international conferences.

Gogineni Krishna Chaitanya received his bachelor's degree in computer science from Acharya Nagarjuna University and master's degree from JNTUK. He is currently pursuing Ph.D degree with Department of Computer Science and Engineering Koneru Lakshmaiah Education Foundation, Vaddeswaram, 522502 Andhra Pradesh, India. His research interests include digital forensics, Biometrics, Authentication and Machine Learning.

Dr. Moparthi Nageswara Rao (born on 15th February 1974) he is an Indian academician who is serving as Professor KL University Vijayawada, Andhra Pradesh, India. I have over all 19.7 years' experience out of Teaching cum Research is 7. 5 years of experience along with 12.2 years **IT industry** from major MNC's like IBM, Sony, Mphasis an HP Company, Birla soft India with a onsite(USA) of 3 years including. I got the Doctoral - Ph.D. in computer science and Technology from Sri Krishandevaraya. The major domain/specialization of doctorate is Software Engineering application data mining. I had 2 patents was published (IPR's) on the same of software Engineering domains and 2 books published Currently I am an associated with different Scopus Int. journal Reviewers like IJAIP, IJDS, CIT& IGI Global publishing (IJORIS) along with 2 SCI journals called IEEE Access and JBD.

Genetic Programming-Based Code Generation for Arduino

Wildor Ferrel[1]
Departamento Académico de Ingeniería Electrónica
Universidad Nacional de San Agustín de Arequipa
Arequipa, Perú

Luis Alfaro[2]
Departamento Académico de Ingeniería de Sistemas
Universidad Nacional de San Agustín de Arequipa
Arequipa, Perú

***Abstract*—This article describes a methodology for writing the program for the Arduino board using an automatic generator of assembly language routines that works based on a cooperative coevolutionary multi-objective linear genetic programming algorithm. The methodology is described in an illustrative example that consists of the development of the program for a digital thermometer organized on a circuit formed by the Arduino Mega board, a text LCD module, and a temperature sensor. The automatic generation of a routine starts with an input-output table that can be created in a spreadsheet. The following routines have been automatically generated: initialization routine for the text LCD screen, routine for determining the temperature value, routine for converting natural binary code into unpacked two-digit BCD code, routine for displaying a symbol on the LCD screen. The application of this methodology requires basic knowledge of the assembly programming language for writing the main program and some initial configuration routines. With the application of this methodology in the illustrative example, 27% of the program lines were written manually, while the remaining 73% were generated automatically. The program, produced with the application of this methodology, preserves the advantage of assembly language programs of generating machine code much smaller than that generated by using the Arduino programming language.**

Keywords*—Genetic programming; Arduino mega board; multi-objective linear genetic programming; cooperative coevolutionary algorithm; automatic generation of programs; Arduino based thermometer

I. Introduction

Arduino is an open-source, free hardware, microcontroller-based electronic board that has a series of analog and digital pins that can be used to connect sensors, peripheral devices, or actuators [1]. The program, which the user stores in the Arduino memory, allows this board to perform various functions such as controller functions, measurement instrument functions, communications equipment functions, etc.

Due to their relatively easy-to-use hardware and software, Arduino boards, in addition to being applied to everyday tasks, are also being applied in scientific instruments such as the measurement of transendothelial/epithelial resistance [2], in the analysis of the production volume of breast milk [3], in the measurement of the methane content of biogas samples [4], etc., which means that there is significant interest from many users and researchers in the use of this free software platform.

There are various Arduino boards such as Arduino Uno, Arduino Mega, Arduino Nano, Arduino Leonardo, Arduino Micro, etc. Because it is one of the fastest Arduino boards on the market [5] and due to the amount of digital and analog pins it has, in this work the Arduino Mega board is used, which is built based on the ATmega 2560 microcontroller with AVR architecture. The main features of the Arduino Mega 2560 are: it has 54 digital input/output pins, 16 analog inputs, 4 UARTs (serial ports in hardware), a 16 MHz crystal oscillator, a USB connection.

Arduino board programming is done through free software that is now accessible On-Line: Arduino Web Editor [6]. To program this board, knowledge of the Arduino programming language is required, which is similar to C++ [7]. In our research work, a program development methodology is proposed for the Arduino Mega board based on program synthesis. Program synthesis aims to automatically produce a program from a specification called "user intent". There are many ways to represent the specification, it can be a sketch [8], a sequence [9], or a table of input-output examples [10][11]. In our work, the starting point for the automatic generation of a program is an input-output table.

The rest of this paper is organized as follows: Section II discusses the related work. In Section III, the theoretical foundations of the proposed methodology are summarized. In Section IV, the fitness evaluation algorithms are detailed. Section V describes the automatic generation of the illustrative example routines. In Section VI, the experimental work carried out is described. In Section VII, the conclusions and recommendations are presented.

II. Related Work

In the classification of program synthesis techniques, genetic programming is within the group of stochastic search techniques [12]. There are efforts to use genetic programming in general-purpose synthesizers [13] and microcontroller program synthesizers. Genetic programming that evolves programs in an imperative language is called linear genetic programming. There are two types of linear genetic programming [14]: a machine code genetic programming, where each instruction is directly executable by the CPU, and an interpreted linear genetic programming. Due to the large difference in the clock frequency of the computer and the

Arduino Mega microcontroller, we use interpreted linear genetic programming.

Dias and Pacheco [15] proposed to apply linear genetic programming in the automatic synthesis of programs in assembly language for the PIC 18F452 microcontroller, showing as examples, the implementation of optimal time control strategies in two cases: in the cart-centering problem and in the problem of balancing an inverted pendulum in a minimum amount of time. The authors in [16] used linear genetic programming in the automatic synthesis of assembly program for the PIC 18F452 microcontroller for optimized control of a water bath plant. In both research works [15][16], a classical evolutionary algorithm is used, and two simulators for the fitness evaluation of an individual have been organized: a simulator of the microcontroller CPU; and, based on its dynamic equations, a simulator of the plant.

Serruto and Casas [17] have proposed to apply linear genetic programming with multi-objective optimization for the automatic synthesis of programs for the AT89S52 microcontroller of 8051 architecture. The generated programs are: 4x3 matrix keyboard scan program, initialization program of the text LCD screen, and a character display program on the LCD screen. For the first case, the authors have proposed the fitness evaluation based on an exhaustive search of the bits of the result, and for the last two cases, the fitness evaluation is carried out by comparing the timing diagrams produced by the genetic program with the target timing diagrams. Serruto and Casas in [18] improved the program generator by introducing the cooperative coevolutionary algorithm, which allowed generating the 4x4 matrix keyboard scanning program and the character display program with a better hit rate. To apply the cooperative coevolutionary algorithm, a machine code program is considered to be made up of program segments, and each segment corresponds to a species. To calculate the fitness of an individual (program segment) a complete program is formed with the individual and the representatives of the other species.

In the proposed work, a methodology of programming of the Arduino Mega board using cooperative coevolutionary multi-objective linear genetic programming is described. The application of the methodology is shown in an illustrative example of developing the program for a digital thermometer circuit based on the Arduino Mega board.

III. Theoretical Fundamentals

A. Circuit with the Arduino Mega Board

The program that will be developed using genetic programming will run on the Arduino Mega board, which will be part of a circuit that will also include other devices connected to the digital or analog pins of the board. This circuit is a system based on the ATmega 2560 microcontroller. In Fig. 1 we show an example of a circuit made up of the Arduino Mega board, a text LCD module connected to digital pins, and a temperature sensor connected to an analog pin. The details of the connection are given in Table I. The program that will be elaborated, following the proposed methodology, as an illustrative example, will allow the circuit of Fig. 1 to function as a thermometer. The concepts on which the proposed methodology is based are inductive programming, linear genetic programming, multi-objective optimization, and cooperative coevolution. Next, these concepts are formulated adapting them to the problem of the synthesis of programs for a microcontroller.

B. Inductive Programming

In the inductive programming method, the starting point is an input/output table [11], from which, the inductive programming technique allows generating a program that makes each input correspond to the output given in the table, and also extrapolates values for other inputs. An input-output table is used in some functions in everyday spreadsheets (Flash-Fill in Microsoft Excel), in which a string processing program is automatically generated, from one or more examples provided by the user [10]. In [19] a formulation of the problem for programming-by-example is shown: Given a set of *M* input-output examples (desired input-output table):

$$(\mathrm{E}_0, S_0), (\mathrm{E}_1, S_1), \ldots, (\mathrm{E}_{M-1}, S_{M-1})$$

a *P* program must be found that performs all the transformations correctly:

$$P(\mathrm{E}_0) \rightarrow S_0;\ P(\mathrm{E}_1) \rightarrow S_1; \ldots;\ P(\mathrm{E}_{M-1}) \rightarrow S_{M-1}$$

To find the *P* program, in the proposed work a multi-objective cooperative coevolutionary linear genetic programming algorithm is used.

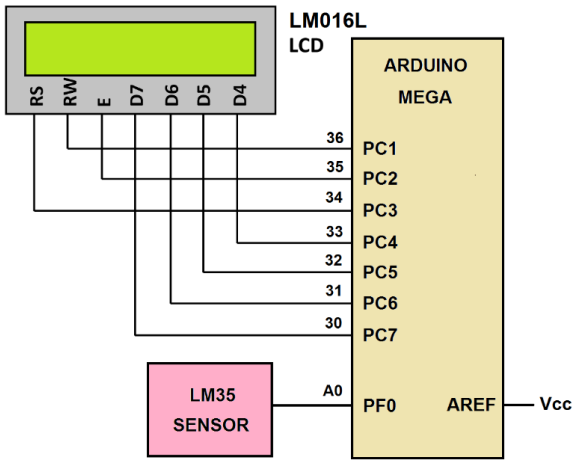


Fig. 1. Circuit Example based on the Arduino Mega.

TABLE I. Connections in the Circuit Example based on the Arduino Mega

Device	Device Line	Arduino Mega Pin	ATmega 2560 Microcontroller Pin
LM35 sensor	Output	A0	PF0/ADC0
LM016L LCD Module	RS	34	PC3/A11
	E	35	PC2/A10
	RW	36	PC1/A9
	D4	33	PC4/A12
	D5	32	PC5/A13
	D6	31	PC6/A14
	D7	30	PC7/A15

C. Linear Genetic Programming

Linear genetic programming (LGP) is the branch of genetic programming that evolves sequences of instructions from an imperative programming language or machine language [20]. In the automatic generation of routines for Arduino Mega, a subset of the ATmega 2560 microcontroller instruction set is used [21][22]. For the generation of programs, which do not interact with the input/output ports, the instructions in Table II are used. When generating programs that produce timing diagrams on Port C, in addition to the instructions in Table II, the instructions in Table III must be used. If the timing diagrams are generated on another port, then in the instructions in Table III, the operand "PORT C" must be changed to the corresponding Port.

TABLE II. INSTRUCTIONS USED IN THE SYNTHESIS OF PROGRAMS

Instruction	Instruction	Instruction	Instruction
NOP	AND R0,R20	INC R20	ASR R1
ADD R0,R1	AND R1,R0	DEC R0	ASR R20
ADD R0,R20	AND R1,R20	DEC R1	SWAP R0
ADD R1,R0	AND R20,R0	DEC R20	SWAP R1
ADD R1,R20	AND R20,R1	TST R0	SWAP R20
ADD R20,R0	ANDI R20,K	TST R1	BST R0,b
ADD R20,R1	OR R0,R1	TST R20	BST R1,b
ADC R0,R1	OR R0,R20	CLR R0	BST R20,b
ADC R0,R20	OR R1,R0	CLR R1	BLD R0,b
ADC R1,R0	OR R1,R20	CLR R20	BLD R1,b
ADC R1,R20	OR R20,R0	SER R20	BLD R20,b
ADC R20,R0	OR R20,R1	MUL R0,R1	SEC
ADC R20,R1	ORI R20,K	MUL R1,R20	CLC
SUB R0,R1	EOR R0,R1	MUL R0,R20	SEN
SUB R0,R20	EOR R0,R20	MULS R20,R20	CLN
SUB R1,R0	EOR R1,R0	LSL R0	SET
SUB R1,R20	EOR R1,R20	LSL R1	CLT
SUB R20,R0	EOR R20,R0	LSL R20	SEH
SUB R20,R1	EOR R20,R1	LSR R0	CLH
SUBI R20,K	COM R0	LSR R1	MOV R0,R1
SBC R0,R1	COM R1	LSR R20	MOV R0,R20
SBC R0,R20	COM R20	ROL R0	MOV R1,R0
SBC R1,R0	NEG R0	ROL R1	MOV R1,R20
SBC R1,R20	NEG R1	ROL R20	MOV R20,R0
SBC R20,R0	NEG R20	ROR R0	MOV R20,R1
SBC R20,R1	CBR R20,K	ROR R1	LDI R20,K
SBCI R20,K	INC R0	ROR R20	
AND R0,R1	INC R1	ASR R0	

TABLE III. ADDITIONAL INSTRUCTIONS USED IN THE SYNTHESIS OF PROGRAMS

Instruction
OUT PORTC,R0
OUT PORTC,R1
OUT PORTC,R20
SBI PORTC,b
CBI PORTC,b

D. ATmega 2560 Microcontroller used Registers

In the microcontrollers of the AVR architecture [23], there are 32 general-purpose registers with names R0, R1, R2,…, R31. These registers are mapped to memory occupying the lowest 32 addresses. The working registers used in the evolutionary process, in genetic programs, are R0, R1, and R20. Register R20 has been included to use the instructions with immediate addressing ANDI, ORI, SUBI, LDI that only work with registers R16 to R31. In code conversion problems, at the completion stage of the program, K registers are used from register R5 to register R(K + 4), where K is the number of bits in the result. In the generation of programs that interact with an input-output Port, the PORT, DDR, and PIN registers of this Port are used. All genetic programs use the SREG register that stores the status of the program.

E. Multi-Objective Evolutionary Optimization

In this work, to find the program P, multi-objective optimization is used. For this purpose, we form K objective functions. In the generation of code conversion routines, each objective function corresponds to one bit of the objective code. In the generation of routines that produce timing diagrams in an input/output Port, each objective function corresponds to a pin of the Port. Next, we formulate the multi-objective optimization problem, adapted to the synthesis problem of a microcontroller program, based on the formulation given in [24]:

Maximize

$$f(P) = (f^0(P), f^1(P), \dots, f^{K-1}(P))$$

Subject to condition

$$P \in U$$

where $P = [I_0, I_1, \dots, I_{N-1}]$ is a genetic program in machine language, I_i is an instruction from Table II or Table III, U is the feasible set, $f(P) = (f^0(P), f^1(P), \dots, f^{K-1}(P))$ is the vector of objective functions.

The target vector $f(P)$ indicates the degree of similarity of the input-output table, generated after running the genetic program P, and the desired input-output table. The multi-objective genetic programming algorithm aims that the generated table is equal to the desired one. When this occurs, the objective functions have the highest value. Thus, in the microcontroller program synthesis problem, studied in this work, multi-objective optimization seeks to maximize the vector of objective functions.

F. Multi-Objective Cooperative Coevolutionary Linear Genetic Programming Algorithm (MOCCLGPA)

The automatic generation of a routine for Arduino is a complex problem, so in addition to multi-objective optimization, in the proposed work, a cooperative coevolutionary algorithm is used. Cooperative coevolutionary algorithms are based on decomposing the problem into subcomponents also called "species" that evolve in collaboration with each other [25].

To apply cooperative coevolution in program synthesis, a machine code program is considered to be made up of program segments, and each segment corresponds to a species. The calculation of the fitness of an individual of a species is carried out by previously forming a complete solution (genetic program) combining the individual (program segment) with the selected representatives of the other species as detailed in [18]. In the proposed work, the number of species is 10 and the representatives are the two best individuals of each species. The Multi-Objective Cooperative Coevolutionary Linear Genetic Programming Algorithm (MOCCLGPA) used in the proposed work is Algorithm 1 taken and modified from [18].

In Algorithm 1, each time the fitness of a genetic program is evaluated, the number of evaluations *n* is increased, the values of *f, fsum, fop,* and *fopsum* are determined, and the *Pbest* program is updated if a better one was found. The sorting of a list is done according to the value of the scalar *fopsum*. The insertion of an individual in a list is carried out using the vector *fop* and the concept of Pareto dominance. The algorithms for the selection of representatives, selection of parents, and variation are described in [18]. At the end of Algorithm 1, the synthesized *Pbest* program, after the completion operation, becomes the generated program.

Algorithm 1. Multi-objective cooperative coevolutionary linear genetic programming algorithm

S_t is a species (program segment) $(t = 0, \dots, T-1)$
Each species has two non-Pareto-dominated fronts P_{1t} and P_{2t} and two temporary lists Q_t and D_t
Pbest is the best program found so far
Nlimit is the limit number of evaluations

1. **for** $t = 0$ **to** $T - 1$ **do**
2. Random generation of P_{1t} and P_{2t}
3. Selection of representatives of S_t
4. **end for**
5. **for** $t = 0$ **to** $T - 1$ **do**
6. Fitness evaluation of each individual of P_{1t} and P_{2t}
7. Sorting of P_{1t}
8. Sorting of P_{2t}
9. **end for**
10. **while** *n* < *Nlimit* **do**
11. **for** $t = 0$ **to** $T - 1$ **do**
12. Selection of representatives of S_t
13. $Q_t \leftarrow$ Parent selection (P_{1t}, P_{2t})
14. $Q_t \leftarrow$ Variation (Q_t)
15. **end for**
16. **for** $t = 0$ **to** $T - 1$ **do**
17. Fitness evaluation of each individual of P_{1t}, P_{2t} and Q_t
18. **end for**
19. **for** $t = 0$ **to** $T - 1$ **do**
20. Insertion of the best from Q_t in P_{1t} and those discarded in D_t
21. Insertion of the best from D_t in P_{2t}
22. Sorting of P_{1t}
23. Sorting of P_{2t}
24. **end for**
25. **end while**

G. Structure of the Automatic Routine Generator

The automatic routine generator consists of an evolutionary program synthesizer and a program completion block. In turn, the synthesizer is made up of the program that implements the MOCCLGPA algorithm and a simulator of the ATmega 2560 microcontroller CPU. The fitness evaluation, in the MOCCLGPA algorithm, and the completion of the program, is carried out according to the type of routine that is generated.

IV. Fitness Evaluation

A. Fitness Evaluation in the Conversion of One Code to Another

Authors in [17] proposed to evaluate the fitness through an exhaustive search of each bit of the binary representation of the output. In the algorithm description, the output value corresponding to the E_j input, expressed in binary, is represented as $\left(S_j^{K-1}, \dots S_j^t, \dots S_j^0\right)$. In this way, each output bit S^t is a combinational function.

The algorithm uses the *RVM* register value matrix that contains the value of the three working registers R0, R1, and R20, at the end of each I_i instruction of the *P* program, and for each value of the E_j input, therefore, the matrix has three dimensions. Each element of the *RVM* matrix is a binary value that we represent RVM_{ij}^b where the index *i* corresponds to the instruction number in the genetic program, *j* corresponds to the input E_j in the input-output table, and *b* is the number of bit in the range 0 to 23 (bit numbers 0 to 7 correspond to register R0, 8 to 15 to R1, and 16 to 23 to R20). Algorithm 2 describes the fitness evaluation of a genetic program in the generation of a code conversion program.

For each combinational function S^t corresponding to an output bit, with the following formula, the most similar combinational function is found in *RVM* matrix:

$$f = (f^0, f^1, \dots, f^{K-1}) \; \forall \; t = 0, \dots, K-1$$
$$f^t = \max_{\substack{0 \le b \le 23 \\ 0 \le i \le N-1}} \sum_{j=0}^{M-1} \left\{ RVM_{ij}^b \odot S_j^t \right\} \quad (1)$$

where the operator ⊙ represents the nor-exclusive operation. The result of this operation, which can be "0" or "1", is then considered an integer value that participates in the arithmetic sum represented by the summation symbol.

Algorithm 2. Fitness evaluation of a genetic program of code conversion

IOT is the input-output table with the outputs in binary representation:

$$\left(\mathrm{E}_0, (S_0^{K-1}, \dots S_0^t, \dots S_0^0)\right),$$
$$\dots$$
$$\left(\mathrm{E}_j, (S_j^{K-1}, \dots S_j^t, \dots S_j^0)\right),$$
$$\dots$$
$$\left(\mathrm{E}_{M-1}, (S_{M-1}^{K-1}, \dots S_{M-1}^t, \dots S_{M-1}^0)\right),$$

$fmax = K \cdot M$ is the maximum value of *fsum*.
$P = [I_0, I_1, \dots, I_i, \dots, I_{N-1}]$ is the genetic program to evaluate.
Pbest is the best program found so far of *NEbest* size with fitness *fsumbest*
RVM is the array of register values

1. Clear (*RVM*)
2. **for** each E_j of *IOT* **do**
3. SREG ← 80H
4. R0 ← E_j
5. R1 ← E_j
6. R20 ← E_j
7. **for** *i* = 0 **to** *N*-1 **do**
8. Execute (I_i)
9. RVM_{ij} ← (*R20*) (*R1*) (*R0*)
10. **end for**
11. **end for**
12. Calculate *f*, *fsum*, *BLM*, *NE*, *fop* and *fopsum* using formulas (1), (2), (3), (4), (5) and (6) respectively
13. **if** (*fsum* > *fsumbest*) **or** ((*fsum* = *fmax*) **and** (*NE* < *NEbest*)) **then**
14. *Pbest* ← *P*; *NEbest* ← *NE*; *fsumbest* ← *fsum*
15. **end if**
16. Return *BLM*, *NE*, *fop*, *fopSum*

To find out if the maximum value has been reached, the sum of all the elements of fitness *f* is calculated:

$$fsum = \sum_{t=0}^{K-1} f^t \quad (2)$$

The best bit location for each element of fitness *f* is stored in the bit location matrix (*BLM*):

$$BLM = [(i^0, b^0), \dots, (i^t, b^t), \dots, (i^{K-1}, b^{K-1})] \quad (3)$$

The effective size of the program is:

$$NE = 1 + \max_{0 \le t \le K-1} (i^t) \quad (4)$$

For smaller programs to have better fitness we use the formula:

$$f_{op} = \left(f_{op}^0, f_{op}^1 \dots, f_{op}^{K-1}\right) \forall\, t = 0, \dots, K-1$$

$$f_{op}^t = f^t - \alpha \cdot NE \quad (5)$$

α has been assigned a value of 0.001.

The fitness f_{op} is used in insertion operations of individuals on Pareto fronts. On Pareto fronts the sorting operations are based on the scalar value:

$$fopSum = \sum_{t=0}^{K-1} f_{op}^t \quad (6)$$

B. Fitness Evaluation in the Generation of Timing Diagrams without Input Values

We represent the timing diagrams of the pins of a microcontroller Port as a string of decimal values in which two consecutive values are not equal.

The algorithm for fitness evaluation of a program that generates timing diagrams proposed in [17], adapted to the AVR architecture, is shown in Algorithm 3, where *L* represents the number of values in the timing diagrams and *K* is the number of pins where the timing diagrams are generated. When the generated timing diagrams *G* are updated, a new value is recorded only if it is different from the previous one by at least one bit. Each component of the fitness vector corresponds to the timing diagram of a Port pin. After the execution of the genetic program, the generated timing diagrams (*G*) are compared with the target timing diagrams (*S*) in binary representation with the formula:

$$f = (f^0, f^1, \dots, f^{K-1}) \text{ for all } p = 0, \dots, K-1$$

$$f^p = \sum_{d=0}^{L-1} \left\{ (L-d)\left(G_d^p \odot S_d^p\right) \right\} \quad (7)$$

where (*L* - *d*) is a weight assigned to *d* time. The first bits have greater weight compared to the last ones. This allows that, in the evolutionary process, the correct values are established, with greater probability, starting with the previous times and ending with the later ones, to improve the speed of convergence of the algorithm.

During the execution of the program, a *VNI* vector is formed with the indices of the instructions, whose execution has produced a change in some Port pin:

Algorithm 3. Fitness evaluation of a genetic program for the generation of timing diagrams without input values

$S = (S_0, , \dots S_d, \dots, S_{L-1},)$ are the target timing diagrams.
$G = (G_0, , \dots G_d, \dots, G_{L-1},)$ are the generated timing diagrams.
Each value of S is represented in binary $S_d = [S_d^{K-1} \quad \dots \quad S_d^0]$
Each value of G is represented in binary $G_d = [G_d^{K-1} \quad \dots \quad G_d^0]$
$P = [I_0, I_1, \dots, I_i, \dots, I_{N-1}]$ is the genetic program to be evaluated
$fmax = K \cdot (L+1) \cdot L/2$ is the maximum value of *fsum*.
Pbest is the best program found so far of *NEbest* size with fitness *fsumbest*

1. SREG ← 80H
2. PORT ← 00H
3. DDR ← FFH
4. R0 ← 0
5. R1 ← 0
6. R20 ← 0
7. Clear (*G*)
8. Clear (*VNI*)
9. **for** *i* = 0 **to** *N* - 1 **do**
10. Execute (*Ii*)
11. Update (*G*)
12. Update (*VNI*)
13. **end for**
14. Calculate *f*, *fsum*, *NE*, *fop* y *fopsum* using formulas (7), (2), (9), (5) and (6) respectively
15. **if** (*fsum* > *fsumbest*) **or** ((*fsum* = *fmax*) **and** (*NE* < *NEbest*)) **then**
16. *Pbest* ← *P*; *NEbest* ← *NE*; *fsumbest* ← *fsum*
17. **end if**
18. Return *NE*, *fop*, *fopSum*

$$VNI = [i_0, i_1, \dots, i_d, \dots, i_{L-1}] \quad (8)$$

The effective size of the program is:

$$NE = 1 + \max_{0 \le d \le L-1}(i_d) \quad (9)$$

C. Fitness Evaluation in the Generation of Timing Diagrams According to Input Values

Algorithm 4 describes the fitness evaluation of a program that generates timing diagrams according to input values. In the algorithm, L represents the number of values in each timing diagram, M is the number of input values, and K is the number of pins where the timing diagrams are generated. To compare the generated timing diagrams with those desired for the E_j input, the formula is:

$$f_j = \left(f_j^0, f_j^1, \dots, f_j^{K-1}\right) \ \forall p = 0, \dots, K-1$$

$$f_j^p = \sum_{d=0}^{L-1}\left\{(L-d)\left(G_{j,d}^p \odot S_{j,d}^p\right)\right\} \quad (10)$$

The fitness vector f is equal to the sum of all the fitness vectors f_j. For the fitness vector f the sum is calculated:

$$f = (f^0, f^1, \dots, f^{K-1})$$

$$f_{sum} = \sum_{p=0}^{K-1} f^p \quad (11)$$

Algorithm 4. Fitness evaluation of a genetic program for the generation of timing diagrams according to input values.

In the input-output table (*IOT*) each output is a sequence of values:

$$\left(E_0, (S_{0,0}, \dots S_{0,d}, \dots S_{0,L-1})\right),$$
$$\dots$$
$$\left(E_j, (S_{j,0}, \dots S_{j,d}, \dots S_{j,L-1})\right),$$
$$\dots$$
$$\left(E_{M-1}, (S_{M-1,0}, \dots S_{M-1,d}, \dots S_{M-1,L-1})\right),$$

Each value in the sequence is represented in binary:

$$S_{j,d} = [S_{j,d}^{K-1} \quad \dots \quad S_{j,d}^0]$$

$P = [I_0, I_1, \dots, I_i, \dots, I_{N-1}]$ is the genetic program to be evaluated

$fmax = M \cdot K \cdot (L+1) \cdot \frac{L}{2}$ is the maximum value of *fsum*.

Pbest is the best program found so far of *NEbest* size with fitness *fsumbest*

1. Clear (*f*)
2. **for** each input E_j **do**
3. DDR ← FFH
4. PORT ← 00H
5. SREG ← 80H
6. R0 ← E_j;
7. R1 ← E_j;
8. R20 ← E_j;
9. Clear (*Gj*);
10. Clear (*VNIj*)
11. **for** i = 0 to N-1 **do**
12. Execute (I_i)
13. Update (*Gj*)
14. Update (*VNIj*)
15. **end for**
16. f_j is calculated with the formula (10)
17. $f \leftarrow f + f_j$
18. **end for**
19. Calculate *fsum, NE, fop, fopsum* using formulas (11), (12), (13) and (14) respectively
20. **if** (*fsum* > *fsumbest*) **or** ((*fsum* = *fmax*) **and** (*NE* < *NEbest*)) **then**
21. *Pbest* ← *P*; *NEbest* ← *NE*; *fsumbest* ← *fsum*
22. **end if**
23. Return *NE, fop, fopsum*

As in the generation of timing diagrams without input values, for each input E_j there is a vector of indices to instructions:

$$VNI_j = \left[i_{j,0}, i_{j,1}, \dots, i_{j,d}, \dots, i_{j,L-1}\right]$$

based on which the scalar $NUI_j = \max_{0 \le d \le L-1}\left(i_{j,d}\right)$ is calculated which is the index of the instruction that produced the last change in the timing diagrams for E_j. Therefore, the effective size of the program is:

$$NE = 1 + \max_{0 \le j \le M-1}\left(NUI_j\right)$$

To prevent that, for different E_j inputs, the vectors VNI_j are different, and the program sizes are also different, based on all the vectors VNI_j, the *VDIF* vector of size *L* is formed, in which each element of position *d* is equal to the difference between the maximum value and the minimum value of all the values of that position in the vectors VNI_j:

$$VDIF_d = \max_{0 \le j \le M-1}\left(i_{j,d}\right) - \min_{0 \le j \le M-1}\left(i_{j,d}\right) \quad (12)$$

Then it is calculated:

$$DIFmax = \max_{0 \le d \le L-1}(VDIF_d),$$

When the evolutionary process ends, $DIFmax$ must be equal to 0. Using NE and $DIFmax$, the fitness vector f_{op} and its sum f_{opsum} are calculated with the following formulas:

$$f_{op} = \left(f_{op}^0, f_{op}^1 \dots, f_{op}^{K-1}\right)$$

$$f_{op}^p = f^p - \alpha \cdot (DIFmax + NE) \quad (13)$$

$$f_{opsum} = \sum_{p=0}^{K-1} f_{op}^p \quad (14)$$

V. Automatic Routine Generation for Arduino MEGA

A. Generation of the Routine for Determining the Temperature Value (ADC_BIN)

As can be seen in the circuit in Fig. 1, the LM35 temperature sensor is connected to the A0 analog input of the Arduino Mega, which in the ATmega 2560 microcontroller is an input to the analog-digital converter (ADC). The ADC converts the voltage provided by the sensor into an integer that we represent as *ADCvalue*.

If the reference voltage on the ADC is Vcc = 5V, and the resolution is 10 bits; then the ADC converts the voltage range from 0V to 5V, proportionally, in the integer range from 0 to 1023. When the LM35 temperature sensor is configured to produce 10mV per °C, to convert the output value of the ADC in temperature value, in degrees centigrade, the following formula is used:

$$Temperature(°C) = \frac{500 \cdot ADCvalue}{1024} \quad (15)$$

The LM35 sensor, as configured, allows a temperature measurement range of 0° C to 150°C. To simplify the input-output table, we set the temperature range from 0°C to 99°C. In a spreadsheet, formula (15) is calculated for all the ADC

output values in the range from 0 to 203, obtaining the column "Temperature (°C)" of Table IV. Then, these values are rounded to the nearest one, obtaining the column "Rounded Temperature". The input-output table for the routine generator is made up of the columns "ADC Value" (Input) and "Rounded Temperature" (Output) of Table IV.

The evolutionary process follows Algorithm 1 with the fitness evaluation described in Algorithm 2. In the completion stage, to the synthesized program *Pbest*, instructions are inserted and added with the information from the bit location matrix (*BLM*). For each (i^t, b^t) pair of *BLM*, after the instruction with index i^t, a MOV instruction is inserted to copy the register containing the bit b^t into a temporary register. For each bit, a different register is used from register R5 to register R(K + 4). At the end of the synthesized program *Pbest*, clear instruction of the R0 register is concatenated followed by pairs of BLD and BST instructions that copy the bits stored in the temporary registers into the R0 register. The completion of the program ends by inserting at the beginning, the instructions that place the initial values in the registers as indicated in Algorithm 2 on lines 3-6.

Before the execution of the ADC_BIN routine, the output value of the ADC must be placed in the R0 register. After executing the ADC_BIN routine, the temperature value rounded to the nearest one is obtained in the same R0 register as a number in natural binary code in the range from 0 to 99.

B. Generation of the Conversion Routine from a Natural Binary Number to Unpacked 2-Digit BCD (BIN_BCD2)

The routine converts a natural binary number in the range 0 to 99 to unpacked BCD code. The input-output table for the generator of this routine is shown in Table V, where for the numbers from 0 to 99, after separating the tens digit and the units digit, the operation *Tens* · 256 + *Units* is calculated placing the tens digit in the high byte and the units digit in the low byte that corresponds to the representation of the number in unpacked BCD code.

The evolutionary process follows Algorithm 1 with the fitness evaluation described in Algorithm 2. The completion of the program is done similarly to that carried out in the generation of the ADC_BIN routine, with the difference that now the most significant bits are placed in the register R1.

To invoke the BIN_BCD2 routine, the natural binary value is placed in register R0. The result of the conversion is obtained in registers R1 and R0. In register R1 the BCD tens digit is obtained and in register R0, the BCD units digit.

C. Generation of the "Home" Command Routine for the LCD Screen (LCD_HOME)

As can be seen in the circuit of Fig. 1 and Table I, the LCD text display module is connected to digital pins on the Arduino Mega board that correspond to Port C of the ATmega 2560 microcontroller. The "Home" routine performs the command to return the LCD screen to the initial state causing the cursor to return to the first left position of the first row. For the execution of this command, the microcontroller must generate in Port C the timing diagrams corresponding to the command with hexadecimal code 02 of the LCD screen controller.

According to the datasheet of the LCD module, in the timing diagrams, the RS and RW signals have value "0", while the enable signal E, for each nibble of the command, during a time interval has value "1" and in the next interval, the value "0". Therefore, in Table VI, at times 1 and 2 when the lines D7-D4 have a value of 0 (high nibble of the hexadecimal value 02), E = 1 at time 1, and E = 0 at time 2. It is important that at the beginning (time 0) and at the end (time 5) all signals are deactivated with a value of "0". The row "Port C" of Table VI is obtained by expressing the value of the Port C pins in decimal representation. The input-output table is made up of the row "Time" (Input) and the row "Port C" (Output) of Table VI.

The evolutionary process follows Algorithm 1 with the fitness evaluation of a genetic program according to Algorithm 3. When the stop condition is met, the synthesized program is *Pbest* of *NE* size.

The completion of the program consists of inserting at the beginning the instructions that place the initial values in the registers as indicated in Algorithm 3 on lines 1-6.

TABLE IV. OBTAINING THE INPUT-OUTPUT TABLE FOR THE GENERATION OF THE ADC_BIN ROUTINE

ADC value	Temperature (°C)	Rounded Temperature
0	0.0000	0
1	0.4883	0
2	0.9766	1
3	1.4648	1
4	1.9531	2
…		
201	98.1445	98
202	98.6328	99
203	99.1211	99

TABLE V. INPUT-OUTPUT TABLE FOR THE BIN_BCD2 ROUTINE GENERATION

Temperature	Unpacked BCD
0	0
1	1
2	2
…	
97	2311
98	2312
99	2313

TABLE VI. OBTAINING THE INPUT-OUTPUT TABLE FOR THE GENERATION OF THE "HOME" COMMAND ROUTINE

Time		0	1	2	3	4	5
D7-D4	PC7-PC4	0	0	0	2	2	0
RS	PC3	0	0	0	0	0	0
E	PC2	0	1	0	1	0	0
RW	PC1	0	0	0	0	0	0
-	PC0	-	-	-	-	-	-
Port C		0	4	0	36	32	0

In the generation of routines that produce timing diagrams on a Port, such as Port C, to overcome timing problems, after each instruction with the PORTC operand the CALL DELAY instruction is inserted.

D. Generating the LCD Screen Initialization Routine (LCD_INI)

According to the datasheet of the LCD screen module, to configure the module with a 4-bit interface and non-visible cursor, the microcontroller must generate timing diagrams on Port C corresponding to the hexadecimal sequence of commands: 33, 32, 28, 0C.

As in the generation of the "Home" routine, the RS and RW signals must have a value of "0" and the E signal for each nibble of command must have a value of "1" during a time interval, and then it must change to "0". In Table VII, the obtaining of the decimal sequence corresponding to the timing diagrams for the initialization of the LCD screen is shown. The input-output table consists of the row "Time" (Input) and the row "Port C" (Output) of Table VII.

The evolutionary process follows Algorithm 1 with the fitness evaluation according to Algorithm 3. When the stop condition is met, the synthesized program is *Pbest* of *NE* size. The program is completed with the insertion of the instructions that place the initial values in the registers, as indicated in lines 1-6 of Algorithm 3, at the beginning of the program. As in the generation of the LCD_HOME routine, after each instruction with the PORTC operand the CALL DELAY instruction is inserted.

E. Generating the Symbol Writing on the LCD Screen Routine (BCD1_LCD)

This routine allows the display of symbols on the LCD screen such as decimal digits from 0 to 9, space, decimal point, and capital letter C. To be displayed, the symbols must be sent to the LCD screen in ASCII code. As an example, Table VIII shows the obtaining of the timing diagrams in Port C to display the decimal digit "1". Proceeding similarly with all symbols, the timing diagrams are obtained in columns S0 to S5 of Table IX. The input-output table for the routine generator consists of the column "Code" (Input) and the columns from S0 to S5 (Output).

TABLE VII. OBTAINING THE INPUT-OUTPUT TABLE FOR THE GENERATION OF THE LCD_INI ROUTINE

Time		**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
D7-D4	**PC7-PC4**	0	3	3	3	3	3	3	2	2	2	2	8	8	0	0	12	12	0
RS	**PC3**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	**PC2**	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0
RW	**PC1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-	**PC0**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Port C		0	52	48	52	48	52	48	36	32	36	32	132	128	4	0	196	192	0

TABLE VIII. OBTAINING THE TIMING DIAGRAMS TO DISPLAY THE DECIMAL DIGIT "1" ON THE LCD SCREEN

Time		**0**	**1**	**2**	**3**	**4**	**5**
D7-D4	**PC7-PC4**	0	3	3	1	1	0
RS	**PC3**	0	1	1	1	1	0
E	**PC2**	0	1	0	1	0	0
RW	**PC1**	0	0	0	0	0	0
-	**PC0**	-	-	-	-	-	-
Port C		0	60	56	28	24	0

TABLE IX. OBTAINING THE INPUT-OUTPUT TABLE FOR THE BCD1_LCD ROUTINE GENERATION

Symbol	**Code**	**ASCII (Hex)**	**S0**	**S1**	**S2**	**S3**	**S4**	**S5**
0	0	30	0	60	56	12	8	0
1	1	31	0	60	56	28	24	0
2	2	32	0	60	56	44	40	0
3	3	33	0	60	56	60	56	0
4	4	34	0	60	56	76	72	0
5	5	35	0	60	56	92	88	0
6	6	36	0	60	56	108	104	0
7	7	37	0	60	56	124	120	0
8	8	38	0	60	56	140	136	0
9	9	39	0	60	56	156	152	0
Space	10	20	0	44	40	12	8	0
.	11	2E	0	44	40	236	232	0
C	12	43	0	76	72	60	56	0

The evolutionary process is developed according to Algorithm 1 with the fitness evaluation according to Algorithm 4. The completion of the program is carried out with the insertion, at the beginning of the program, of the instructions that perform the operations indicated in rows 3-8 of Algorithm 4. As explained before, after each instruction with the PORTC operand, the CALL DELAY instruction is inserted. To invoke the BCD1_LCD routine, the input value (symbol code) must be in the R0 register.

VI. Experimental Work

In the design of systems based on the Arduino Mega board, when the circuit includes devices such as LCD screen module, matrix keyboard, seven-segment indicators, sensors, and others, the methodology proposed in this article for developing the program consists of executing the following steps: 1) the global task is divided, if possible, into subtasks that are expressed through an input-output table; 2) the input-output tables are made with the help of a spreadsheet; 3) the routines described by input-output tables are automatically generated by the generator implemented based on the algorithms shown in previous sections; 4) the main program and routines that were not described with tables are manually written in assembly language. The main program has two parts: the first run only once and sets the system configuration, the second is a loop that repeats indefinitely. For the implementation of the digital thermometer, the main program, shown in Fig. 2(a), which was manually written, mainly contains register transfer instructions (MOV) and routine call instructions (CALL). The routines whose names are in green have been written manually, while the routines with names in red have been generated automatically. Within the infinite loop, the tasks are sequentially executed: place the LCD screen cursor in the initial state, read the ADC value, convert the ADC value into natural binary temperature value, convert the natural binary value into two-digit unpacked BCD code, display the tens digit, display the units digit, and display the letter C.

When the entire assembly language program is written manually, it is necessary to write routines to perform multiplication and/or division operations in the determination of temperature, and routines that perform the operations for handling the LCD screen. In the proposed methodology, these operations are carried out at the stage of making the input-output tables in a spreadsheet; therefore, the application of the proposed method is less complex compared to the manual writing of the entire program in assembly language.

Fig. 2(b) shows the program written in Arduino programming language for the circuit in Fig. 1, with the modification of connecting the RW signal of the LCD module to the ground. Because it is a high-level programming language, writing the program in this language is less complex than writing according to the proposed methodology. In the program in Fig. 2(b), the setup function configures the system and is executed only once; the loop function repeats indefinitely. In the loop function, the following operations are carried out: placing the LCD screen cursor in the first position of the first row, reading the microcontroller ADC value, calculating the temperature, displaying the temperature, and displaying the letter C.

```
.INCLUDE
"M2560DEF.INC"
LDI R16,HIGH(RAMEND)
OUT SPH,R16
LDI R16,LOW(RAMEND)
OUT SPL,R16
CALL SYS_INI
CALL ADC_INI
CALL LCD_INI
LOOP:
CALL ADC_HOME
CALL ADC_READ
MOV R0,R24
CALL ADC_BIN
CALL BIN_BCD2
MOV R2,R0
MOV R0,R1
CALL BCD1_LCD
MOV R0,R2
CALL BCD1_LCD
LDI R24,0X0C
MOV R0,R24
CALL BCD1_LCD
JMP LOOP
```

(a)

```
int ADCvalue;
float Temper;
#include <LiquidCrystal.h>
LiquidCrystal lcd(34, 35, 33, 32, 31, 30);
void setup() {
 lcd.begin(16, 2);
}
void loop() {
 lcd.setCursor(0, 0);
 ADCvalue = analogRead(0);
 Temper=round((500.0*ADCvalue)/1024);
 lcd.print(Temper,0);
 lcd.print("C");
}
```

(b)

Fig. 2. Program for the Digital Thermometer for the Arduino Mega Board. (a) In Assembly Programming Language following the Proposed Methodology (Only the main Program). (b) In Arduino Programming Language.

As can be seen, the two programs in Fig. 2 have similar structures. The programmer who follows the proposed methodology must have basic knowledge about the assembly programming language and the microcontroller architecture to manually write the main program, system configuration routine, ADC initialization routine, ADC reading routine, and a delay routine. All the other routines, which represent a significant percentage of the entire program (73%), have been automatically generated. The complete program, obtained with the proposed methodology, is shown in Fig. 3, with the routines automatically generated on a light blue background. Compiling this program generates a 614-byte machine code. On the other hand, writing the program in the Arduino programming language requires knowledge of this language, which is similar to the C++ language, and its compilation produces a 3788-byte machine code, which means that the use of the proposed methodology preserves the advantage of assembly language to produce smaller machine code compared to other programming languages.

In the stage of convergence and stability tests of the algorithms used in the proposed methodology, the generator of each routine has been executed 10 times. Table X shows the most important characteristics of the tests carried out in the generation of these routines. As can be seen in Table X, when the input-output table has a large number of rows (ADC_BIN routine), or when the table has several output columns (BCD1_LCD routine) the hit rate may be low (30%); this also occurs when the output has a low correlation or is not correlated with the input. For the generator of the BIN_BCD2 routine, the hit rate is the highest (100%) since the input and the output are quite correlated. The test parameters are: The size of the initial population of each species is 100, the number of species (program segments) is 10, the number of

representatives per species is 2, the initial size of the program segment has a minimum value of 2 and maximum value 4, the coefficient α in the fitness evaluation is 0.001. The programs have been compiled in AtmelStudio software and simulated in Proteus software using the Simulino Mega model. In the simulation or the implementation, if necessary, in the DELAY procedure, the waiting time can be varied by modifying the initial value placed in register R21 or R22.

```
.INCLUDE "M2560DEF.INC"
 LDI R16,HIGH(RAMEND)
 OUT SPH,R16
 LDI R16,LOW(RAMEND)
 OUT SPL,R16
 CALL SYS_INI
 CALL ADC_INI
 CALL LCD_INI
LOOP:
 CALL LCD_HOME
 CALL ADC_READ
 MOV R0,R24
 CALL ADC_BIN
 CALL BIN_BCD2
 MOV R2,R0
 MOV R0,R1
 CALL BCD1_LCD
 MOV R0,R2
 CALL BCD1_LCD
 LDI R24,0X0C
 MOV R0,R24
 CALL BCD1_LCD
 JMP LOOP
SYS_INI:
 IN R20,MCUCR
 ANDI R20,0XEF
 OUT MCUCR,R20
 LDI R20,255
 OUT DDRC,R20
 LDI R20,0X00
 OUT PORTC,R20
 RET
ADC_READ:
 LDS R22,ADCSRA
 ORI R22,(1<<ADSC)
 STS ADCSRA,R22
SALTO:
 LDS R22,ADCSRA
 SBRC R22,ADSC
 RJMP SALTO
 LDS R24,ADCL
 LDS R25,ADCH
 RET
DELAY:
 IN R21,SREG
 PUSH R21
 LDI R21,1;*
SAL1:
 LDI R22,10;*
SAL2:
 LDI R23,255
SAL3:
 DEC R23
 BRNE SAL3
 DEC R22
 BRNE SAL2
 DEC R21
 BRNE SAL1
 POP R21
 OUT SREG,R21
 RET
ADC_INI:
 LDI R22,0X00
 STS ADMUX,R22
 LDS R22,ADCSRB
 ANDI R22,0B11110111
 STS ADCSRB,R22
 LDI R22,0B10000111
 STS ADCSRA,R22
 ANDI R22,0B11011111
 STS ADCSRA,R22
 RET
ADC_BIN:
 LDI R20,0X80
 OUT SREG,R20
 MOV R20,R0
 MOV R1,R0
 LDI R20,250
 MUL R0,R20
 INC R1
 MOV R5,R1
 MOV R6,R1
 MOV R7,R1
 MUL R0,R20
 IN R21,SREG
 MOV R8,R1
 MUL R1,R20
 MOV R9,R1
 MOV R6,R1
 MOV R10,R1
 MOV R7,R1
 MOV R11,R1
 MOV R8,R1
 CLR R0
 CLR R0
 BST R5,1
 BST R5,0
 BLD R0,0
 BLD R0,0
 BST R6,2
 BST R6,1
 BLD R0,1
 BLD R0,1
 BST R7,3
 BST R7,2
 BLD R0,2
 BLD R0,2
 BST R8,4
 BST R8,3
 BLD R0,3
 BLD R0,3
 BST R9,5
 BST R9,7
 BLD R0,4
 BLD R0,4
 BST R10,6
 BST R10,7
 BLD R0,5
 BLD R0,5
 BST R11,7
 BST R11,7
 BLD R0,6
 BLD R0,6
 RET
BIN_BCD2:
 LDI R20,0X80
 OUT SREG,R20
 MOV R20,R0
 MOV R1,R0
 ROR R1
 MOV R5,R0
 MOV R9,R0
 MOV R10,R0
 MOV R11,R0
 MOV R12,R0
 INC R1
 LDI R20,51
 MUL R1,R20
 MUL R1,R20
 CLR R1
 BST R13,0
 BLD R1,0
 BST R14,1
 BLD R1,1
 BST R15,2
 BLD R1,2
 BST R16,3
 BLD R1,3
 RET
LCD_HOME:
 LDI R20,0XFF
 OUT DDRC,R20
 MOV R13,R1
 MOV R14,R1
 MOV R15,R1
 MOV R16,R1
 MUL R0,R20
 MUL R1,R20
 MOV R6,R1
 MOV R7,R1
 MOV R8,R1
 CLR R0
 BST R5,0
 BLD R0,0
 BST R6,1
 BLD R0,1
 BST R7,2
 BLD R0,2
 BST R8,3
 BLD R0,3
 BST R9,7
 BLD R0,4
 BST R10,7
 BLD R0,5
 BST R11,7
 BLD R0,6
 BST R12,7
 BLD R0,7
 CLR R1
 BST R13,0
 BLD R1,0
 BST R14,1
 BLD R1,1
 BST R15,2
 BLD R1,2
 BST R16,3
 BLD R1,3
 LCD_HOME:
 LDI R20,0XFF
 OUT DDRC,R20
 LDI R20,51
 MUL R1,R20
 LDI R20,0X00
 OUT PORTC,R20
 CALL DELAY
 MOV R0,R20
 MOV R1,R20
 LDI R20,0X80
 OUT SREG,R20
 LDI R20,0
 MOV R0,R20
 MOV R1,R20
 SBCI R20,219
 SBI PORTC,2
 CALL DELAY
 CALL DELAY
 CBI PORTC,2
 CALL DELAY
 CBI PORTC,2
 CALL DELAY
 OUT PORTC,R20
 CBI PORTC,2
 CALL DELAY
 CBI PORTC,2
 CALL DELAY
 OUT PORTC,R0
 CALL DELAY
 RET
BCD1_LCD:
 LDI R20,0XFF
 OUT DDRC,R20
 LDI R20,0X00
 OUT PORTC,R20
 CALL DELAY
 LDI R20,0X80
 OUT SREG,R20
 MOV R20,R0
 MOV R1,R0
 LDI R20,0X00
 OUT PORTC,R20
 CALL DELAY
 LDI R20,0X80
 OUT SREG,R20
 LDI R20,0
 MOV R0,R20
 MOV R1,R20
 SUBI R20,208
 SBCI R20,219
 SBI PORTC,2
 CALL DELAY
 CBI PORTC,2
 CALL DELAY
 CBI PORTC,2
 CALL DELAY
 ADD R20,R1
 AND R20,R0
 ADD R1,R20
 OUT PORTC,R1
 CALL DELAY
 CBI PORTC,2
 CALL DELAY
 MULS R20,R20
 LDI R20,133
 OUT PORTC,R20
 CALL DELAY
 LDI R20,196
 CBI PORTC,2
 CALL DELAY
 CBI PORTC,2
 CALL DELAY
 LDI R20,0X80
 CALL DELAY
 CBI PORTC,2
 CALL DELAY
 OUT PORTC,R1
 CALL DELAY
 CALL DELAY
 CALL DELAY
 OUT PORTC,R20
 CALL DELAY
 LDI R20,0X80
 CALL DELAY
 OUT PORTC,R20
 CALL DELAY
 CBI PORTC,2
 CALL DELAY
 OUT PORTC,R20
 CALL DELAY
 LDI R20,0X80
 CALL DELAY
 CBI PORTC,2
 CALL DELAY
 MOV R1,R20
 SUBI R20,203
 CALL DELAY
 OUT PORTC,R20
 RET
 CALL DELAY
 CBI PORTC,2
```

Fig. 3. The Complete Program, Obtained with the Proposed Methodology, for the Example of the Digital Thermometer based on the Arduino Mega Board.

TABLE X. FEATURES AND RESULTS OF THE ROUTINE GENERATION TESTS

Feature/Result	**ADC_BIN**	**BIN_BCD2**	**LCD_HOME**	**LCD_INI**	**BCD1_LCD**
Number of rows in the input-output table	204	100	6	18	13
Number of bits of the input value	8	7	-	-	4
Number of output columns in the input-output table	1	1	1	1	6
Number of output bits or number of output pins	7	12	7	7	7
Hit rate	30%	100%	60%	40%	30%
Minimum number of evaluations	500279	247441	524155	2915945	22868727
Maximum number of evaluations	2390847	1990565	2707987	4030899	39725057
Minimum program size	30	48	22	51	42
Limit number of evaluations (stop condition)	$5x10^6$	$5x10^6$	$5x10^6$	$5x10^6$	$40x10^6$

The limitations of the proposed methodology are: 1) the application of the methodology depends on the existence of subtasks that are described by input-output tables; 2) the size of the input-output tables cannot be too large, tables of up to 204 rows and up to 6 columns of output have been used in the tests; 3) depending on the speed of the computer, in some cases, the generation of the program may take a long time; 4) in the generation of peripheral device management routines, in addition to the simulator of the microcontroller CPU, it is necessary to simulate the interface of the microcontroller with the peripheral device.

VII. Conclusions and Suggestions

In this article, a programming methodology for the Arduino Mega board has been described. This methodology is applicable in cases when, in addition to the Arduino Mega board, the circuit includes devices such as LCD screen, matrix keyboard, seven-segment indicators, sensors, etc. The methodology makes use of an automatic generator of assembly language routines, whose operating algorithms, based on multi-objective cooperative coevolutionary linear genetic programming, are described in this work.

The application of the methodology has been shown in an illustrative example that consists of the development of the program for a digital thermometer organized on a circuit formed by the Arduino Mega board, an alphanumeric LCD module, and a temperature sensor. The result is an assembly language program where 73% of the program lines have been generated automatically; which means that writing the program following the proposed methodology is less complex than manually writing the entire program in assembly language. The example has also shown that the application of the proposed methodology preserves the advantage of assembly language programming of generating machine code of a much smaller size than that generated by the Arduino programming language.

In the illustrative example, the temperature range is 0°C to 99°C displaying only the integer value of the temperature; however, it is also possible to display the values with tenths and hundredths, but with a lower temperature range. As part of future work, the authors intend to follow the methodology for programming other Arduino-based measurement instruments.

In the ATmega 2560 microcontroller instruction set, there are fractional multiplication instructions FMUL, FMULS, and FMULSU, which have not been used in this work. In future work, we recommend the inclusion of these instructions in the table of instructions used by the microcontroller simulator, which could lead to an improvement in the performance of the automatic routine generator, especially when generating routines for calculating mathematical formulas.

References

[1] Arduino, http://www.arduino.cc/, Accessed: 2020-09-29.

[2] Curtis G. Jones, Chengpeng Chen, An arduino-based sensor to measure transendothelial electrical resistance, Sensors and Actuators A: Physical, Volume 314, 2020, 112216, ISSN 0924-4247.

[3] Nur Aliya Arsyad, Syafruddin Syarif, Mardiana Ahmad, Suryani As'ad, Breast milk volume using portable double pump microcontroller Arduino Nano, Enfermería Clínica, Volume 30, Supplement 2, 2020, Pages 555-558, ISSN 1130-8621.

[4] Shunchang Yang, Yikan Liu, Na Wu, Yingxiu Zhang, Spyros Svoronos, Pratap Pullammanappallil, Low-cost, Arduino-based, portable device for measurement of methane composition in biogas, Renewable Energy, Volume 138, 2019, Pages 224-229, ISSN 0960-1481.

[5] Congduc Pham, Communication performances of IEEE 802.15.4 wireless sensor motes for data-intensive applications: A comparison of WaspMote, Arduino MEGA, TelosB, MicaZ and iMote2 for image surveillance, Journal of Network and Computer Applications, Volume 46, 2014, Pages 48-59, ISSN 1084-8045.

[6] Arduino Web Editor, https://www.arduino.cc/en/Main/Software, Accessed: 2020-09-29.

[7] Bob Dukish,Coding the Arduino: Building Fun Programs, Games, and Electronic Projects, Canfield, Ohio, USA, 2018, ISBN-13 (pbk): 978-1-4842-3509-6 ISBN-13 (electronic): 978-1-4842-3510-2.

[8] Alexandre Mota, Juliano Iyoda, Heitor Maranhão, Program synthesis by model finding, Information Processing Letters, Volume 116, Issue 11, 2016, Pages 701-705, ISSN 0020-0190.

[9] De Ridder L., Vercammen T. (2019) Deriving Formulas for Integer Sequences Using Inductive Programming. In: Atzmueller M., Duivesteijn W. (eds) Artificial Intelligence. BNAIC 2018. Communications in Computer and Information Science, vol 1021. Springer, Cham.

[10] Sumit Gulwani, José Hernández-Orallo, Emanuel Kitzelmann, Stephen H. Muggleton, Ute Schmid, and Benjamin Zorn. 2015. Inductive programming meets the real world. Commun. ACM 58, 11 (November 2015), 90–99. DOI:https://doi.org/10.1145/2736282.

[11] Flener P., Schmid U. (2017) Inductive Programming. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA.

[12] Sumit Gulwani, Oleksandr Polozov and Rishabh Singh, "Program Synthesis", Foundations and Trends® in Programming Languages: Vol. 4: No. 1-2, pp 1-119. (2017).

[13] Alexandre Correia, Juliano Iyoda, Alexandre Mota, Combining model finder and genetic programming into a general purpose automatic program synthesizer, Information Processing Letters, Volume 154, 2020, 105866, ISSN 0020-0190, https://doi.org/10.1016/j.ipl.2019.105866.

[14] Grochol D., Sekanina L. (2017) Comparison of Parallel Linear Genetic Programming Implementations. In: Matoušek R. (eds) Recent Advances in Soft Computing. ICSC-MENDEL 2016. Advances in Intelligent Systems and Computing, vol 576. Springer, Cham.

[15] Douglas Mota Dias, Marco Aurélio C. Pacheco, José F. M. Amaral, "Automatic synthesis of microcontroller assembly code through linear genetic programming", In Genetic Systems Programming: Theory and Experiences, Springer Berlin Heidelberg, Berlin, 2006, pp 193 – 227.

[16] Dias D.M., Pacheco M.A.C., Amaral J.F.M. (2006) Genetic Programming of a Microcontrolled Water Bath Plant. In: Gabrys B., Howlett R.J., Jain L.C. (eds) Knowledge-Based Intelligent Information and Engineering Systems. KES 2006. Lecture Notes in Computer Science, vol 4253. Springer, Berlin, Heidelberg.

[17] W. F. Serruto and L. A. Casas, "Automatic Code Generation for Microcontroller-Based System Using Multi-objective Linear Genetic Programming," 2017 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, 2017, pp. 279-285, doi: 10.1109/CSCI.2017.47.

[18] Serruto, Wildor Ferrel and Alfaro, Luis, Many-Objective Cooperative Co-evolutionary Linear Genetic Programming applied to the Automatic Microcontroller Program Generation, International Journal of Advanced Computer Science and Applications, 2019, Volume 10, Number 1, Pages 21-31.

[19] Jacob Devlin, Jonathan Uesato, Surya Bhupatiraju, Rishabh Singh, Abdel-rahman Mohamed, Pushmeet Kohli, RobustFill: Neural Program Learning under Noisy I/O, 2017.

[20] Oliveira, V.P.L., Souza, E.F.d., Goues, C.L. *et al.* Improved representation and genetic operators for linear genetic programming for automated program repair. Empir Software Eng 23**,** 2980–3006 (2018).

[21] Atmel Corporation, ATmega640/V-1280/V-1281/V-2560/V-2561/V, 8-bit Atmel Microcontroller with 16/32/64KB In-System Programmable

Flash datasheet, 1600 Technology Drive, San Jose, CA 95110 USA, 2014.

[22] Atmel Corporation, AVR Instruction Set Manual, 1600 Technology Drive, San Jose, CA 95110 USA, 2016.

[23] Sarmad Naimi, Muhammad Ali Mazidi, Sepehr Naimi, The AVR Microcontroller and Embedded Systems Using Assembly and C: Using Arduino Uno and Atmel Studio, Microdigitaled, 632 pages, 2017.

[24] Nyoman Gunantara, Qingsong Ai (Reviewing editor) (2018) A review of multi-objective optimization: Methods and its applications, Cogent Engineering, 5:1, DOI: 10.1080/23311916.2018.1502242.

[25] Rodriguez-Coayahuitl L., Morales-Reyes A., Escalante H.J., Coello Coello C.A. (2020) Cooperative Co-Evolutionary Genetic Programming for High Dimensional Problems. In: Bäck T. et al. (eds) Parallel Problem Solving from Nature – PPSN XVI. PPSN 2020. Lecture Notes in Computer Science, vol 12270. Springer, Cham.

Drop-Out Prediction in Higher Education Among B40 Students

Nor Samsiah Sani[1], Ahmad Fikri Mohamed Nafuri[2], Zulaiha Ali Othman[3]
Mohd Zakree Ahmad Nazri[4], Khairul Nadiyah Mohamad[5]
Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia[1, 2, 3, 4]
Unit Pemodenan Tadbiran dan Perancangan Pengurusan Malaysia (MAMPU), Cyberjaya, Selangor, Malaysia[5]

***Abstract*—Malaysia citizens are categorized into three different income groups which are the Top 20 Percent (T20), Middle 40 Percent (M40), and Bottom 40 Percent (B40). One of the focus areas in the Eleventh Malaysia Plan (11MP) is to elevate the B40 household group towards the middle-income society. In 2018, it was estimated that 4.1 million households belong to this group. The government of Malaysia has widened access to higher education for the B40 group in an effort to reduce the gaps in socioeconomics and to improve their living standards. Statistical data shows that since 2013, a yearly intake of students in bachelor's degree programs in Malaysia's public universities amounts to more than 85,000. Despite this huge number of enrolments, not all were able to graduate, including students from low-income family background. Data mining approach with machine learning techniques has been widely used effectively and accurately to predict students at risk of dropping out in general education. However, machine learning related works on student attrition in Malaysia's higher education is generally lacking. Therefore, in this research, three machine learning models were developed using Decision Tree, Random Forest and Artificial Neural Network algorithm in order to classify attrition among B40 students in bachelor's degree programs in Malaysia's public universities. Comparative performance analysis between the three models indicates that the Random Forest model is the best model in predicting student attrition in this study. Random Forest model outperforms the other two models in terms of accuracy, precision, recall and F-measure with the value of 95.93%, 97.10%, 81.26% and 88.50%, respectively. Nevertheless, there is a statistically significant difference in performance between the Random Forest model and Decision Tree model but no statistically significant difference between Random Forest models and Artificial Neural Network model.**

Keywords*—Machine learning; prediction; student attrition; student drop-out; B40; random forest; decision tree; artificial neural network

I. INTRODUCTION

Malaysia's household income is classified into three groups, which are Bottom 40% (B40), Middle 40% (M40) and Top 20% (T20). According to the Department of Statistic, Malaysia (2017), generally in Malaysia, B40 household income is not more than RM 4,630.00. Approximately, there were 2.7 million households belonging to B40 group in 2014. The figures increased in 2018, as the government announced that 4.1 million households will continue to benefit from Bantuan Sara Hidup (Household Living Aid) (BSH) which is specially allocated to B40 group [1].

B40 had been selected as a focus group in Rancangan Malaysia 11 2016-2020 (The Eleventh Malaysia Plan) (RMK-11). Through RMK-11, the government used education as one of the strategies to boost B40 household's income and ultimately narrowing socioeconomic gap [1]-[2]. Higher education institution and skills training institutes were encouraged to allocate more seats and allowing admission by special allocation for B40 students in an effort to ensure their access to higher education is secured. As reported in Higher Education Statistic, from the year 2011 to 2015, the total number of students intake for bachelor's degree programmes in Malaysia Public Universities is more than 85,000 students yearly. The highest number of students intake was recorded in the year 2011 with 99,862 students. Nonetheless, not all were able to graduate on time. In a worse case, some dropped out voluntarily or were expelled from by university.

Student's attrition in university will negatively affect B40 students financially. The family financial burden will increase as student's education loan has to be paid even if they fail to graduate. Furthermore, it will affect a student's chances on securing a high-income job. Students drop out would also lead to a huge loss in human capitals to the nation as fewer professionals and expert skills will be produced by public universities.

Hence, a proactive approach is desperately needed in identifying students who are at risk at dropping out. An effective prediction model using machine learning technique can be implemented for that purpose. Thus, the aim of this paper is to conduct a comparative study for machine learning models in predicting attrition among B40 students, particularly in the bachelor's degree programme in Malaysia Public Universities. Decision Tree (DT) Random Forest (RF), and Artificial Neural Networks (ANN) algorithms were adopted in constructing the models.

The remainder of this paper is organized as follows. Section 2 presents previous research articles related to classification technique in education and student drop-out prediction in higher learning institutions. Section 3 describes the methodology used in predicting student's drop out in this research. Results and discussion will be discussed in Section 4, while the conclusion of this paper and further works is outlined in Section 5.

II. Literature Review

Classification is a machine learning technique that can be used to predict students drop out rate accurately to help reducing student attrition rate. The task is crucial as the ability to predict students at risk the earliest possible is a great help to keep students from leaving their studies and overcome attrition among B40 students. Classification technique had been developed and applied successfully to a wide range of real-world domains [3] – [8]. Also, the classification is playing an important role in the education domain, especially in predicting student's academic performance, whether in school or higher education institution [9]. The research had review 30 studies carried out in between year 2002 until early 2015 and discovered that Artificial Neural Network (ANN), Decision Tree (DT), Naïve Bayes (NB), k-Nearest Neighbour (k-NN) and Support Vector Machine (SVM) were often used in building prediction models. However, findings showed that ANN and DT models produced higher accuracy results than the others.

Over recent years, there has been a significant growth of research published in predicting student performance, focusing on course drop-out/ retention using the technique of classification (supervised learning). These researches concentrate on predicting final grade or Cumulative Grade Point Average (CGPA) of students by utilizing classifier algorithm [10]-[13], predicting student's performance in *Massive Online Open Courses* (MOOC) environment [14], and predicting students at risk of not graduating high school on time [15].

In Malaysia, classification techniques had been applied in education domain, but the focus was more on student performance rather than attrition. Reference [16] in their work designed a model to identify key factors that influence the drop-out rates in Computer Science course. They collected student's demographic information and transcript records which focused on the core courses offered as it gives more impact on the drop-out case. Four different classification techniques namely k-NN, DT, NN and Logistic Regression (LR) are utilized to classify the dataset. The results show that LR classifier is the most accurate (91%) as compared to other techniques used in this work. The outcome of this work reveals that there are five important courses that the student must score higher to lower the chance of dropping out.

Bedregal Alpaca et al. [17] proposed classification models based on academic information provided by university to identify a student at risk of drop-out. The student's demographic, academic performance, admission test and course information data are considered for the evaluation. From the result, it is observed that the model is able to determine the most significant variable that affects academic performance, which is the abandoned subjects.

Gil, Delima, and Vilchez [18] adapted DT and NB to identify the underlying factors of student drop-out in a public school in the Philippines. They used Weka tool kit to utilize the classifier algorithm on the selected dataset and produced a comparative result of each algorithm performance in terms of recall, precision and accuracy. Meanwhile, [19] only concentrated on k-NN to perform extensive evaluation and predict student drop-out at an early stage of study. The technique is versatile, simple and can handle different type of data. The results can help teachers to identify a student at risk of drop-out and check on their welfare.

Mardolkar and Kumaran [20] adapted data mining technique to find comprehensive prediction models of student drop-out as early as possible. The model with sufficiently high accuracy will be used in an early warning system as an effort to detect students at high risk of drop-out as soon as possible. They explored the academic variables (both at universities and former school), sociodemography, behaviour and extracurricular activities that may influence student drop-out. However, only a subset of attributes that has a very high predictive contribution on the student drop-out.

Tomasevic, Gvozdenovic and Vranes [21] conducted a research with an objective to provide a comprehensive analysis and comparison of supervised machine learning techniques for discovering students at a high risk of dropping out from the course. For this, they used various classifier such as k-NN, SVM, ANN, DT, NB and LR as the classification tool. The overall highest precision was obtained with ANN by feeding the algorithm with student engagement data in online learning and past performance data.

Viloria and Padilla [22] in their study applied NN, DT and Bayesian Network to predict drop-out among engineering students in India. As a result, it was found that academic results and socioeconomic situation have an influence on students and managing these variables helps reduce the drop-out rate.

Sangodiah et al. [23] used SVM to predict academic performance for students under probation in a private higher learning institution. The model gained 89.84% of accuracy. Likewise, [24] also used single classifier to predict postgraduate doctoral degree students that will complete their study on time by using Binary Linear Regression. The outcome revealed that only 6.8% of the students in the year 2014 were able to graduate on time.

Table I described 16 studies conducted in predicting student drop out in higher learning institution from the year 2015 until 2020. The studies indicated that academic and sociodemographic data were important features used in predicting student. Other than that, there was only one research in predicting student drop-out that uses data from server logs containing student's activities for online courses offered from various universities. All of the research reviewed here were targeting students from only one course/major in one faculty or similar institution. However, in this research, the focus will be shifted to predicting drop out among B40 students by using academic or sociodemographic data from various majors and various higher learning institutions (public universities).

TABLE I. RESEARCH IN STUDENT DROP-OUT PREDICTION IN HIGHER INSTITUTION USING CLASSIFICATION TECHNIQUES

Author, Year	Objective	Data	Algorithm	Result
Bedregal-Alpaca et al. (2020) [17]	To generate a classification model and implement them on academic information provided by the university.	Demographic, academic, admission test, course information	ANN, DT	The generated model is able to determine the most significant variable that affects academic performance, which is the abandoned subjects.
Gil et al. (2020) [18]	To identify the underlying factors of drop-out students and apply the different approach of data mining algorithms.	Academic, student attendance, sociodemographic,	DT, NB	DT model produces the best result. The model identified key factors that affect students drop-outs.
Mardolkar & Kumaran (2020) [19]	Evaluate and propose k-NN method to predict students' drop-out	Student welfare feature and academic performance	k-NN	The technique is versatile, simple and can handle different type of data.
Tomasevic et al. (2020) [21]	To provide a comprehensive analysis and comparison of supervised machine learning techniques applied for discovering students at a high risk of dropping out of the course.	Demographic, student engagement in virtual learning, academic performance	k-NN, SVM, ANN, DT, NB, LR	ANN gave the highest precision by feeding the engagement data and past performance data.
Wan Yaacob et al. (2020) [16]	To identify key factors that influence the drop-out rates in Computer Science Program and which data mining technique is the most suitable approach.	Demographic, academic (CGPA and transcript records)	k-NN, DT, NN, LR,	LR classifier was the best technique with 91% accuracy. The models identified key factors/courses that have a greater impact on drop-out.
Viloria et al. (2019) [22]	Student drop-out prediction using data mining techniques.	Academic (university), academic (school)	NN, DT, Bayesian Network	All predictive models produced similar results, but Bayesian Network has slightly higher precision.
Limsathitwong, Tieatthanont & Yatsungnoen (2018) [25]	To develop web-based system with the ability to predict students who are at risk to drop-out in Information Technology major.	Academic (First and second year students)	DT, RF	RF accuracy higher than DT
Chen, Johri & Rangwala (2018) [26]	Performance comparison between survival analysis framework and machine learning approach in predicting student attirition in Science, Technology, Engineering and Mathematic (STEM) major.	Academic and sosiodemografic	*survival analysis*, Linear Regression (LR), DT, NB, RF, Adaboosting	Survival analysis outperforms other classifiers. Important features that influence student attrition was a student's age when enrolled in university and CGPA.
Ortiz-Lozano et al. (2018) [27]	Student drop-out prediction from school of engineering in one of the universities in Spain..	Academic and sosiodemografic	DT – CART, QUEST	Model accuracy is 70%. Academic results were a good feature in predicting student drop-out.
Aulck et al. (2016) [28]	Student drop-out prediction in one of thepublic universities in United States of America.	Academic and sosiodemografic	LR, RF, k-NN	The prediction accuracy of LR model was higher than the other two models.
Liang et al. (2016) [29]	Predicting student who is at risk of leaving a course in ten days on Edx MOOC platform	Students activities record in server logs.	SVM, LR, RF, DT with *gradian boosting*	DT with the gradian boosting prediction model outperformed the others in accuracy.
Pokrajac et al. (2016) [30]	To develop student drop-out prediction model at Delaware State University.	Academic and high school data.	ANN	The accuracy of the model increased with CGPA and the number of credit hours were included in developing the model.
Márquez-Vera et al. (2016) [20]	Propose an algorithm to obtain a reliable and comprehensible classification with sufficiently high accuracy.	Academic (university), academic (school), sociodemography, behaviour and extracurricular activities	NB, SVM, k-NN, DT	Focus on early detection to be used in an Early Warning System. Classification performance is near to 100%.
S. Abu-Oda & M. El-Halees (2015) [31]	Predicting attrition for science computer students in University Al-Aqsa.	Academic , sosiodemografic and student admission.	DT, NB	DT performed better than NB with an accuracy of 98.14%.
Strecht et al. (2015) [32]	Comparing prediction models' performance for student drop-out at Porto University.	Sosiodemografic, student admission, financial aid	k-NN, RF, Adaboost, CART, SVM, NB	SVM outperformed other models based on F-measure score, but the differences were not significant.
Siri (2015) [33]	Predicting student drop-out for bachelor degree programme (healthcare) in University of Genoa, Itali.	Academic , sosiodemografic and phone conversation	ANN	Able to predict student drop-out with 76% accuracy.

Comparative studies of two or more prediction models had been the core for 13 studies while the remaining used single classifier. In comparing prediction models performance, DT was the main choice among the researchers which was used in 11 studies followed by NB/ k-NN (six studies), RF (five studies) and ANN/ SVM (four studies). Based on the review, it can be concluded that DT is the most popular choice among the researchers in predicting student drop out as it is easy to comprehend and produce high-performance prediction results. Other than that, over the recent years, classification model using ensemble learning, especially RF had been increasingly popular among researchers because the performance outcome is very high as compared to a single classifier. Thus, a comparative study between classifier, particularly DT, ANN and RF is very much needed to discover the best prediction models for student drop-out among B40 students.

III. Research Methodology

In general, this research was conducted in three phases which were Phase I – Feasibility Study, Phase II – Data Preparation ad Phase III – Modeling and Evaluation. Fig. 1 shows detailed activities for each phase in research methodology. Fig. 1 illustrated phases and details of activities for each phase for research methodology in this study. There were three softwares used in this study, which were RapidMiner for prediction models construction and performance evaluation, MariaDB database to store and pre-process data, along with SPSS for attribute selection and statistical test.

A. Data Preparation

1) Data acquisition: The dataset was provided by Bahagian Pembangunan dan Perancangan Dasar (BPPD), Kementerian Pendidikan Malaysia (Pendidikan Tinggi) which consists of 44,406 records with 23 attributes. The dataset holds student's records from 20 public universities for bachelor degree programmes, who have dropped out or graduated from the year 2014 to 2017 intake.

2) Data Pre-processing: Pre-processing of data is a method of transforming a dataset in order to better expose the information quality to the mining tool. Real world data is often incomplete, incoherent and can contain noise such as errors and outliers. Pre-processing data is therefore required to ensure that data is formatted for a given miner tool and must be adequate for a given method. Data cleaning was performed using dimension reduction process. Attributes with more than 20,000 data unavailable, redundant or obsolete were deleted from the dataset. Incomplete records or outliers were also discarded (Table II). Data cleaning also ensured that the dataset included only student records with B40 household income (not more than RM4,387.00).

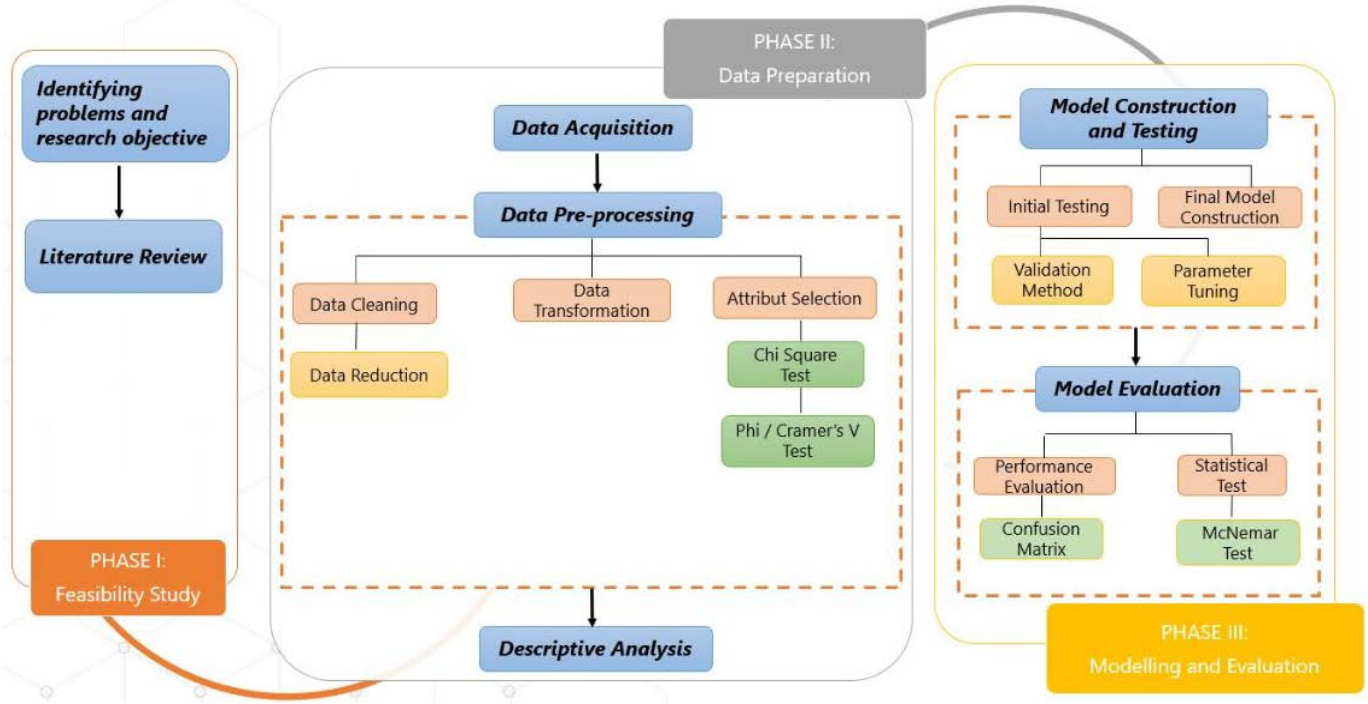


Fig. 1. Research Methodology.

TABLE II. RAW DATA ATTRIBUTES

Bil	Attribute Name	Data Type
1	student_id	nominal
2	date_of_birth	number (date)
3	gender	nominal
4	maritial_status	nominal
5	place_of_birth	nominal
6	postcode	nominal
7	state_legislative_assembly	nominal
8	parliament	nominal
9	country	nominal
10	institution_code	nominal
11	institution	nominal
12	citizenship	nominal
13	programme_code	nominal
14	programme	nominal
15	study_mode	nominal
16	study_method	nominal
17	sponsorship	nominal
18	family_income	nominal
19	student_status	nominal
20	session_intake_id	nominal
21	end_date	number (date)
22	qualification	nominal
23	cgpa_t1	number

Next, the data were transformed into a structure or understandable format befitting data mining. There were two new attributes constructed from existing attributes which were age from date_of_birth and class from student_status. Attribute class was the class label in this study. In attribute student_status, records with data' Berhenti' or 'Diberhentikan' were translated into 'C' in attribute class which represented students who dropped out while records with data' Tamat' were translated into 'G' which represented students who manage to graduate. Furthermore, attributes with varieties of data were aggregated or generalized by using the hierarchical concept.

Only relevant attributes were selected and used in building the prediction models. For this reason, the Chi-Square test was used to assess the relationship between attributes and the class label. Attributes with test result $\boldsymbol{p} < 0.05$ were considered having significant association with the class label. Afterwards, further tests were performed using Phi (ϕ) or Cramer's V (V) to each associated attribute in order to measure the strength of association. The value of ϕ or V is between 0 to 1, with 1 being the strongest and 0 being the weakest. The interpretation of the association between attributes and class label is, as shown in Table III.

Refer to Table IV, Chi-Square test results showed that all attributes had a significant association with the class label as the ***p***-value for each attribute that was less than 0.05. However, based on Cramer's V/Phi test result, *place_of_birth* and *family_income* were discarded from the dataset as their association level with the class label can be ignored. The final dataset for model construction consisted of 28,844 records with 9 regular attributes and 1 special attribute (class) (Table V).

B. Descriptive Analysis

Table VI and Table VII report the statistical analysis for each attribute after pre-processing data activity. Students who dropped out (class 'C') only represented 19.22% data in this study as compared to 80.78% of students who managed to graduate (class 'G'). More than half of the students (59.47 %) come from UiTM while UKM had the lowest number of students (0.21%).

TABLE III. PHI OR CRAMER'S V INTERPRETATION

Nilai Phi / Cramer's V	**Interpretation of Association**
0.00–0.10	Negligible
0.10–0.20	Weak
0.20–0.40	Moderate
0.40–0.60	Relatively Strong
0.60–0.80	Strong
0.80–1.00	Very Strong

Source : Lee 2016.

TABLE IV. X2 AND PHI/CRAMER'S V RESULTS (SORT BY ATTRIBUTE RANKING)

Attribute	**$\chi 2$**	***p*-Value**	**Cramer's V / Phi**	**Interpretation of Association**
cgpa_t1	18090.327	0.0000	0.79	Strong
institution	7084.965	0.0000	0.50	Relatively Strong
study_mode	2394.091	0.0000	0.29	Moderate
programme	2307.152	0.0000	0.28	Moderate
sponsorship	1838.139	0.0000	0.25	Moderate
qualification	1690.690	0.0000	0.24	Moderate
marital_status	706.977	0.0000	0.16	Weak
gender	587.057	0.0000	0.14	Weak
age	429.402	0.0000	0.12	Weak
place_of_birth	126.723	0.0000	0.07	Negligible

family_income	62.190	0.0000	0.05	Negligible

TABLE V. ATTRIBUTES IN FINAL DATASET (AFTER PRE-PROCESSING ACTIVITY)

Attribute	**Data Type**	**Details / Data**
age	nominal	Age group during enrolment in bachelor's degree Programme (< =20, >= 21)
gender	nominal	Gender (Male, Female)
marital_status	nominal	Marital Status (Single, Married, Divorced/Widowhood)
institution	nominal	Public Universities
programmes	nominal	Student's bachelor's degree programmes group (Engineering, Humanities, Art, Health, Journalism & Information, Computing, Mathematics & Statistics, Manifacturing & Processing, Education, Security Service, Transport Service, Social Service, Environmental Protection, Business & Administration, Agriculture, Forestry & Fisheries, Sains – Broad Programme, Physical Science, Life Science, Social Science, Social & Behavioural Science, Architecture & Building, Law, Veterinary)
study_mode	nominal	Student's mode of study (Part Time, Full Time)
sponsorship	nominal	Student's education financing (Government Agencies, Private Institutions, Self Funding, Education Loan, Foundation, Others)
qualification	nominal	Student's qualification for bachelor's degree admission (Pre-universiy, Diploma, Matriculation, SPM, STPM, Others)
cgpa_t1	nominal	Students's first year CGPA (Below 2.00, 2-2.99, 3-3.49, 3.5-4.00)
class	nominal	Class label (C, G)

TABLE VI. STATISTICAL DATA FOR ATTRIBUTE CGPA_T1, STUDY_MODE, SPONSORSHIP, QUALIFICATION, MARITAL STATUS, GENDER, AND AGE

Attribute	**Data**	**Sum**	**% (Sum)**	**Class**		**% C**
				C	**G**	
cgpa_t1	Below 2.00	3778	13.10	**3734**	44	**98.84**
	2-2.99	9573	33.19	1255	8318	13.11
	3-3.49	**11406**	**39.54**	429	10977	3.76
	3.5-4.00	4087	14.17	127	3960	3.11
study_mode	Part Time	566	1.96	563	3	**99.47**
	Full Time	**28278**	**98.04**	**4982**	23296	17.62
sponsorship	Government Agencies	1080	3.74	305	775	28.24
	Private Institutions	21	0.07	0	21	0.00
	Self Funding	**18348**	**63.61**	**3348**	15000	18.25
	Education Loan	7735	26.82	954	6781	12.33
	Foundation	24	0.08	1	23	4.17
	Others	1636	5.67	937	699	**57.27**
qualification	SPM	118	0.41	114	4	**96.61**
	STPM	5930	20.56	1320	4610	22.26
	Matriculation	3998	13.86	1449	2549	36.24
	Pre-university	502	1.74	161	341	32.07
	Diploma	**18004**	**62.42**	**2420**	15584	13.44
	Others	292	1.01	81	211	27.74
marital_status	Single	**28592**	**99.13**	**5331**	23261	18.65
	Married	238	0.83	203	35	**85.29**
	Divorced/Widowhood	14	0.05	11	3	78.57
gender	Male	8448	29.29	2362	6086	**27.96**
	Female	**20396**	**70.71**	**3183**	17213	15.61
age	< = 20	12348	42.81	**3060**	9288	**24.78**
	> = 21	**16496**	**57.19**	2485	14011	15.06

TABLE VII. NUMBER OF DROPPED OUT AND GRADUATED STUDENTS ALONG WITH DROP OUT PERCENTAGE BY PROGRAMME GROUPS

Programme Groups	Class		Sum	Drop Out Percentage
	C	G		
Sains - Broad Programme	34	0	34	100.00
Veterinary	10	0	10	100.00
Environmental Protection	2	0	2	100.00
Transport Service	1	0	1	100.00
Social Service	90	19	109	82.57
Agriculture, Forestry & Fishery	61	27	88	69.32
Security Service	53	25	78	67.95
Life Science	116	57	173	67.05
Education	67	79	146	45.89
Law	63	104	167	37.72
Architecture & Buliding	374	633	1007	37.14
Engineering	**1313**	2663	3976	33.02
Journalism & Information	230	843	1073	21.44
Personal Service	156	609	765	20.39
Social & Behavioural Science	261	1029	1290	20.23
Computing	364	1440	1804	20.18
Art	315	1418	1733	18.18
Health	117	564	681	17.18
Physical Science	168	972	1140	14.74
Manifacturing & Processing	169	1203	1372	12.32
Humanities	151	1086	1237	12.21
Business & Administration	1303	**9506**	10809	12.05
Mathematics & Statistics	127	1022	1149	11.05

Majority of B40 students managed to obtain a CGPA higher than 2.00 in the first year of their study. Nearly 40% of the students obtained their first year CGPA between 3.00 – 3.49. Even though students with first CGPA lower than 2.00 percentage is the lowest (13%), this group is most likely to drop-out as almost all of the students (98.84 %) did not continue their study. Business and Administration programme group contributed the largest number of students (10,809 students followed by Engineering (3,976 students) and the lowest was Transport Service with only one student. Further analysis also found that six out of ten programme groups with most numbers of students that dropped out and obtained CGPA lower than 2.00 were from Science, Technology, Engineering and Mathematics (STEM) major (Engineering, Computing, Manifacuturing and Processing, Mathematics and Statistics, Physical Science and Architecture).

Almost two-third of B40 students (63.61%) self-funded their study while the balance (36.39%) used education loan or received financial aid from government agencies, a private institution, foundation or other sources. Self-funded students were also presumed to drop-out as 60% of the students quit their study. Being a part-time student also can be a disadvantage, as 99.47% of them failed to graduate. Students who were single got a higher chance of finishing their degree as their percentage of dropping out was very low as compared to married or divorced/widowed students. When it comes to gender, over 70% of students in this study were female, but their drop out rate is 15.61% lower than male students. Finally, students in the age group o 20 years old and below when enrolled in a bachelor degree programme are most likely to drop-out than students in the age group 21 years and above.

C. Modelling and Evaluation

1) Model construction and testing: Each prediction model (DT, RF and ANN) was tested beforehand to determine the validation method and algorithms parameters that can be used to produce high performance prediction model. All nine attributes were used in validation method testing and parameter tuning. Prediction model validation was tested using holdout (70 %– 30% and 60% – 40%) and 10-folds cross-validation methods, and the latter was chosen as it gave the highest accuracy results for the majority of the models. Next, each prediction model also was constructed repeatedly by using a different parameter to achieve highest accuracy result. Parameter tuning results are as shown in Table VIII, and these parameters were used in building the final prediction models.

TABLE VIII. PARAMETER TUNING RESULTS

Prediction Models	Parameter
Decisin Tree (DT)	*criterion*: *information_gain* *maximal_depth*: 30
Randon Forest (RF)	*number_of _tree*: 90 *criterion: information_gain* *maximum_depth*: 50
Artificial Neural Network (ANN)	*Hidden layer*: 8 *Learning rate*: 0.01

Different numbers of attributes were used in building final prediction models. At first, the models were build with attributes that had moderate to strong relationship with class label. Subsequently, attributes with weak relationship were added one-by-one based on attribute ranking. The importance of weak attributes can't be neglected as they might be useful in producing high performance prediction models. Table IX shows attribute representation for final models construction.

2) Model evaluation: Prediction model's performance was evaluated by comparing the value of accuracy, precision, recall and F measure. Those values were calculated based on the confusion matrix technique, as shown in Table X. Prediction results and actual class were put in a matrix for comparison depending on a positive and negative value. Class 'C' was marked as positive value while class 'G' was negative.

In addition to performance comparison, the statistical test was performed to decide the best prediction model. This study used the McNemar test to determine if there was a significant difference statistically to the proportion of error between two prediction models with a significance level of 0.05 ($\alpha = 0.05$). The significant difference between the proportion of error of two prediction models is also interpreted as a significant difference in performance between two prediction models (Dietterich, 1998).

TABLE IX. ATTRIBUTE REPRESENTATION FOR FINAL MODELS CONSTRUCTION

Attribute Representation	**Attributes**
6 Attributes	*cgpa_t1, institution, study_mode, programme, sponsorship, qualification*
7 Attributes	*cgpa_t1, institution, study_mode, programme, sponsorship, qualification, marital_status*
8 Attributes	*cgpa_t1, institution, study_mode, programme, sponsorship, qualification, marital_status, gender*
9 Attributes	*cgpa_t1, institution, study_mode, programme, sponsorship, qualification, marital_status, gender, age*

TABLE X. CONFUSION MATRIX

		Prediction Class	
		Positive (C)	**Negative (G)**
Actual Class	**Positive (C)**	True Positive (TP)	True Negative (TN)
	Negative (G)	False Negative (NP)	True Negative (TN)

IV. RESULTS AND DISCUSSION

The results indicated that RF model gives the highest accuracy in predicting student drop-out with 95.93%, followed by ANN with 95.86% and DT with 95.84%. The highest accuracy for RF model was produced with seven attributes while the others by using six attributes. However, the accuracy for RF model with six attributes was higher than ANN and DT models with the same number of attributes (refer Fig. 2). Consistently, RF also yields a higher accuracy rate than the other two models, even by applying different numbers of attributes. This showed that prediction performance could be improved with the use of ensemble learning. This result is also inline with research outcome by [20], which predicts students' drop-out in higher learning institution, revealing that the accuracy of the prediction model using RF l was higher than DT.

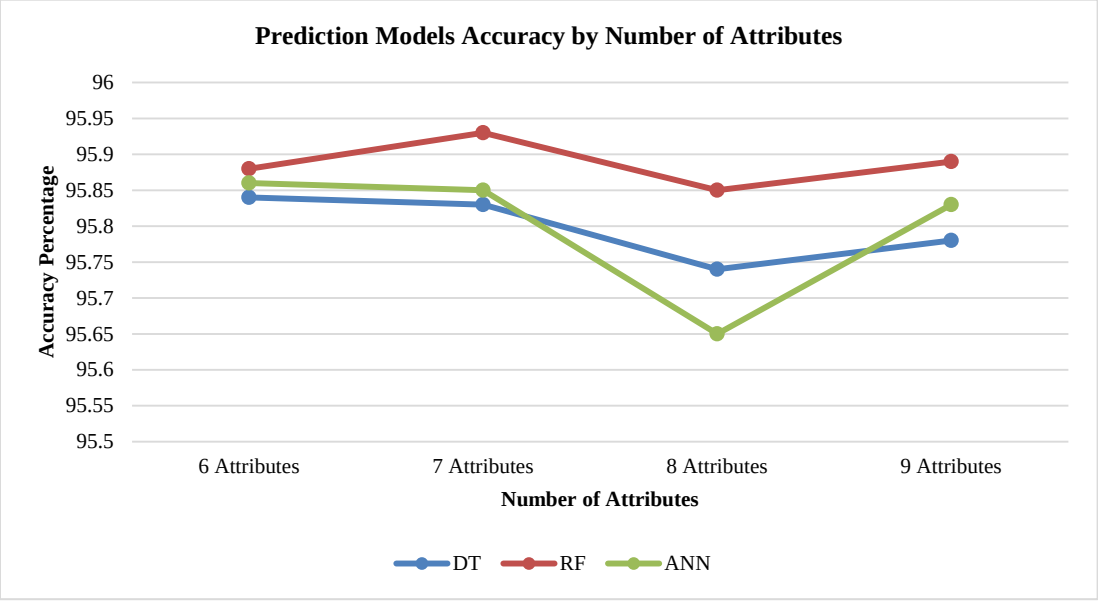


Fig. 2. Prediction Model Accuracy based on Number of Attributes.

TABLE XI. PERFORMANCE RESULTS COMPARISON BETWEEN STUDENTS DROP-OUT PREDICTION MODELS WITH HIGHEST ACCURACY

Prediction Models	Evaluation Parameters			
	Accuracy (%)	Recall (%)	Precision (%)	F Measure
DT (6 attributes)	95.84	80.99	96.83	0.882
RF (7 attributes)	**95.93**	**81.26**	**97.10**	**0.885**
ANN (6 attributes)	95.86	81.08	96.89	0.883

Performance between prediction models was evaluated by comparing the value of accuracy, recall, precision and F measure (refer Table XI). Aside from accuracy, the results also showed that RF leads ANN and DT with regards to recall value with 81.26%, 81.03% and 80.99%. This means that RF model succeeded in predicting more students who will drop-out (class 'C') correctly from the total number of student who were actually dropped out in this study. Likewise, the highest value for precision was also recorded by RF with 97.10% which means that the model was able to predict more class 'C' precisely from the total number of students who were predicted to drop-out. ANN took second place in precision with 96.89% while DT the last place with 96.83%. When comparing F measure, RF also the highest with 0.885, followed by ANN with 0.883 and DT with 0.882.

Generally, RF is the best model in predicting drop-out among B40 students in this study as it outperformed the other two models with reference to accuracy, recall, precision and F measure, subsequently ANN and DT models. Nevertheless, the difference in accuracy and F measure value between the three models were very narrow, with 0.07% to 0.09% and 0.002% to 0.003%, respectively. Hence, the statistical test (McNemar) results were referred to in determining a significant difference in performance between the prediction models.

Based on Table XII, McNemar test results proved that statistically:

1) DT and RF models had a significant difference in proportion of error;

2) DT and ANN models had no significant difference in proportion of error; and

3) ANN and RF models had no significant difference in proportion of error.

This implied that even though RF is the best model in predicting drop-out among B40 students in this study, there is a significant difference in performance only between RF and DT, but contrarily, no significant difference in performance between RF and ANN.

TABLE XII. MCNEMAR'S *P*-VALUE (*2 SIDED*) RESULTS

Prediction Models	DT	RF	ANN
DT		**0.004**	0.719
RF	**0.004**		0.224
ANN	0.719	0.224	

V. CONCLUSION

Drop-out prediction among B40 students in bachelor's degree programmes can be implemented by using classification technique. Prediction model using RF was selected as the best model in this study as it outperformed ANN and DT in accuracy, recall, precision and F measure. However, statistically, the difference in performance was only significant between RF and DT, not between RF and ANN.

Results of this research are expected to benefit B40 students, public universities and the government. Early prevention steps can be deployed by public universities to avoid drop-out to produce more graduates. B40 students who are at risk to drop-out will be able to graduate with the help of their university and getting better job opportunities that will improve their socioeconomic status. These students also will become assets to the government as professionals and skilful worker that can be contributed to the nation's future development.

In future, this study can be furthered by applying regression technique to predict when attrition will happen with the additional data of students who are still studying and the exact date of drop-out. Besides, the association rule technique can be applied to discover hidden patterns that can be used to identify students at risk, and the results can be verified by the experts from the ministry or universities.

ACKNOWLEDGMENT

The authors would like to thank Universiti Kebangsaan Malaysia (UKM) and Ministry of Education, Malaysia (MOE) under the Research University Grant (project code: GUP-2019-060 and FRGS/1/2018/ICT02/UKM/02/6) for funding and supporting this research.

REFERENCES

[1] Abu, R. Hamdan, R. and N.S. Sani, "Ensemble Learning for Multidimensional Poverty Classification," Sains Malaysiana., vol. 49(2), pp.447-459 2020.

[2] N. S. Sani, M. A. Rahman, A. A. Bakar, S. Sahran, and H. M. Sarim, "Machine learning approach for bottom 40 percent households (B40) poverty classification," IJASEIT, vol. 8, pp. 1698-1705, 2018.

[3] J. D. Holliday, N. Sani, and P. Willett, "Calculation of substructural analysis weights using a genetic algorithm," J. Chem. Inf. Model, vol. 55, pp. 214-221, 2015

[4] J. D. Holliday, N. Sani, and P. Willett, "Ligand-based virtual screening using a genetic algorithm with data fusion," Match-Commun. Math. Co., vol. 80, pp. 623-638, 2018.

[5] N. S. Sani, I. I. S. Shamsuddin, S. Sahran, A. H. A. Rahman, and E. N. Muzaffar, "Redefining selection of features and classification algorithms for room occupancy detection," IJASEIT, vol. 8, pp. 1486-1493, 2018.

[6] S. Shabudin, N. S. Sani, K. A. Z. Ariffin and M. Aliff, "Feature Selection for Phishing Website Classification," International Journal of Advanced Computer Science and Applications, vol. 11(4), pp. 587-595, 2020.

[7] T. K. M. Zali, N. S. Sani, A. H. Abd Rahman, and M. Aliff, "Attractiveness Analysis of Quiz Games," International Journal of Advanced Computer Science and Applications, vol. 10(8), pp. 205-210, 2019.

[8] Z. A. Othman, A. A. Bakar, N. S. Sani, and J. Sallim, "Household Overspending Model Amongst B40, M40 and T20 using Classification Algorithm," International Journal of Advanced Computer Science and Applications, vo1. 11(7), pp. 392-399, 2019.

[9] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," Procedia Comput. Sci., vol. 72, pp. 414–422, January 2015.

[10] M. A. Al-Barrak and M. Al-Razgan, "Predicting students final GPA using decision trees: A case study," Int. J. Inf. Educ. Technol., vol. 6, pp. 528–533, July 2016.

[11] E. A. Amrieh, T. Hamtini and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's performance," 2015 IEEE Conf. Appl. Electr. Eng. Comput.Technol. (AEECT), Amman, Jordan, pp. 1–5. November 2015.

[12] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," Comput. Educ., vol. 113, pp. 177–194, October 2017.

[13] F. Widyahastuti and V. U. Tjhin, "Predicting students performance in final examination using linear regression and multilayer perceptron," 2017 10th Int. Conf. Human Syst. Interact. (HSI), pp. 188–192, July 2017.

[14] R. Umer, T. Susnjak, A. Mathrani, and S. Suriadi, "Predicting student's academic performance in a MOOC environment," 11th Int. Conf. Data Mining, Comput., Commun. Ind. Appl. (DMCCIA-2017), pp. 119–124, December 2017. [Umer, R., Science, M., Susnjak, T., Mathrani, A., Science, M., & Suriadi, S].

[15] E. Aguiar, H. Lakkaraju, N. Bhanpuri, D. Miller, B. Yuhas, and K. L. Addison, "Who, when and why: A machine learning approach to prioritizing students at risk of not graduating high school on time," Proc. 5th Int. Conf. Learning Anal. Knowl., New York, pp. 93–102, March 2015.

[16] W. W. Yaacob, N. M. Sobri, S. M. Nasir, N. D. Norshahidi, and W. W. Husin, "Predicting student drop-out in higher institution using data mining techniques," J. Physics: Conf. Series 2020, vol. 1496, 012005, March 2020.

[17] N. Bedregal-Alpaca, V. Cornejo-Aparicio, J. Zárate-Valderrama, and P. Yanque-Churo, "Classification models for determining types of academic risk and predicting drop-out in university students, Int. J. Adv. Comput. Sci. Appl., vol. 11, pp. 266–272, 2020.

[18] J. S. Gil, A. J. P. Delima, and R. N. Vilchez, "Predicting students' drop-out indicators in public school using data mining approaches." Int. J. Adv. Trends in Computer Sci. Eng., vol. 9, pp. 774–778, 2020.

[19] M. Mardolkar and N. Kumaran, "Forecasting and avoiding student drop-out using the K-nearest neighbor approach," SN Computer Sci., vol. 1, pp. 1–8, March 2020.

[20] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S.Ventura, "Early drop-out prediction using data mining: A case study with high school students," Expert Systems, vol. 33, pp. 107–124. February 2016.

[21] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," Comput. Educ., vol. 143, p. 103676, January 2020.

[22] A. Viloria, J. G. Padilla, C. Vargas-Mercado, H. Hernández-Palma, N. O. Llinas, and M. A. David, "Integration of data technology for analyzing university drop-out," Procedia Comput. Sci., vol. 155, pp. 569–574, January 2019.

[23] A. Sangodiah, P. Beleya, M. Muniandy, L. E. Heng, and C. Ramendran, "Minimizing student attrition in higher learning institutions in Malaysia using support vector machine," J. Theoritical Appl. Inf. Technol., vol. 71, pp. 377–385, January 2015.

[24] S. S. Shariff, N. A. Rodzi, K. A. Rahman, S. M. Zahari, and S. M. Deni, "Predicting the "graduate on time (GOT)" of PhD students using binary logistics regression model," AIP Conf. Proc. 2016, vol. 1782, p. 050015, October 2016.

[25] K. Limsathitwong, K. Tiwatthanont, and T. Yatsungnoen, "Drop-out prediction system to reduce discontinue study rate of information technology students," Proc. 2018 5th Int. Conf. Business and Industrial Research: Smart Technol. Next Generation of Information, Eng., Business and Social Sci. (ICBIR 2018), pp. 110–114, May 2018.

[26] Y. Chen, A. Johri, and H. Rangwala, "Running out of STEM: A comparative study across STEM majors of college students at-risk of dropping out early," Proc. 8th Int. Conf. Learn. Anal. Knowl., pp. 270–279, March 2018.

[27] J. M. Ortiz-Lozano, A. Rua-Vieites, P. Bilbao-Calabuig, and M. Casadesús-Fa, "University student retention: Best time and data to identify undergraduate students at risk of drop-out," Innovations Educ. Teach. Int., vol. 57, 74–85, January 2020.

[28] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student drop-out in higher education," arXiv preprint arXiv:1606.06364., June 2016.

[29] J. Liang, J. Yang, Y. Wu, C. Li, and L. Zheng, "Big data application in education: Drop-out prediction in edx MOOCs," Proc. - 2016 IEEE 2nd Int. Conf. Multimedia Big Data, pp. 440–443, April. 2016.

[30] D. D. Pokrajac, K. R. Sudler, P. Y. Edamatsu, and T. Hardee, "Prediction of retention at historically black college/university using artificial neural networks," 2016 13th Symp. Neural Networks and Applications (NEUREL), pp. 1–6, November 2016.

[31] G. S. Abu-Oda, and A. M. El-Halees, "Data mining in higher education: University student drop-out case study," Int. J. Data Mining & Knowl. Manage. Proc., vol. 5(1), pp. 15–27, January 2015.

[32] P. Strecht, L. Cruz, C. Soares, J. Mendes-Moreira, and R. Abreu, "A comparative study of classification and regression algorithms for modelling students' academic performance," Proc. 8th Int. Conf. Educ. Data Mining, 392–395, January 2015.

[33] A. Siri, "Predicting drop-out at university using Artificial Neural Networks," Italian J. Soc. Educ., vol. 7, pp. 225–247, June 2015.

Proficiency Assessment of Machine Learning Classifiers: An Implementation for the Prognosis of Breast Tumor and Heart Disease Classification

Talha Ahmed Khan[1]
British Malaysian Institute, Universiti Kuala Lumpur Kuala Lumpur, Malaysia, Usman Institute of Technology (NED), Karachi, Pakistan

Kushsairy A. Kadir[2]
British Malaysian Institute, Universiti Kuala Lumpur Kuala Lumpur, Malaysia

Shahzad Nasim[3]
Faculty of Engineering Sciences, Technology and Management, Ziauddin University Karachi, Pakistan

Muhammad Alam[4]
CCSIS, Institute of Business Management (IoBM), Karachi, Pakistan, MIIT-Universiti Kuala Lumpur, Kuala Lumpur, Malaysia

Zeeshan Shahid[5]
Electrical Engineering, Institute of Business Management (IoBM), Karachi, Pakistan

M.S Mazliham[6]*
Malaysian France Institute, Universiti Kuala Lumpur Kuala Lumpur, Malaysia

***Abstract*—Breast cancer and heart disease can be acknowledged as very dangerous and common disease in many countries including Pakistan. In this paper classifiers comparative study has been performed for the tumor and heart disease classification. Around one lac women are diagnosed annually with this life-threatening disease having no family history of the disease. If it is not treated on time it may grow and spread to the other parts of human body. Mammograms are the X-rays of the breast which can be used for the screening of cancer tumor. Prior identification of breast cancer may increase the chance of survival up to 70 percent. Tumors which causes cancer can be categorized into two types: a) Benign and b) Malignant. Benign tumor can be explained as the tumor which are not attached to neighbor tissues or spread in the other parts of the body. In Malignant tumor, other parts may be affected by it as it can grow and spread in the other parts of the body. To classify the tumor as Malignant or Benign is very complex as the similarities of cancer tumor and tumor caused by the skin inflammation are almost same. The early identification of Malignant is mandatory to protect the patient life. Diversified medical methods based on deep learning and machine learning have been developed to treat the patients as cancer is a very serious and crucial issue in this era. In this research paper machine learning algorithms like logistic regression, K-NN and tree have been applied to the breast cancer data set which has been taken from UCI Machine learning repository. Comparative study of classifiers has been performed to determine the better classifier for the robust prediction of breast tumors. Simulated results proved that using Logistic regression, ninety-one percent accuracy was achieved. The research showed that logistic regression can be applied for the accurate and precise early prediction of breast cancer. Cardiovascular disease is very common throughout the world. It has been noticed that health in cardiac patients that there are so many factors which causes heart disease or heart attack. The factors leading to the heart failure includes varying blood pressure, high sugar, cardiac pain, and heart rate, high cholesterol level (LDL), artery blockage and irregular ECG signals. Many researchers proved that stress in patients can also be the reason for the heart disease. Higher numbers of cardiac surgeries like angioplasty and heart by-pass are performed on annual basis. Actually, people don't care about their lifestyle and diet and fully ignore the symbols. It can be early predicted and cured if proper testing and medication for heart is done. Sometimes there is a false pain which has the same feeling like angina pain depicting cardiovascular disease. To reduce the false alarm and robustly classify the heart disease, several machine learning approaches have been adopted. In proposed research for the accurate classification of heart disease comparison has been performed among support vector machine (SVM), K-nearest neighbors K-NN and linear discriminant analysis. Simulated results demonstrated that Support vector machine was found to be a better classifier having an accuracy of 80.4%.**

Keywords*—Breast cancer; benign; malignant; logistic regression; cardiovascular disease; heart disease diagnosis; support vector machine; classifiers; k-nearest neighbors

I. Introduction

Twenty-five percent of women die due to the breast cancer in the ages of thirty-five to forty. Mammography is usually performed to enhance the radiographic decisions. SENOLOG was developed for the breast therapy assistance using SENOBASE [1]. RF-ELM classifier was applied to find out the tumor from the digital mammogram. Mammogram images were taken from MIAS database. Kurtosis, mean, standard deviation, correlation coefficient, entropy and variance were chosen for the accurate classification. RF-ELM was found to be very competent classifier for the diagnosis of breast cancer [2]. This research paper is divided into four main sections. Section one explains the introduction. Second part discusses the implementation of classifier algorithms for breast tumor identification. Section three elaborated the heart disease

*Corresponding Author

classification and its implementation. Results and conclusion have been discussed in section four.

A. Existing Methods for the Identification of Malignant Tumor

Local binary patterns (LBP) were applied using mammograms. Data set was collected from DDSM. LBP using mammograms achieved the accuracy of 84% [3]. Mammography is acknowledged as a good strategy for the screening of tumors. Generally, mammogram analysis is very complex as the image comprises of various little differences of different tissues [4]. A novel hybrid approach of digital image processing was adopted to analyze the mammograms. Using this novel approach, the early identification of breast cancer at the stage of micro calcification was achieved leading to the higher accuracy of proposed technique [5]. Digital image based elasto-tomography was developed as a prototype for the evaluation of breast cancer. Segmentation was performed to identify the model of breast. Using this system up to 10 mm tumor could be detected in a silicon phantom breast [6]. Ensemble empirical mode decomposition (EEMD) has been proposed for the prior determination of breast tumors using ultra-wide band (UWB) microwave imaging. Approximately 4 mm tumor has been identified in inside the glandular tissue whose di-electric constant was 35 in a breast model [7]. The pulsed confocal approach was proposed to improve the identification of breast tumors. Two-dimensional finite difference time domains (FDTD) analysis was conducted to determine the 2mm tumor in the presence of clutter [8]. It has been observed that rumors possess different permittivity and conductivity with respect to the surrounding tissues. Electrical impedance spectroscopy (EIS) was used to classify the normal tissues and malignant [9].

Fig. 1(a) and 1(b) elaborated that homogenous breast model was designed. Incident wave of 6 Ghz having vertical polarization was tested. Artificial Neural network was designed to evaluate the scattered electromagnetic waves. The dielectric values for malignant and normal tissue was randomly chosen [11-12].

B. Problem Statement

Early detection of Breast cancer has become very crucial issue in the medical science as 30% women die annually due to the breast cancer. Women usually ignore the tumor because of the lack of awareness for the breast cancer. To classify the tumor as Malignant or Benign is very complex as there is a misconception and confusion regarding these two classes [1-4]. The last stage symptoms are also similar with the normal inflammatory conditions. Therefore, vigorous early breast cancer detection was needed. Mammography based screening is usually used to evaluate the cancer tumor on early basis as well when it is small. Many clinical laboratories are there to record the mammograms of breast which is the X-ray of breast. Data acquisition for the breast tumor can also be possible as the size, shape adhesion, location and other attributes related to cancer tumor can also be recorded [2-6]. To make it more certain and enhance the accuracy of Malignant tumor identification Machine Learning based decision making was required. The similarities of the tumor's symptoms are almost same as the inflammation of the skin problem. Breast pain, swelling, and reddish skin are very common symptoms of cancer but people ignore it as they take it as normal skin inflammatory problem [9-11].

C. Methodology

Fig. 2 elaborates the main fundamental block diagram for the proposed breast cancer classification using machine learning. Clinical data acquisition was performed and collected from the UCI machine learning repository for applying the proposed classification models. Logistic regression, K-NN and decision tree classifiers were applied to determine the best predictive model for the breast cancer. Cross validation curve was also obtained for the comparison. Results were obtained in terms of accuracy, precision, prediction speed, ROC, true positive rate and false positive rate.

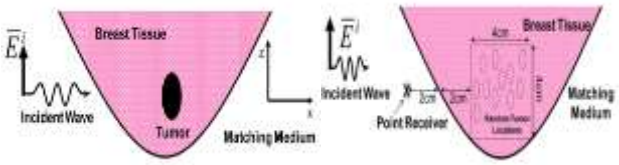


Fig. 1. (a); and (b); Breast Tumors at Random Locations [10].

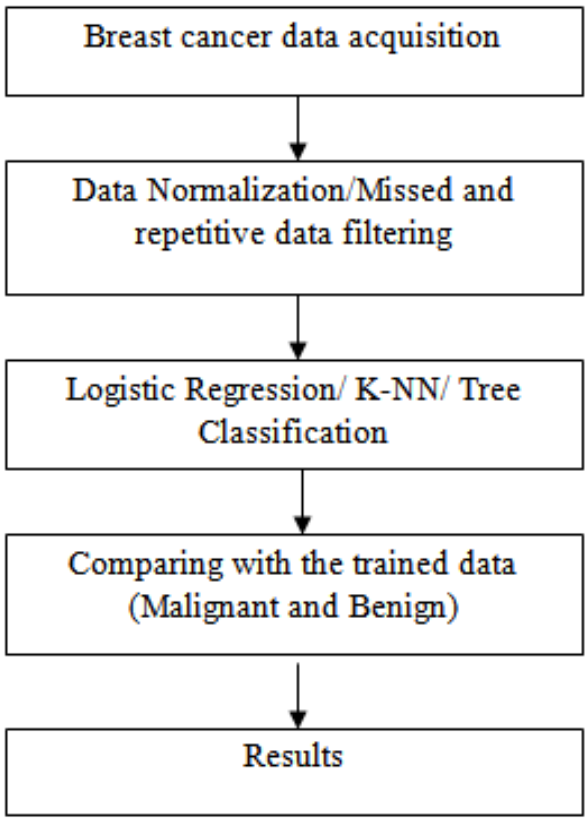


Fig. 2. Main Block Diagram.

D. Data Acquisition for Breast Cancer

Table I represents that Breast cancer dataset has been gathered from the database of UCI machine learning repository for the Benign and Malignant classification purpose. The data set has been used in many researches by using neural networks and machine learning based classification [13-14]. The data set comprised of eleven attributes including the patient ID. Column 1 describes the patient ID for each individual patient. Clump thickness has been mentioned in column 2. Clump is the bunch of close roots which have been grown with the tumor tissue. Uniformity of cell size has been represented in the column no. 3. Cell shape of tumor has been described in column no. 4. Column 5 shows the marginal adhesion for the tumor. Column no. 6 represents the single epithelial cell size. Epithelial cell is defined as the cell which protects the upper surface of the skin against germs and bacteria. Bare Nuclei has been demonstrated in the column no. 7. Bare nuclei are the cytology preparation which can be observed in the degeneration of cell. Bland chromatin has been defined in the column no. 8. Bland chromatin explains the pattern and texture of the Benign tumor. Usually in cancer cell the texture is found to be rough and harsh. Column no. 9 displays the normal nucleoli. Nucleoli depict the cell's response to the stress. Column no 10 displays the Mitosis attribute of breast tumor. Mitosis can be defined as the two daughter cells having same properties and number of chromosomes in parent cell like an ordinary tissue. Output results have been mentioned in the last column. Column 11 shows the output classes 1 and 2, 1 for Benign and 2 for Malignant [13-14].

TABLE I. BREAST CANCER DATA ACQUISITION [13-14]

Patient ID	CL Thick	Cell size U	Cell shape	Adhesion	Cell size	Bare Nuclei	Bland Chromatin	Nucleoli	Mitosis	Class
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1035283	4	2	1	1	2	1	2	1	1	2
1036172	1	1	1	1	1	1	3	1	1	2
1041801	2	1	1	1	2	1	2	1	1	2
1043999	1	1	1	1	2	3	3	1	1	2
1044572	8	7	5	10	7	9	5	5	4	4
1047630	7	4	6	4	6	1	4	3	1	4

II. IMPLEMENTATION OF CLASSIFIER ALGORITHMS FOR BREAST TUMOR IDENTIFICATION

A. Decision Tree Implementation

Decision tree is same as the tree in which the outputs are represented by leaves. Decision tree algorithm is able to classify and sort from roots to the leaf. For the information gain the entropy is estimated in information coding theory.

$$E = \sum \boldsymbol{i} = \mathbf{1} \text{ - } P \times \log_2(p_i) \quad (1)$$

Probabilistic classification can be positive or negative. Entropy for" t" training sample can be explained as:

$$(t) = -p+\log_2 p + -p-\log_2 p- \quad (2)$$

t = training of sample data.

$-p+\log_2 p$ = negative examples defined in data.

$+ -p-\log_2 p-$ = positive examples in data.

Fig. 3 graphically illustrated the ROC curve for the decision tree algorithm. ROC curve was plotted between true positive rate and false positive rate to assess the performance of the classifier. The area under the curve (AUC) defines the value up to which the classification model can classify. The values 0.40 and 0.92 were found to be false positive rate and true positive rate respectively in the graph. AUC was determined 0.70 that can be considered as a good classification model for the breast cancer prediction.

Fig. 4 represented decision tree confusion matrix. Usually confusion matrix is observed diagonally; all the values in diagonal show the true positive classes. Confusion matrix shows the performance evaluation of decision tree classifier. In Class "2- Benign" classification, true positive rate was found to be 92% and false positive rate was determined as 8%. For the class "4-Malignant" classification, 60% true positive rate was observed with 40% false positive rate.

B. K-Nearest Neighbors K-NN Classifier

Fig. 5 portrays the confusion matrix of K-NN algorithm. 88.2% accuracy was achieved by the K-NN for the classification of Benign and Malignant Tumor. The efficiency was estimated using true positive rate and false positive rate

values. Confusion matrix assessed the K-NN classifier and displayed that 92% true positive rate was achieved with 8% false positive rate for the classification of "class 1-Benign". In "Class-Malignant" classification, 80% true positive rate was achieved with 20% false positive rate.

Euclidean distance is estimated to determine the closest distance with the value of the K.

$$d = \sqrt{(x1 - xA1)^2 + (x2 - xA2)^2} \quad (3)$$

Fig. 6 shows ROC and area under the curve (AUC) for the K-NN classifier. The ROC curve has been plotted between true positive rate and false positive rate. Area under the curve (AUC) was found to be 0.86. It is slightly away from the 1. For good classification AUC must be close to 1.

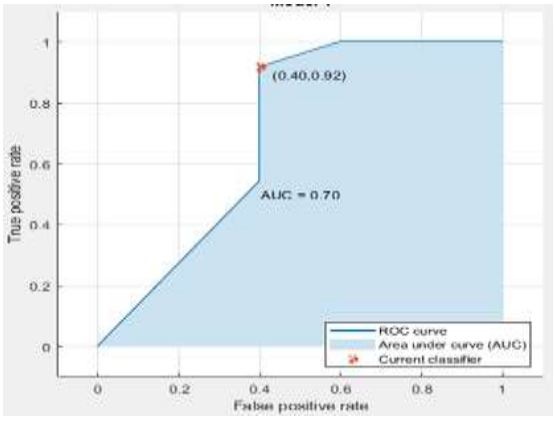


Fig. 3. Decision Tree ROC Curve.

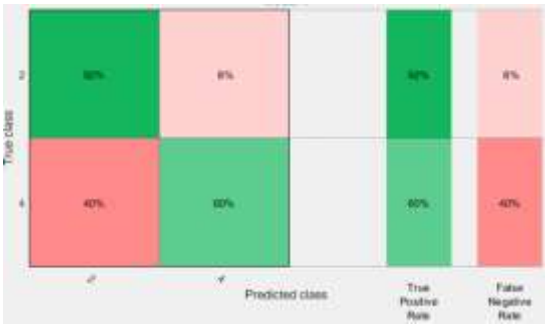


Fig. 4. Decision Tree Confusion Matrix.

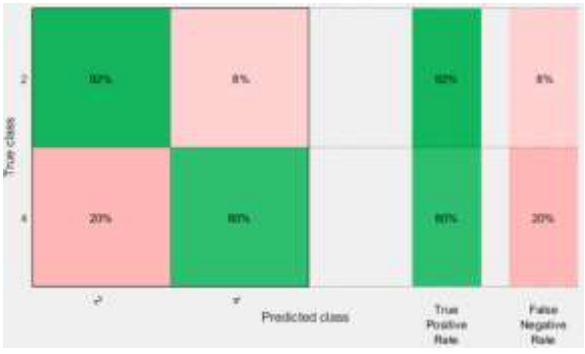


Fig. 5. K-NN Classifier Confusion Matrix.

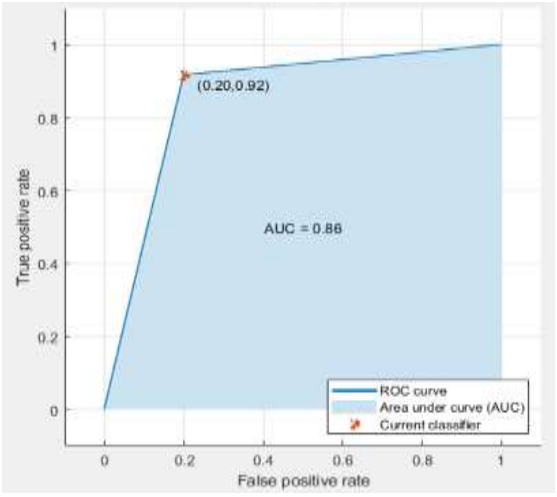


Fig. 6. K-NN ROC and AUC.

C. Logistic Regression

Logistic regression is defined as the regression model in which predictive model predict the output in binary. Outliers and missing values must be filtered out before processing the predictive regression analysis. Logistic function may be defined as:

$$\text{Logistic Regression}(x) = \frac{1}{1+\exp(-x)} \quad (4)$$

Fig. 7 elaborates the regression model for the logistic regression. It can be seen that all the attributes or the parameters related to the classification of classes have been entered to the logistic regression model for the prediction of tumor analysis. The weighted sum is converted into probability by logistic function.

$$o = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)}) \quad (5)$$

Where o is the predicted output, b0 is the bias and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

In Fig. 8, Logistic Regression confusion matrix portrayed that mean accuracy of 91.2% was achieved based on the true positive rate (TPR) and false positive rate (FPR). For the classification of Benign, confusion matrix showed that 100% true positive rate was achieved in classifying class 2 while 0% of false positive rate was achieved in the classification of class 1. It was also observed that 70% true positive rate was estimated in the classification of "class 2-Malignant" with the 30% false positive rate.

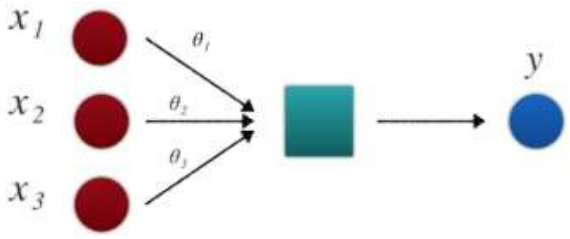


Fig. 7. Logistic Regression Sample Model.

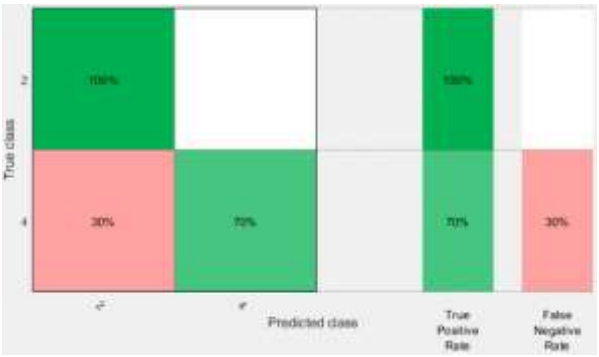


Fig. 8. Logistic Regression Confusion Matrix.

Fig. 9 illustrated graphically that area under the curve (AUC) for the logistic regression was found to be 0.89.

D. Fine Gaussian

In Fig. 10, Confusion matrix of fine Gaussian SVM elaborated that the algorithm performed very poor for the prediction of breast cancer as it classified all classes as class 1. Class 2 was not predicted at all therefore false negative rate was found to be 100% and false positive rate for class 2 was found to be 0%.

E. Comparative Study of Classifiers for the Prognosis of Breast Tumor

Fig. 11(a) shows the accuracy, prediction speed and training time for the Decision tree. Fig 11(b) explains the parametric analysis for the logistic regression. It can be observed from the parametric analysis that 91.2% accuracy has been achieved by logistic regression. Moreover, it can also be noticed that 88.2% accuracy was achieved in the trained predictive model for the breast tumor classification.

Table II showed that the Logistic Regression and K-NN performed better classification compared to the Decision tree and Fine Gaussian SVM in terms of Accuracy, prediction speed, training elapsed time, precision and area of under the curve. 91.2% accuracy was achieved by logistic regression for the benign and malignant classification.

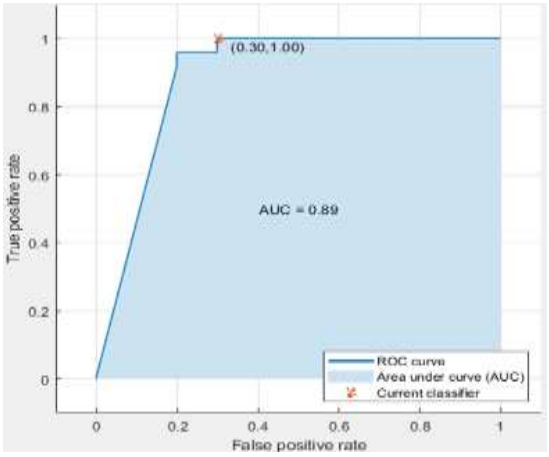


Fig. 9. ROC and AUC Curve for Logistic Regression.

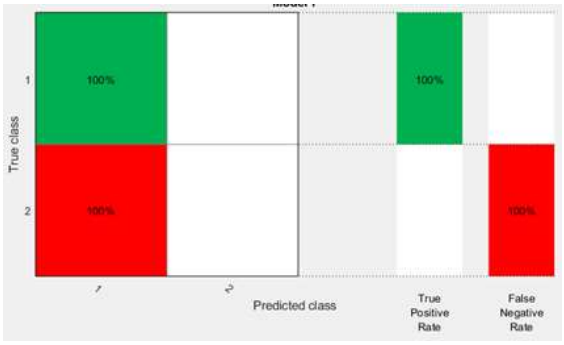


Fig. 10. Confusion Matrix of fine Gaussian SVM.

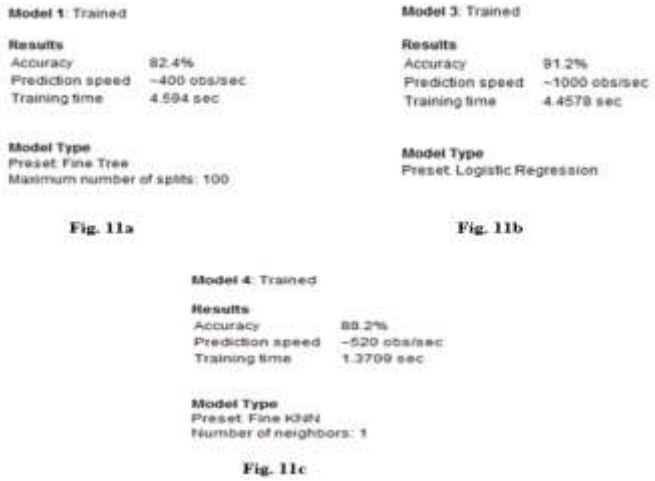


Fig. 11a Fig. 11b

Fig. 11c

Fig. 11. (a), (b), and (c) Displays the Trained Model for the Predictive Analysis of Breast Cancer.

TABLE II. TREE, K-NN AND LOGISTIC REGRESSION COMPARISON

Parameters	Decision Tree	K-NN	Logistic Regression	Fine Gaussian SVM
Accuracy	82.4%	88.2%	91.2%	59.8%
Prediction Speed	400 obs/sec	520 obs/sec	1000 Obs/sec	1700 obs/sec
Training time	4.594 seconds	1.3709 seconds	4.4578 seconds	0.8763 sec
Precision	0.82	0.89	0.90	0.59
AUC	0.70	0.86	0.89	0.60
Precision	0.83	0.89	0.93	0.55
F1-score	0.84	0.88	0.91	0.45
Recall	0.85	0.87	0.90	0.50

III. RECENT TRENDS FOR HEART DISEASE CLASSIFICATION

Cardiovascular disease is very common throughout the world even in United states of America in every thirty-four second one patient losses his or her life due to this silent disease [15]. Electrocardiography signals (P-wave, QRS complexes) and cardiac arrhythmias have also been processed and classified through convolutional neural network (CNN) to identify the heart disease [16]. Smoking and hypertension may also increase the chances of heart disease. Data mining

techniques were applied to the heart disease data (HDD) for the classification of heart disease [17]. Quadratic support vector machine and discriminant analysis have been performed in the MATLAB environment for the classification of heart disease [18]. Fuzzy based K-NN classifier was developed for the classification of pure cardiovascular disease. Training, testing and validation curve demonstrated that fuzzy based K-NN classifier worked better [19]. A wearable gadget was also fabricated for the real time data transmission. The health parameters including blood pressure, temperature and heart rate were optimized using particle swarm optimization for the optimal results in previous research [20].

Fig. 12 illustrated that cardiac arrhythmias and ECG signals classification were performed using, Fuzzy logic controller, MLP-PSO, Improved PSO(ImpSO) and Genetic algorithm [21-22]. The data was collected from PHYSIONET. Heart rate variability (HRV) has been used as a yardstick to measure the heart health. Heart rate variability signals have been processed for the classification of heart disease using artificial neural network (ANN) [23-26]. Cardiac arrhythmia can be categorized into following categories, Asystole, Bradycardia, Tachycardia, Ventricular Tachycardia and Ventricular flutter. Fuzzy logic was used to classify the heart disease as cardiac arrhythmia can be used for measuring the heart health [27]. Rs and QRS complexes were cleaned and classified using Pan and Tompkins algorithm [28]. Cardiac abnormalities have been observed for the heart rate classification [29]. ECG signals have been classified using fuzzy network for the heart disease classification [30]. Cardiac arrhythmias based on rate time series were forecasted using radial basis function [31]. 116 heart sounds were segmented using classification and regression trees [32]. An optimal architecture of multi-layer perceptron with the combination of particle swarm optimization has been designed for the prediction of cardiac arrhythmias [33]. A stand-alone system using DSK6713 was developed to measure the abnormalities in heart sound [34].

A. Problem Statement for Heart Disease Classification

Cardiovascular disease is very common throughout the world. Usually people ignore the early symptoms of the heart disease like people think the actual cardiac chest pain as a typical angina or non-cardiac chest pain. Ignorance of blood pressure, sugar and cholesterol serum may lead towards the heart disease. Sometimes gastric pain and non-cardiac chest pain occurred as a false alarm for the heart disease identification. The accurate and proper prediction of heart disease may be performed using machine learning based on the patient historical data set with respect to the age.

B. Implementation of Classifiers for Heart Disease Classification

Table III explained that dataset has been collected from the database of UCI machine learning repository for the heart disease classification purpose. The data set has been used in many researches by using neural networks and ensemble classification [35]. The data set contained thirteen attributes to predict the heart disease. Output results were represented as two classes 1 or 2. Class "1" confirms the absence of heart disease and heart is working fine while class "2" indicates the high risk of heart attack as the heart disease has been found and it needs urgent consultation or treatment with the doctor or heart specialist. The Age and sex of patients have been mentioned in the column 1 and 2, respectively. Column no. 3 defines the type of the chest pain (CP) as the chest pain has four types. The chest pain types include typical angina, atypical angina, non-angina pain and asymptomatic. A heart pain or chest trouble caused by the muscles of heart due to the less oxygen in the blood is usually referred as typical angina pain. Atypical angina pain is a symbol of the problem that is not related to the heart actually. Non-angina pain is also acknowledged as non-cardiac chest pain (NCCP) that has the same feel like heart pain but that doesn't describe the heart disease. Asymptomatic shows that there was no heart disease detected. Column no. 4 shows the blood pressure of the patients in the rest condition. The attribute no. 5 describes the cholesterol serum level value in mg/dl. The sixth column explains that blood sugar (FBS) was measured in fasting of patients to identify the sugar greater than 120 mg/dl. The electrocardiographic results of patients in rest have been observed in the column no. 7 having the values of 0, 1 and 2. The maximum peak heart rate of patient was measured in the column no. 8. The column no. 9 shows the data of those patients who got induced angina pain due to the exercise. In column no. 10 old peak which is related to the ST depression achieved by exercise at rest position. Slope of the peak exercise has been mentioned in the column no. 11 (up, flat, down). Number of the main vessels (0-3) has been recorded in the column no. 12 that has been colored by the fluoroscopy. Thallium is the stress scintigraphy which elaborates that the heart rate is normal, fixed defect or reversible defect. The predicted results have been described in the 14th column having two classes. Class "1" shows that there was no heart disease identified and the heart is working properly. Class "2" indicates that the presence of heart disease has been confirmed therefore emergency consultation or treatment will be needed to cure it.

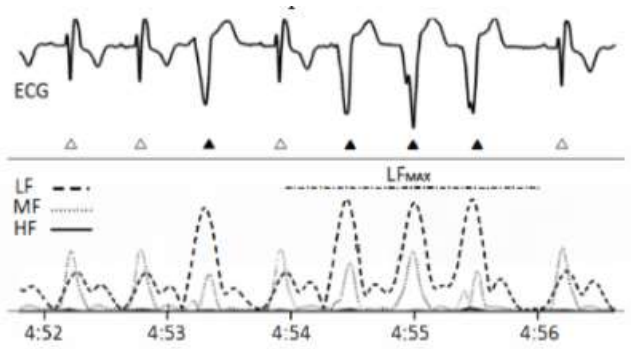


Fig. 12. Electrocardiogram Signals [21].

TABLE III. DATA SET FOR HEART DISEASE CLASSIFICATION [19]

Age	Sex	CP	RBP	SCHL	FBS	RECR	MHR	EIA	OP	SP	MV	TH	Output
67	0	3	15	564	0	2	60	0	1	62	0	7	1
57	1	3	24	261	0	0	41	0	0	31	0	7	2
67	1	2	28	263	0	0	5	1	0	22	1	7	1
74	0	4	20	269	0	2	21	1	0	21	1	3	1
65	1	2	20	77	0	0	40	0	0	41	0	7	1
56	1	4	30	256	0	2	42	1	0	62	1	6	2
59	1	3	10	239	0	2	42	1	1	22	1	7	2
60	1	4	40	293	0	2	70	0	1	22	2	7	2
63	0	4	50	407	0	2	54	0	4	42	3	7	2
59	1	4	35	234	0	0	61	0	0	52	0	7	1
53	1	4	42	226	0	2	11	1	0	22	0	7	1
44	1	4	40	235	0	2	80	0	0	40	0	3	1
61	1	3	34	234	0	0	45	0	2	62	2	3	2
57	0	1	28	303	0	2	59	0	0	59	1	3	1
71	0	4	12	49	0	0	25	0	1	62	0	3	1
46	1	4	40	311	0	0	20	1	1	82	2	7	2
53	1	4	40	203	0	2	55	1	3	13	0	7	2
64	1	4	10	211	0	2	44	1	1	82	0	3	1
40	1	4	40	99	0	0	78	1	1	41	0	7	1

C. Support Vector Machine (SVM)

Generally, Support Vector Machine (SVM) classifiers are applied to resolve complicated engineering problems of the real world; it has been observed in many classification applications that SVM performed better classification. A support vector mechanism produces a hyperplane or a series of hyperplanes in a high or infinite dimension area that can be used for classification, regression or detection of outliers. In the proposed research two hyper planes have been designed for the two classes.

H1 and H2 are the planes:

$$H1: w \bullet x_i + b = 1 \quad (6)$$

$$H2: w \bullet x_i + b = 2 \quad (7)$$

The plane H0 is the median in between, where $w \bullet xi + b = 0$

$$w^T x + b \geq 0 \text{ for } di = 1 \quad (8)$$

$$w^T x + b \geq 1 \text{ for } di = 2 \quad (9)$$

For the maximization of the margin, ||w|| can be minimized. Having the condition that there will be no data points between H1 and H2.

Non-Linear SVMs also used to separate the classes linearly by using the quadratic equation.

$$(x-a)(x-b) = x^2 - (a+b)x + ab \quad (10)$$

Optimization issue of the weight values can be resolved by using the following equations foe SVMs:

For the maximization;

$$\frac{1}{\|w\|} \quad (11)$$

Min. $| w^T x + b |= 0$ for n = 1,2,3……n

For the minimization;

$$\frac{1}{2} W^T.W \quad (12)$$

$y_n = | w^T x + b |= 0$ for n = 1,2,3……n

The preprocessed data was made ready for analysis. The data was filtered and the missing values were recovered. The data was given in the form of numbers and fractions and can be used for training.

Fig. 13 represents the proposed model data points of parameters. Proposed predictive model was trained with different algorithms and cores to verify performance. The method of teaching the algorithm also makes a big difference. If the wrong training data type is provided, the algorithm cannot achieve useful results.

Fig. 14 shows the Support Vector Machine confusion matrix. Commonly confusion matrix is seen diagonally; all the values in diagonal show the true positive classes. The confusion matrix shows the performance of the classifier. The confusion matrix of SVM elaborates that classification has been performed for the two classes. According to the confusion matrix 89% true positive rate with the 11% false positive rate have been predicted while classifying class 1 for the absence of heart disease. It can be easily observed that

68% true positive rate has been achieved with 32% false positive rate for the classification of class 2. 80.4% accuracy for SVM algorithm was experienced in the classification of heart disease. The training time for the proposed algorithm was found to be 0.75427 with the prediction speed of 5900 observations per second.

Fig. 15 illustrates SVM receiver operating characteristics (ROC). To evaluate the multi-class classifier performance, it must be visualized for the analysis. The area under the curve (AUC) determines the degree up to which scale it can classify. Receiver operating characteristics (ROC) is usually measured as the probability of classification. The graph of ROC is plotted between true positive rate and false positive rate. The area under the curve (AUC) was determined as 0.88 and it may be acknowledged as a competent classifier due to the closeness of AUC to 1.

D. K-NN Classifier

Fig. 16 shows the confusion matrix of K-NN algorithm. 76.1% accuracy has been achieved by the K-NN algorithm for the classification of heart disease. The efficiency was calculated using true positive rate and false positive rate. Confusion matrix evaluated the K-NN algorithm and showed that 85% class 1 was classified as true positive rate and false positive rate was achieved as 15% for class 1. For the class 2 classification using K-NN, 62% positive rate was achieved with the 38% of false positive rate.

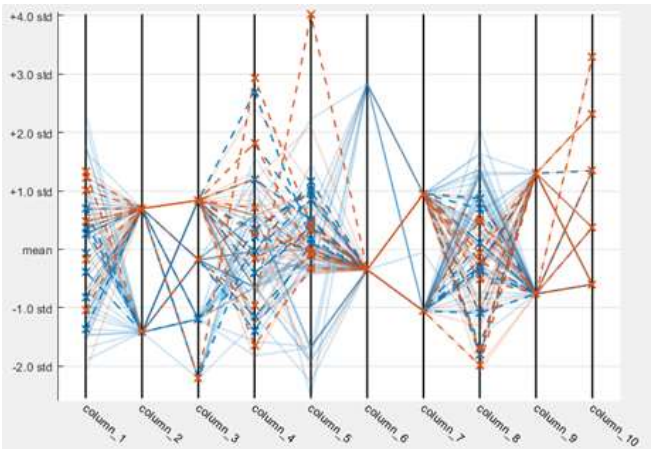


Fig. 13. Prediction Model.

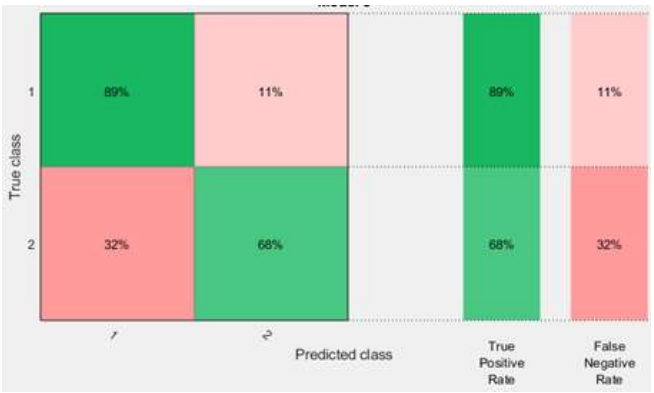


Fig. 14. SVM Confusion Matrix.

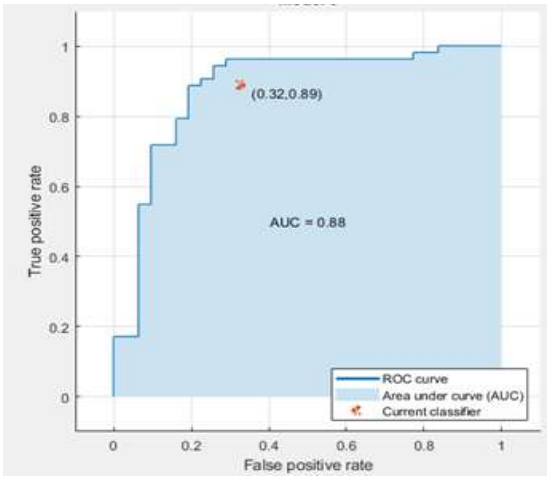


Fig. 15. Support Vector Machine ROC Curve.

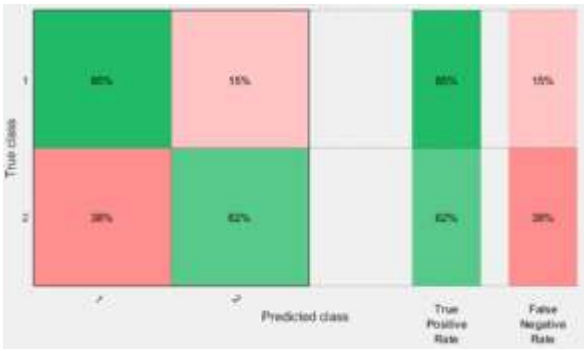


Fig. 16. K-NN Classifier Confusion Matrix.

Usually Euclidean distance is calculated to find out the closest distance with the value of the K.

$$d = \sqrt{(x1 - xA1)^2 + (x2 - xA2)^2} \quad (13)$$

Fig. 17 shows ROC and area under the curve (AUC) for the K-NN classifier. The ROC curve has been plotted between true positive rate and false positive rate. Area under the curve (AUC) was found to be 0.80. It is slightly away from the 1. For good classification AUC must be close to 1.

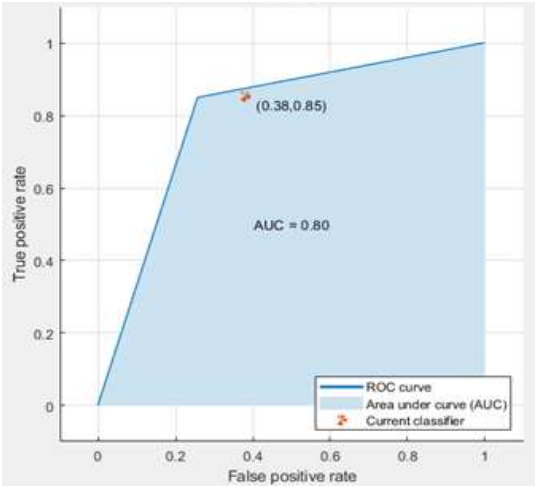


Fig. 17. K-NN ROC and AUC.

E. Linear Discriminant Analyss (LDA)

Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are the two most popular classifiers which are based on probabilistic method. For each class predictions can be easily computed using the following mathematical formulation of Baye's rule.

In Fig. 18, linear discriminant analysis (LDA) confusion matrix demonstrated that overall accuracy of 79.3% was achieved based on the true positive rate (TPR) and false positive rate (FPR). For the classification of heart disease, confusion matrix showed that 87% true positive rate was achieved in classifying class 1 while 13% of false positive rate was achieved in the classification of class 1.

Fig. 19 demonstrated that area under the curve (AUC) for the linear discriminant analysis was found to be 0.85.

F. Fine Gaussian SVM

Fig. 20 demonstrated the confusion matrix for fine Gaussian SVM. Confusion matrix of fine Gaussian SVM elaborated that the algorithm performed very poor as it classified all classes as class 1. Class 2 was not predicted at all therefore false negative rate was found to be 100% and false positive rate for class 2 was found to be 0%.

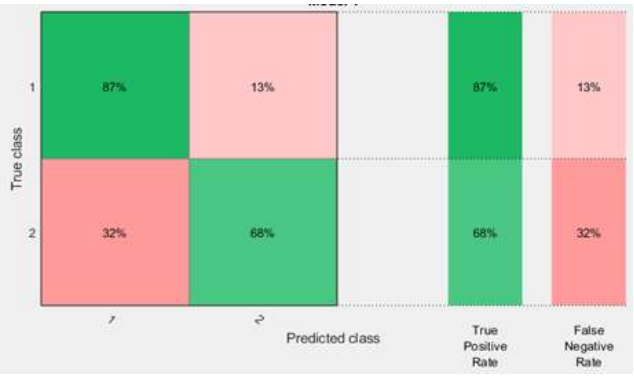


Fig. 18. LDA Confusion Matrix.

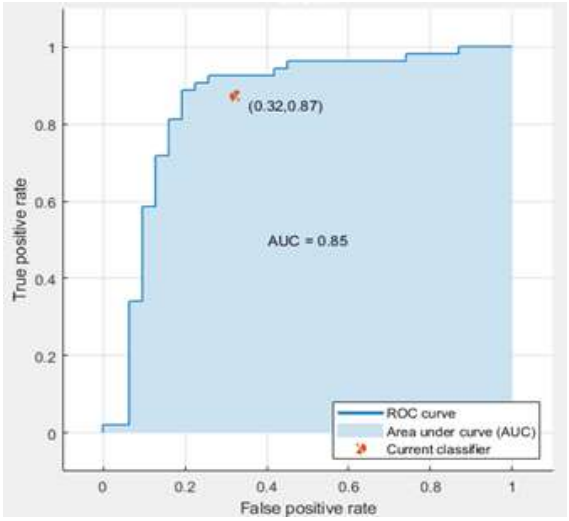


Fig. 19. ROC and AUC Curve for LDA.

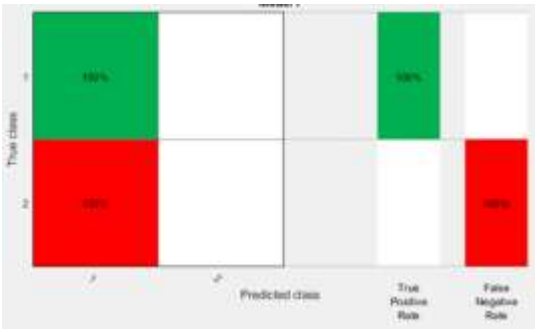


Fig. 20. Confusion Matrix of Fine Gaussian SVM.

G. Performance Comparison of Classifiers for Heart Disease Classification

Table IV proved that the SVM performed better classification compared to the K-NN and LDA in terms of Accuracy, prediction speed, training elapsed time, precision and area of under the curve. 80.4% accuracy was achieved by SVM for the heart disease classification which was greater than the accuracies of K-NN and LDA.

TABLE IV. SVM, KNN AND LDA COMPARISON

Parameters	SVM	K-NN	LDA	Fine Gaussian SVM
Accuracy	80.4%	76.1%	79.3%	59.8%
Prediction Speed	5900 obs/sec	1900 obs/sec	2100 obs/sec	6700 obs/sec
Training time	0.742347 seconds	1.2347 seconds	1.3924 seconds	0.8763 sec
Precision	0.89	0.80	0.83	0.59
AUC	0.88	0.80	0.85	0.60

IV. RESULTS AND CONCLUSION

Comparative study of classifiers was performed to determine the better classifier for the breast cancer prediction. It has been proved from the results that Logistic regression gained highest accuracy of 91.2%. K-NN also performed better with the accuracy of 88.25. Research study shows that logistic regression may be adopted on the real time data set of the patients to reduce the false alarm rate in the prediction of breast cancer tumors. Moreover, simulated results on real time data showed that SVM performed classification rapidly in very less time of 0.74237 seconds compared to the K-NN and LDA for heart disease classification. Prediction was observed 5900 observations per second which is higher than the LDA and K-NN classification algorithms. Accuracies and area under the curve of SVM were found to be 80.4% and 0.88 respectively. SVM proved to be a better and robust classifier for the heart disease classification.

REFERENCES

[1] E. Magnin, D. Vray and A. Brémond, "Early detection of breast cancer using computer assisted diagnosis," 1992 14th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Paris, 1992, pp. 849-850.

[2] R. D. Ghongade and D. G. Wakde, "Detection and classification of breast cancer from digital mammograms using RF and RF-ELM

algorithm," 2017 1st International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech), Kolkata, 2017, pp. 1-6. doi: 10.1109/IEMENTECH.2017.8076982.

[3] P. Král and L. Lenc, "LBP features for breast cancer detection," 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, 2016, pp. 2643-2647. doi: 10.1109/ICIP.2016.7532838.

[4] B. Hela, M. Hela, H. Kamel, B. Sana and M. Najla, "Breast cancer detection: A review on mammograms analysis techniques," 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13), Hammamet, 2013, pp. 1-6. doi: 10.1109/SSD.2013.6563999.

[5] R. Sangeetha and K. S. Murthy, "A novel approach for detection of breast cancer at an early stage using digital image processing techniques," 2017 International Conference on Inventive Systems and Control (ICISC), Coimbatore, 2017, pp. 1-4. doi: 10.1109/ICISC.2017.8068625.

[6] V. Krishnaiah, M. Srinivas, G. Narsimha and N. S. Chandra, "Diagnosis of heart disease patients using fuzzy classification technique," International Conference on Computing and Communication Technologies, Hyderabad, 2014, pp. 1-7.

[7] T. Botterill, T. Lotz, A. Kashif and J. G. Chase, "Reconstructing 3-D Skin Surface Motion for the DIET Breast Cancer Screening System," in IEEE Transactions on Medical Imaging, vol. 33, no. 5, pp. 1109-1118, May 2014. doi: 10.1109/TMI.2014.2304959.

[8] Q. Li et al., "Direct Extraction of Tumor Response Based on Ensemble Empirical Mode Decomposition for Image Reconstruction of Early Breast Cancer Detection by UWB," in IEEE Transactions on Biomedical Circuits and Systems, vol. 9, no. 5, pp. 710-724, Oct. 2015.doi: 10.1109/TBCAS.2015.2481940.

[9] S. C. Hagness, A. Taflove and J. E. Bridges, "Two-dimensional FDTD analysis of a pulsed microwave confocal system for breast cancer detection: fixed-focus and antenna-array sensors," in IEEE Transactions on Biomedical Engineering, vol. 45, no. 12, pp. 1470-1479, Dec. 1998. doi: 10.1109/10.730440.

[10] T. Kao et al., "Regional Admittivity Spectra with Tomosynthesis Images for Breast Cancer Detection: Preliminary Patient Study," in IEEE Transactions on Medical Imaging, vol. 27, no. 12, pp. 1762-1768, Dec. 2008. doi: 10.1109/TMI.2008.926049.

[11] D. A. Woten, J. Lusth and M. El-Shenawee, "Interpreting Artificial Neural Networks for Microwave Detection of Breast Cancer," in IEEE Microwave and Wireless Components Letters, vol. 17, no. 12, pp. 825-827, Dec. 2007. doi: 10.1109/LMWC.2007.910466.

[12] P. M. Meaney, M. W. Fanning, D. Li, S. P. Poplack, and K. D. Paulsen,"A clinical prototype for active microwave imaging of the breast,"IEEE Trans. Microw. Theory Tech., vol. 48, no. 11, pp. 1841–1853, Nov. 2000doi: 10.1109/MCS.2009.932223.

[13] E. J. Bond, X. Li, S. C. Hagness, and B. D. Van Veen, "Microwaveimaging via space-time beamforming for early detection of breastcancer," IEEE Trans. Antennas Propag., vol. 51, no. 8, pp. 1690–705, Aug. 2003.

[14] Wolberg, W.H., & Mangasarian, O.L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In Proceedings of the National Academy of Sciences, 87, 9193--9196.

[15] Zhang, J. (1992). Selecting typical instances in instance-based learning. In Proceedings of the Ninth International Machine Learning Conference (pp. 470--479). Aberdeen, Scotland: Morgan.

[16] Talha Khan, Muhammad Alam, Kushsairy Kadir, Zeeshan Shahid and M. S. Mazliham," Artificial Intelligence based Prediction of Seizures for Epileptic Patients: IoT Based Cost Effective Solution ", 2019 IEEE the 7th International Conference on Information and Communication Technology (IEEE-ICOICT), 24-26 July2019, Kuala Lumpur, Malaysia.

[17] Talha Khan, Muhammad Alam, Kushsairy Kadir, Sheroz Khan, M.S Mazliham, Faraz Shaikh, Syed Faiz Ahmed, Zeeshan Shahid" An Implementation of Electroencephalogram signals acquisition to control manipulator through Brain Computer Interface ", 2nd IEEE International Conference on Innovative research and development 2019 (IEEE-ICIRD), 24-26 July 2019, Kuala Lumpur, Malaysia.

[18] Talha Ahmed Khan, Muhammad Alam, Kushsairy Kadir, Zeeshan Shahid, MS. Mazliham, "False Alarm Reduction For The Cardiac Arryhthmias: AI Based Comparative Analysis", Journal of Engineering and Technology", Universiti Kuala lumpur Journal of Engineering and Technology, Vol. 5 (2017).

[19] K. D. Kochanek, J. Xu, S. L. Murphy, A. M. Miniño, and H.-C. Kung, "Deaths: final data for 2009," National Vital StatisticsReports, vol.60, no.3, pp.1–116,2011.

[20] N. Gawande and A. Barhatte, "Heart diseases classification using convolutional neural network," 2017 2nd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, 2017, pp. 17-20. doi: 10.1109/CESYS.2017.8321264.

[21] G. Meena, P. S. Chauhan and R. R. Choudhary, "Empirical Study on Classification of Heart Disease Dataset-its Prediction and Mining," 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), Mysore, 2017, pp. 1041-1043. [4] S. Ekız and P. Erdoğmuş, "Comparative study of heart disease classification," 2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, 2017, pp. 1-4. doi: 10.1109/EBBT.2017.7956761.

[22] V. Krishnaiah, M. Srinivas, G. Narsimha and N. S. Chandra, "Diagnosis of heart disease patients using fuzzy classification technique," International Conference on Computing and Communication Technologies, Hyderabad, 2014, pp. 1-7.

[23] Talha Ahmed Khan, Muhammad Alam , M. Junaid Tahir, Kushsairy Kadir , Zeeshan Shahid, M.S Mazliham, "Optimized health parameters using PSO: a cost effective RFID based wearable gadget with less false alarm rate" Indonesian Journal of Electrical Engineering and Computer Science, Vol. 15, No. 1, July 2019, pp. 230~239, ISSN: 2502-4752, DOI: 10.11591/ijeecs.v15.i1.pp230-239.

[24] Y. Özbay, B. Karlik, "A recognition of ECG arrhythmias using artificial neural networks," Proceedings of the 23rd Annual Conference, IEEE/EMBS, Istanbul, Turkey, 2001.

[25] A. Kampouraki, G. Manis, C. Nikou, "Heartbeat time series classification with support vector machines," IEEE Transactions on Information Technology in Biomedicine, vol. 13, no. 4, 2009.

[26] G. Evensen, "The ensemble Kalman filter for combined state and parameter estimation," in IEEE Control Systems, vol. 29, no. 3, pp. 83-104, June 2009. doi: 10.1109/MCS.2009.932223.

[27] F. Plesinger, P. Klimes, J. Halamek and P. Jurak, "False alarms in intensive care unit monitors: Detection of life-threatening arrhythmias using elementary algebra, descriptive statistics and fuzzy logic,"2015 Computing in Cardiology Conference (CinC), Nice, 2015, pp. 281-284.2009. doi: 10.1109/MCS.2009.932223.

[28] J. Pan, W.J. Tompkins, "A real-time QRS detection algorithm," IEEE Transactions on Biomedical Engineering, vol. 32, no. 3, 1985.

[29] R. Acharya, A. Kumar, P. S. Bhat, C.M. Lim, S.S. Iyengar, N. Kannathal, S.M. Krishnan, "Classification of cardiac abnormalities using heart rate signals," Med. Biol. Eng. Comput., vol. 42, pp. 288–293, 2004.

[30] M. M. Engin, "ECG beat classification using neuro-fuzzy network," Elsevier, Pattern Recognition Letters, vol. 25, pp. 1715–1722, 2004.

[31] J. P. Kelwade and S. S. Salankar, "Radial basis function neural network for prediction of cardiac arrhythmias based on heart rate time series,"2016 IEEE First International Conference on Control, Measurement and Instrumentation (CMI), Kolkata, 2016, pp. 454-458.

[32] A. M. Amiri and G. Armano, "Early diagnosis of heart disease using classification and regression trees," The 2013 International Joint Conference on Neural Networks (IJCNN), Dallas, TX, 2013, pp. 1-4.

[33] J. P. Kelwade and S. S. Salankar, "An optimal structure of multilayer perceptron using particle swarm optimization for the prediction of cardiac arrhythmias,"2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, 2016, pp. 426-430.doi: 10.1109/ICRITO.2016.7784993.

[34] P. Majety and V. Umamaheshwari, "An electronic system to recognize heart diseases based on heart sounds: A stochasatic alogorithm implemented on DSK6713," 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, 2016, pp. 1617-1621.

[35] Gavin Brown. Diversity in Neural Network Ensembles. The University of Birmingham. 2004.

Educational Data Mining for Monitoring and Improving Academic Performance at University Levels

Ezekiel U Okike[1], Merapelo Mogorosi[2]
Department of Computer Science
University of Botswana, Gaborone, Botswana

***Abstract*—This study applied Educational Data Mining on 712 sample of logs extracted from Moodle Learning Management System (LMS) at an African University in order to measure students and staff patterns of use of the LMS resources and hence determine if the quantity of participation measured in the amount of time spent on the use of LMS resources improved academic performance of students. Data collected from Moodle LMS was preprocessed and analyzed using machine learning algorithms of clustering, classification and visualization from WEKA system tools. The dataset consisted of Course tools (Quiz, Assignment, Chat, Forum, URL, Folder and Files), Lecturer and Student usage of the tools. Furthermore, SPSS was used to obtain a matrix for coefficients of correlations for course tools, tests and final grade. Correlation analysis was done to verify if students use of course tools had impact on student's academic performance. Findings indicated the pattern of usage for course1 as Quiz (38358), System (17910), Forum (8663), File (8566), Assignment (1235), Folder (514, File Submission (172), and Chat (37); Course2 as System (11920), Quiz (8208), Forum (4476), File (4394), Assignment (257), Chat (247), URL (125), and File Submission (38); Course3 as System (2622),File (1022), Folder (570), Forum (258), and URL (2). Overall, evaluating the correlation between the use of LMS resources and students' performance, findings indicated there is significant relationship between the use of LMS resources and students' academic performance at 0.01 level of significant. The findings are useful for strategic academic planning purpose with LMS data at the university.**

Keywords*—Educational data mining; learning management systems; Weka system tools; improved academic performance

I. Introduction

Data Mining (DM) and Machine Learning (ML) as sub-disciplines of computer science provide powerful tools for knowledge discovery from massive data sets [1,2]. As a process of discovering patterns in data a DM process must be automatic or semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage [3]. The concept of Educational Data mining (EDM) is defined in this paper as the application of data mining to derive meaningful patterns from educational system repositories which in turn could be used to improve teaching and learning experiences. One such educational teaching and learning system tool is the learning management system (LMS). LMS tools are computerized teaching and learning platforms for creating contents, delivering courses to learners and managing courses in teaching and learning environments [4,5].

A. Statement of the Problem

Many African universities have invested in learning management systems. Over the years, massive data have also been accumulated through the LMS, but which has not been appropriately mined in order to provide information useful for strategic decisions at university levels. The aim of this paper is to demonstrate with empirical evidence the usefulness of EDM in discovering hidden but useful patterns from teaching and learning data accumulated through the LMS. Several studies for example [4, 5, 6, 7, 8, 9] have also shown the central place of LMSs in teaching and learning, although, only a few addressed the need for EDM from logs of LMS's data. The specific objectives of this study were:

The specific objectives of this study were:

1) To measure students and staff use of LMS resources in teaching and learning at university level.

2) To evaluate the correlation between potential use of LMS resources and students' performances.

B. Research Questions

The following research questions were investigated in the study:

1) What were the recognizable patterns in teaching and learning from Moodle LMS logs for the sampled data set?

2) Is there any correlation between potential use of LMS and prevailing students' performance at the university?

The rest of this paper counts of three sections, and a list of references. Section two presents a review of literature relevant to the study. Section three presents the empirical study and the methodology. Section four presents the results and discussion of the study.

II. Literature Review

The application of Learning management systems and their effectiveness in higher education have been widely discussed in the literature [6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17]. A common approach to EDM modeling is a combination of sequential processes which includes data collection from LMSs, data preprocessing, data mining and analysis, preprocessing, result generation, and application (see Fig. 1).

A successful educational data mining (EDM) which produced satisfactory results at the University of Cordoba, Spain was reported in [9]. The EDM approach enabled the discovery of new rules of association which were used to improve the design of online courses at the university.

Similarly, strategies to mine data from activity logs found on Moodle LMS was investigated in [6]. Applying data mining and using simple statistics to analyze the logs the author recommended the use of Access watch analogue and Web start applications to infer student's attitudes to learning and for predicting examination scores through multiple regressions. Fig. 2 shows a data mining model using LMS.

Following a data cycle approach, Fig. 3 demonstrates the usefulness of data mining in the context of this paper. Every decision-making process is based on a data transformed into information culminating in a decision being made. The cycle begins with identifying the problem, collecting and storing data using appropriate tools, preprocessing data, mining data, and discovering new knowledge from data which provides necessary feedback from the system for future activities. Fig. 3 portrays the Data Cycle in knowledge discovery using data mining.

A brief explanation of each stage of the DCKD include:

1) Problem definition: An initial definition of the problem, or the mission, or the purpose, for which data is required.

2) Identifying data sources: Understanding what data are pertinent, and where they can be located.

3) Data collection and storing: Retrieval of data from various sources and storing them in an accessible location.

4) Data mining: Selection of relevant data out of the Big Data using appropriate DM and ML tools.

5) Knowledge discovery: New knowledge discovered and presented to decision makers (Classified, Clustered and Visualized).

6) Learning and decision-making: The final stage that is the purpose of the data cycle. The results are displayed to the decision makers and decisions are taken.

7) Feedback for further cycles: This stage is not always necessary. However, very often, the need to make a certain decision is repetitive, so the customer (the decision maker) can affect the usefulness and the effectiveness of the cycle by forwarding comments and changes.

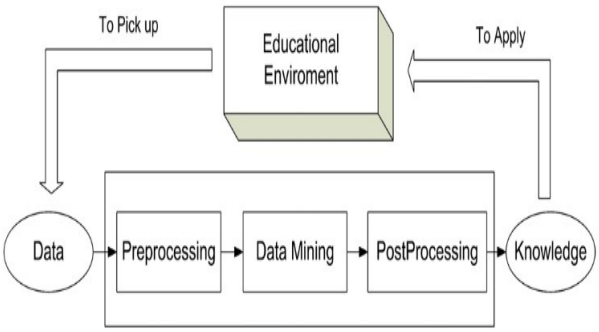


Fig. 1. EDM Processes Source: (García et al., 2011).

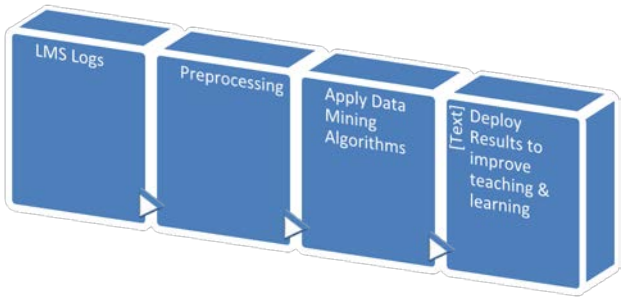


Fig. 2. Mining LMS Data.

Fig. 3. Data Cycle in Knowledge Discovery (DCKD) using Data Mining.

III. Empirical Study

A. Methodology

A sample of 712 Moodle data logs accumulated using Moodle Learning Management System were used in the study. The sample comprised three courses selected randomly from three faculties. The logs comprised of students, academic staff, and course data in the format: user full name, description, time, components, affected user, event context, event name, origin and IP address. The following criteria were observed in the sample selection:

1) Permission was obtained to conduct the study from which the data was obtained.

2) The experience of a lecturer offering a course on Moodle LMS was considered.

3) A selected course was offered in the first semester.

4) A selected course was taught by a lecturer who showed enthusiasm in online activities.

B. Data Collection Instrument

The logs were downloaded as .csv files, prepared and preprocessed using a Waikato Environment for Knowledge Analysis (WEKA) tool.

C. Data Preparation and Preprocessing

The preparation and preprocessing consisted of data extraction, cleaning, aggregation/ integration, filtering and transformation. Data was mined using machine learning schemes selected from the WEKA tool. The selected schemes included tools for pre-processing, classification, clustering and visualization. The algorithms were applied directly to each dataset for each stage of preprocessing, classification, clustering, and visualization invoked from the schemes menu as follows:

1) A data file is selected from the file menu.
2) Important attributes of the data are selected.
3) Aggregates of existing attributes were created using the spreadsheet.
4) A machine learning scheme was selected from the Schemes menu.
5) Results were viewed as trees, text or three- dimensional plots.
6) Attribute/aggregate selections were revised.
7) The scheme was re-run on the revised data.

Furthermore, in order to maintain format independence, the data was converted to an intermediate representation - Attribute Relation File Format (ARFF). Data logs of semester1 of the academic year 2017/18 were downloaded, extracted and used for the study. The extraction process is shown in Fig. 4.

Data filtering was done through the WEKA system filtering tool targeting the attributes required for use in the selected machine learning algorithms ZeroR and J48.

D. Data Mining and Analysis using Machine Learning Algorithms ZERO R and J48

J48 and ZeroR machine learning algorithms were applied on the data sets. The justification for the choice of J48 and ZeroR algorithms in this study was due to their perceived usefulness in classification, clustering and visualization [18, 19] J48 algorithm uses unsupervised learning to form clusters of a data set, while ZeroR algorithm uses supervised learning. Both classifiers differentiate each case according to some set criteria.

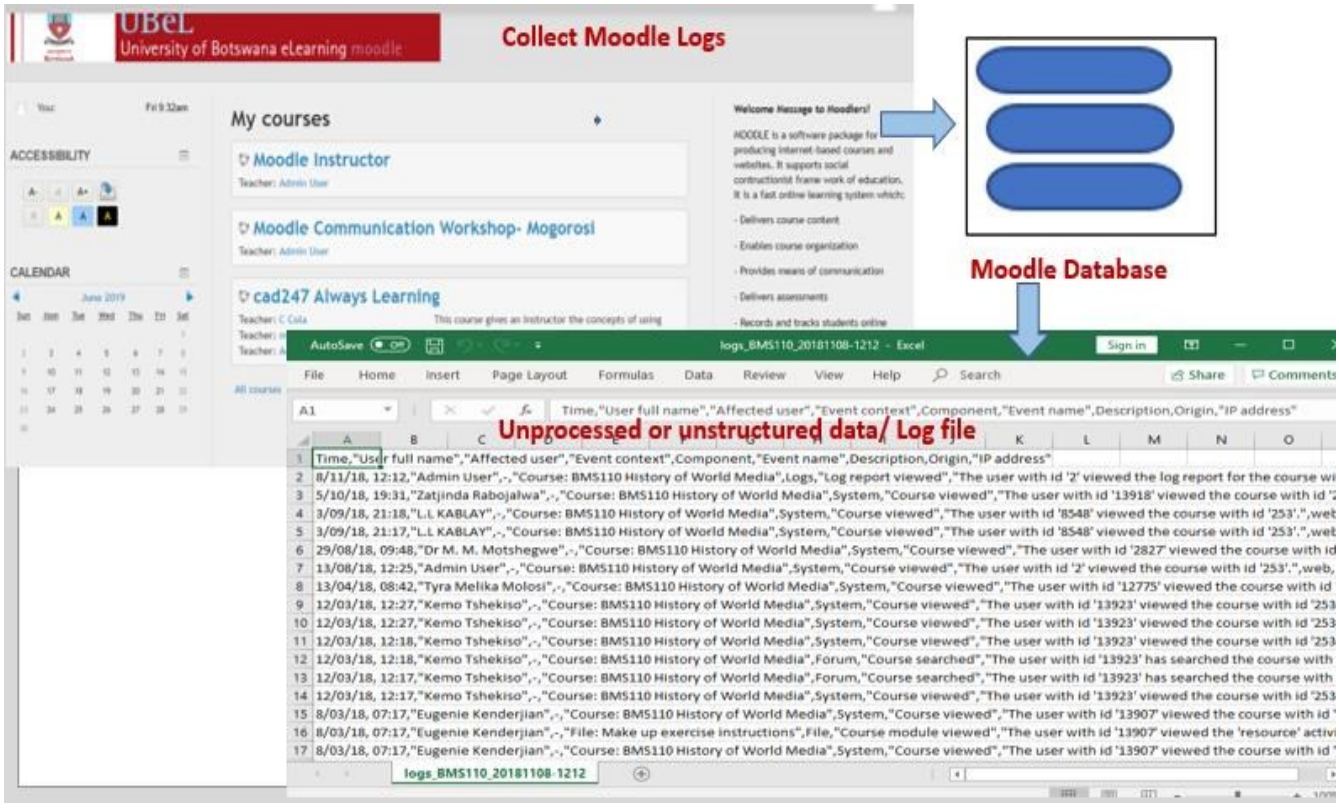


Fig. 4. Extracting Data from Moodle LMS for Data Mining.

V. Results and Discussion

A. Classification with ZeroR and J48 Algorithms

Fig. 5 shows the results of using ZeroR to predict Course1. The prediction class value was Quiz which took 0.09 seconds to build the model. The number of correctly classified instances was 38530 (50.553%) while the number of incorrectly classified instances was 37687 (49.447%).

The mean absolute error and the root mean squared error for these predictions were 0.147 and 0.2711 respectively. There were five attributes (User Full Name, Event context, Components, Event Name and Origin) and the focus was on the Component attribute which explains how the course activities and resources were used during the semester. The accessibility of the logs of the course by a lecturer and the students was 76217 which is shown as the total number of instances. The number of correctly classified instances was 38530 (50.556%) and incorrectly classified instances was 37687 (49.4%). It should be noted that when an algorithm is based on probability, there is the risk of type 1errors (false positives) and type 2 error (false negatives). This accounts for the noticeable mean absolute error, and the root mean squared error above. The detailed accuracy by class is shown below in the confusion matrix (Table I) which shows the activity tools with their accessibility values.

Table II, Table III, and Fig. 6 show the results of predicting activities in course1 using J48 algorithm.

From Table II the statistical summary of the results shows 75463 (99%) as number of correctly classified instances and 754 (0.98%) as incorrectly classified instances. The mean absolute error is 0.0028, root mean squared error is 0.039, and relative absolute error is 1.91%. The errors are minimal. The detailed accuracy by class is shown in the confusion matrix (Table III).

As a machine learning algorithm, J48 uses the decision tree technique. When applied to the LMS logs on course activities and resources, the pattern of activities classification shows that Quiz (column c) has the highest activity at 38358, followed by System (column b) at 17910, Forum (column d) at 8663, File (column f) at 8566, and Assignment (Column e) at 1238, Folder (column g) at 514, File submission (column h) at172 and chat (column i) at 37.

Table IV, and Table V show the result of prediction of course 2 and its confusion matrix in classification using J48 algorithm.

From Table V the number of correctly classified instances is 29765 (99%), while the number of incorrectly classified instances is 276 (0.91%). The mean absolute error was 0.0026, root mean squared error was 0.0387, and relative absolute error was 1.84%. The errors are minimal. The detailed accuracy by class is shown below in Table VI.

Table V above shows the confusion matrix of all classes. The pattern shows that System column b) has the highest activity at 11920, followed by Quiz (column c) at 8308, Forum (column d) at 4476, File (column g) at 4398, Assignment (column f) at 257, Chat (column e) at 247, URL (column h) at 125 and File submission (column i) at 38. Similarly, for ZeroR the prediction of course2 in classification is shown in Fig. 6, Table VI, and Table VII. The same explanation offered in J48 upholds except that the prediction identified the system as predictor as shown in Fig. 6 and Table VI.

The results of ZeroR predicted class value were Quiz (38530), System (18108), Forum (8802), File (8566), Assignment (1395), Folder (514), File submission (172), and Chat (122). ZeroR algorithm predicted Quiz tool to be the highest. This explains the pattern of how the lecturer taught the class. The access of the course (System:18108) led learners to use the quiz tool (38530) more often because they had to discuss (Forum 8802) the concept of the given topic and read notes (File:8566). The learners also had to work on the given assignments (1395) for the lecturer to check whether they had understood the concepts of the topic and submit (172) back their assignments for marking. Detailed accuracy by class is shown in Table VII, while the confusion matrix is shown in Table VIII.

The report on Confusion Matrix above (Table VIII) shows the activity tools with their accessibility values. The results of ZeroR predicted class value as System (12109), Quiz (8308), Forum (4485), File (4394), Assignment (309), URL (125), Chat (125) and File submission (38). ZeroR algorithm predicted System tool to be the highest followed by Quiz, Forum, File, Assignment, URL, and Chat. This explains the pattern of how the lecturer taught the class. The learners accessed the course (System:12109) more often which led them to use the quiz tool . The Table XI above and Table XII shows the results of using ZeroR to predict Course3. The figure has five attributes (User Full Name, Event context, Components, Event Name and Origin) and the focus was on the Component attribute which explains how the course activities and resources were used during the semester. The prediction class value was System which took 0 seconds to build the model. The number of correctly classified instances was 2622 (58.5007%) while the number of incorrectly classified instances was 1860 (41.4993%). The accessibility or the logs of the course by the lecturer and the students was 4482 which is shown as the total number of instances. The ZeroR algorithm is based on probability, hence the likely presence of errors. The Mean Absolute and Root Mean Squared Errors are minimal in the range of 0.1676 and 0.2894 respectively.

Table XII presents a detailed accuracy by class and Confusion Matrix in classification for ZeroR

From Table IX the results above have 5 attributes and focus was on component attribute which explains how the course activities and resources were used during the semester. The course activities were System (2622),

File (1022), Folder (570), Forum (258) and URL (2). This suggests that the Lecturer used the System to create Files, Forum and Folder for the course. Folders were probably used to add notes for students. Files might have been used to create teaching materials like notes while Forum might have been used to create discussion topics.

```
Scheme:       weka.classifiers.rules.ZeroR
Relation:     bms110_sem1_2015-16-weka.filters.unsupervised.attribute.Remove-R1,3,8
Instances:    76217
Attributes:   5
              User full name
              Event context
              Component
              Event name
              Origin
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: Quiz

Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       38530               50.553  %
Incorrectly Classified Instances     37687               49.447  %
Kappa statistic                          0
Mean absolute error                      0.147
Root mean squared error                  0.2711
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances            76217
```

Fig. 5. Prediction of Course1 in Classification – ZeroR.

TABLE I. CONFUSION MATRIX- COURSE1

A	B	C	D	E	F	G	H	I	Classified as
0	0	8	0	0	0	0	0	0	A = Logs
0	0	18108	0	0	0	0	0	0	B=System
0	0	38530	0	0	0	0	0	0	C=Quiz
0	0	8802	0	0	0	0	0	0	D=Forum
0	0	1395	0	0	0	0	0	0	E=Assignment
0	0	8566	0	0	0	0	0	0	F=File
0	0	514	0	0	0	0	0	0	G=Folder
0	0	172	0	0	0	0	0	0	H=File submission
0	0	122	0	0	0	0	0	0	I=Chat

TABLE II. PREDICTION OF COURSE 1 IN CLASSIFICATION USING J48 ALGORITHM

Item	**Value**	
Correctly classified instances	75463	99.0107%
Incorrectly classified instances	754	0.9893 %
Kappa statistic	0.9851	
Mean absolute error	0.0028	
Root mean squared error	0.039	
Relative absolute error	1.9191%	
Root Relative squared error	14.3847%	
Total number of instances	76217	

TABLE III. CONFUSION MATRIX COURSE 1 USING J48 ALGORITHM

A	B	C	D	E	F	G	H	I	Classified as
8	0	0	0	0	0	0	0	0	A = Logs
0	17910	47	111	23	15	2	0	0	B=System
0	46	38358	100	26	0	0	0	0	C=Quiz
0	52	36	8663	35	16	0	0	0	D=Forum
0	30	42	88	1235	0	0	0	0	E=Assignment
0	0	0	0	0	8566	0	0	0	F=File
0	0	0	0	0	0	514	0	0	G=Folder
0	0	0	0	0	0	0	172	0	H=File submission
0	10	0	2	1	72	0	0	37	I=Chat

```
=== Run information ===

Scheme:       weka.classifiers.rules.ZeroR
Relation:     bns201_seml_2014-15-weka.filters.unsupervised.attribute.Remove-R1,3,8
Instances:    30041
Attributes:   5
              User full name
              Event context
              Component
              Event name
              Origin
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: System

Time taken to build model: 0.06 seconds
```

Fig. 6. Prediction of Course 2 in Classification using ZeroR.

TABLE IV. PREDICTION OF COURSE 2 IN CLASSIFICATION USING J48 ALGORITHM

Item	Value	
Correctly classified instances	29765	98.0813%
Incorrectly classified instances	276	0.9187 %
Kappa statistic	0.9872	
Mean absolute error	0.0026	
Root mean squared error	0.0387	
Relative absolute error	1.8404%	
Root Relative squared error	14.46437%	
Total number of instances	30041	

TABLE VI. CONFUSION MATRIX COURSE2 – J48

A	B	C	D	E	F	G	H	I	J	Classified as
0	2	0	0	0	0	0	0	0	0	A = Logs
0	11920	115	41	6	4	22	1	0	0	B=System
0	0	8308	0	0	0	0	0	0	0	C=Quiz
0	1	8	4476	0	0	0	0	0	0	D=Forum
0	0	4	0	247	0	0	0	0	0	E=Chat
0	0	0	0	0	257	0	0	0	0	F=Assignment
0	0	0	0	0	0	4394	0	0	0	G=File
0	0	0	0	0	0	0	125	0	0	H=URL
0	0	0	0	0	0	0	0	38	0	I=File submission
0	1	0	0	0	0	0	0	0	0	J=Activity report

TABLE VII. PREDICTION OF COURSE 2 IN CLASSIFICATION USING ZEROR ALGORITHM

Item	**Value**	
Correctly classified instances	12109	40.30807%
Incorrectly classified instances	17932	59.6918 %
Kappa statistic	0	
Mean absolute error	0.1434	
Root mean squared error	0.2678	
Relative absolute error	100%	
Root Relative squared error	100%	
Total number of instances	30041	

TABLE VIII. CONFUSION MATRIX- COURSE2

A	B	C	D	E	F	G	H	I	J	Classified as
0	2	0	0	0	0	0	0	0	0	A = Logs
0	12109	0	0	0	0	0	0	0	0	B=System
0	8308	0	0	0	0	0	0	0	0	C=Quiz
0	4485	0	0	0	0	0	0	0	0	D=Forum
0	270	0	0	0	0	0	0	0	0	E=Chat
0	309	0	0	0	0	0	0	0	0	F=Assignment
0	4394	0	0	0	0	0	0	0	0	G=File
0	125	0	0	0	0	0	0	0	0	H=URL
0	38	0	0	0	0	0	0	0	0	I=File submission
0	1	0	0	0	0	0	0	0	0	J=Activity report

TABLE IX. PREDICTION OF COURSE 3 IN CLASSIFICATION USING ZEROR ALGORITHM

Item	**Value**	
Correctly classified instances	2622	58.5007%
Incorrectly classified instances	1860	41.4993 %
Kappa statistic	0	
Mean absolute error	0.1676	
Root mean squared error	0.2894	
Relative absolute error	100%	
Root Relative squared error	100%	
Total number of instances	4482	

TABLE X. CONFUSION MATRIX- COURSE2

A	B	C	D	E	F	G	H	I	J	Classified as
0	6	0	0	0	0	0	0	0	0	A = Logs
0	2622	0	0	0	0	0	0	0	0	B=System
0	1022	0	0	0	0	0	0	0	0	C=File
0	258	0	0	0	0	0	0	0	0	D=Forum
0	570	0	0	0	0	0	0	0	0	E=Folder
0	2	0	0	0	0	0	0	0	0	F=Activity report
0	2	0	0	0	0	0	0	0	0	G=URL

Similarly, Fig. 7 shows the J48 pruned tree and the event context is course3. The summary of classification (Table X) shows correctly classified instances as 4482 (100%). The root mean absolute error is 0, the root mean squared error is 0, and the relative absolute error is also 0. Table XI shows the confusion matrix where the leading diagonal displays numbers which represent interaction with the system and the total of these numbers represent the logging behavior of students and the lecturer(s). Therefore, the algorithm J48 predicts System as the major classifier.

Fig. 7 shows prediction of course3 activities in classification using J48 with 5 attributes. Table X shows the summary of the results reported correctly classified instances as 4482 (100%), and 0 (0%) as incorrectly classified instances. The mean absolute error was 0, root mean squared error was 0, and relative absolute error was 0%. The errors are minimal.

Table XI shows 7 classes (Logs, System, File, Forum, Folder, Activity report and URL) in the confusion matrix for Course3. The pattern shows that System (column b) has the highest activity at 2622, followed by File (column c) at 1022, Folder (column e) at 570, Forum (column d) at 258 and, URL (column g) at 2.

B. Pattern of Usage Discuss

1) Research question 1: With reference to research question1, the data analysis revealed a mixed picture. There is a substantial use of the quiz tool (about 70%) for assessment purposes such as gauging the level of achievement of instructional objectives. It is worth noting that another facility which was also used substantially was the Resources tools (File, Folder and URL) which were mainly used for posting notes and communication between lecturer and students. They can be considered as an entry point for lecturers to make teaching and learning digital. In all, it appears that the pattern of usage identified above are complemented by blended learning where both traditional and technology-based approaches are mixed depending on the instructional goals.

2) Research question 2: With reference to research question 2 patterns were already revealed in the results presented. It was clearly observed that student's login into their Moodle portal to check for new course content (system use). Usually, students would have been alerted by the lecturer on the existence of new contents. Students could download the course content into their personal devices to read and to discuss amongst themselves. However, they could also do limited discussions using forums created by the lecturer.. Students might spend minimal time online when the course contents were not engaging. Furthermore, students might have exhibited, "the student syndrome", where there was a rush to do assignments just before the due date. These activities were discernible from the students' activity logs and their dates of submissions.

On the side of lecturers, the typical pattern of usage suggested, posting of notes using the resources tool and creating forums. In addition, it was observed that lecturers made good use of quizzes, assignments, chats and URLs.

In terms of correlation between potential use of LMS and prevailing students' performance (Research question 2).

Fig. 8 and Table VII, Table XIII and Table XIV show student's interaction with LMS and performance at final exam.

Fig. 8 shows that Files, Quiz, Test1, Test2 and Final Exams contributed to final grade. The correlation is explained in Tables XII, XIII and XIV for course 1, course 2 and course 3 respectively. The tables only reflect correlating components at 0.01 level of significance.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     bns305_seml_2014-15-weka.filters.unsupervised.attribute.Remove-R1,3,8
Instances:    4482
Attributes:   5
              User full name
              Event context
              Component
              Event name
              Origin
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===
```

Fig. 7. Prediction of Course3 in Classification – J48.

TABLE XI. SUMMARY OF CLASSIFICATION

Item	**Value**	
Correctly classified instances	4482	100%
Incorrectly classified instances	0	0 %
Kappa statistic	1	
Mean absolute error	0	
Root mean squared error	0	
Relative absolute error	0 %	
Root Relative squared error	0%	
Total number of instances	4482	

TABLE XII. CONFUSION MATRIX USING J48- COURSE3

A	B	C	D	E	F	G	Classified as
6	0	0	0	0	0	0	A = Logs
0	2622	0	0	0	0	0	B=System
0	0	1022	0	0	0	0	C=File
0	0	0	258	0	0	0	D=Forum
0	0	0	0	570	0	0	E=Folder
0	0	0	0	0	2	0	F=Activity report
0	0	0	0	0	0	2	G=URL

TABLE XIII. SUMMARY OF SIGNIFICANT CORRELATIONS IN COURSE 1

Tool	Files	Folder	Quiz	Final Exam	Test1	Test2	Final Grade
Files		.390	.402				
Folder	.390		.869	.180			.171
Quiz	.402	.869		.207	.122	.133	.202
Final Exam		.180	.207			.452	.798
Test1				.417		.334	.557
Test2				.452	.334		.627
Final Grade		.171	.202	.798	.557	.627	

Correlation is significant at the 0.01 level (2-tailed)

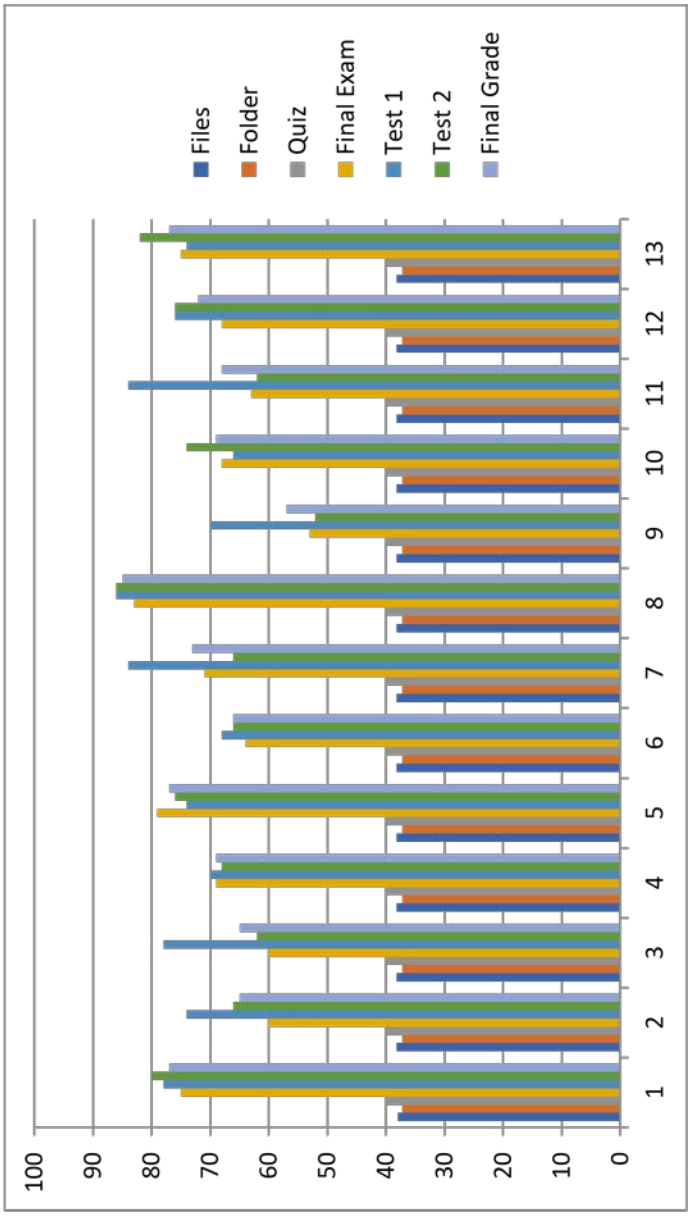


Fig. 8. Data Analysis based on Correlation Matrix of Course Tools and Students Academic Performance.

TABLE XIV. SUMMARY OF SIGNIFICANT CORRELATIONS IN COURSE 2

Tool	Files	Folder	Quiz	Assignment	Forum	Final Exam	Test1	Test2	Final Grade
Files		0.342	0.607	0.675	0.540				0.232
Folder	0.342		0.859	0.816	-0.726	0.288		0.200	0.372
Quiz	0.607	0.859		0.901	-0.786	0.254			0.388
Assignment	0.675	0.816	0.901		-0.885	0.302	0.212	0.241	0.438
Forum	-0.540	-0.726	-0.786	-0.885		-0.281	-0.226	-0.224	0.419
Final Exam		0.288	0.254	-.302	-0.281		0.465	0.563	0.761
Test1				0.212	-0.226	0.465		0.416	0.582
Test2		0.200		0.241	-0.224	0.563	0.416		0.564
Final Grade	0.232	0.372	0.388	0.438	-0.419	0.731	0.582	0.564	

Correlation is significant at the 0.01 level (2-tailed)

TABLE XV. SUMMARY OF SIGNIFICANT CORRELATION OF STUDENTS PERFORMANCE – COURSE3

Tool	Files	Folder	Quiz	Assignment	Forum	Final Exam	Test1	Test2	Final Grade
Files		0.394	0.406	0.540	0.513				0.144
Folder	0.394		0.869	0.858	-0.671	0.184		0.145	0.276
Quiz	0.406	0.869		0.847	-0.724	0.209	0.144	0.156	0.286
Assignment	0.540	0.858	0.847		-0.742	0.216	0.148	0.174	0.298
Forum	-0.513	-0.671	-0.727	-0.742		-0.232	-0.211	-0.202	-0.331
Final Exam		0.184	0.209	0.216	-0.232		0.414	0.432	0.731

Correlation is significant at the 0.01 level (2-tailed)

C. Analysis of Correlations

In Course1 (Table XII) the correlation between students' interactions with LMS and their performance were observed as follows:

There is significant relationship at 0.01 level between Files and Folder (0.390), Files and Quiz (0.402); Quiz and Files (0.402), Quiz and Folder (0.869), Quiz and Final Exam (0.207), Quiz and Final Grade (0.202); Final Exam and Folder (0.180), Final Exam and Quiz (0.207), Final Exam and Test1 (0.417); Final Exam and Test2 (0.452), Final Exam and Final Grade (0.709); Test1 and Final Exam (0.417), Test1 and Test2 (0.334), Test1 and Final Grade (0.557); Test2 and Final Exam (0.452), Test2 and Test1 (0.334), Test2 and Final Grade (0.627); Final Grade and Folder (0,171), Final Grade and Quiz (0.202); Final Grade and Final Exam (0.798), Final Grade and Test1 (0.557), Final Grade and Test2 (0.627); Therefore, the various components influenced final grade. Hence, there is significant relationship between use of LMS tools and students' performance. Since Pearson Coefficient value is greater than 0.01, we accept null hypothesis (research question 2).

In course 2 (Table XIII). There is significant relationship at 0.01 level between Files and Folder (0.342), Files and Quiz (0.607), Files and Assignment (0.675), Files and Forum (-0.540) and Files and Final Grade (0.232); Folder and Files (0.342), Folder and Quiz (0.859), Folder and Assignment (0.816), Folder and Forum (-0.726), Folder and Final Exam (0.288), Folder and Test2 (0.200) and Folder and Final Grade (0.372); Quiz and Files (0.607), Quiz and Folder (0.859), Quiz and Assignment (0.901), Quiz and Forum (-0.786), Quiz and Final Exam (0.254) and Quiz and Final Grade (0.388); Assignments and Files (0.675), Assignments and Folder (0.816), Assignments and Quiz (0.901), Assignments and Forum (-0.885), Assignments and Final Exam (0.302), Assignments and Test1 (0.212), Assignments and Test2 (0.241) and Assignments and Final grade (0.438); Forum and Files (-0.540), Forum and Folder (-0.726), Forum and Quiz (-0.786), Forum and Assignment (-0.885), Forum and Final Exam (-0.281), Forum and Test1 (-0.226), Forum and Test2 (-0.224), Forum and Final Grade (-0.419); Final Exam and Folder (0.288), Final Exam and Quiz (0.254), Final Exam and Assignment (0.302), Final Exam and Forum (-0.281), Final Exam and Test1 (0.465), Final Exam and Test2 (0.563), Final Exam and Final Grade (0.761); Test1 and Assignment (0.212), Test1 and Forum (-0.226), Test1 and Final Exam (0.465), Test1 and Test2 (0.416), Test1 and Final Grade (0.582); Test2 and Folder (0.200), Test2 and Assignment (0.241), Test2 and Forum (-0.224), Test2 and Final Exam (0.563), Test2 and Test1 (0.416), Test2 and Final Grade (0.564); Final Grade and Files (0.232), Final Grade and Folder (0.372), Final Grade and Quiz (0.388), Final Grade and Assignment (0.438), Final Grade and Forum (-0.419), Final Grade and Final Grade (0.761), Final Grade and Test1 (0.582), Final Grade and Test2 (0.564); Therefore, the various components influence final grade. Since Pearson Coefficient value is greater than 0.01, we accept null hypothesis and conclude that there is significant relationship between use of LMS and prevailing students' performance. The results show that the Files, Folder, Quiz, Assignment, Forum, Final Exam, Test1 and Test2 all had an impact on academic performance significantly. In this case the lecturer used the tools (research question2).

In course 3 (Table XIV) There is significant relationship at 0.01 level between Files and Folder (0.394), Files and Quiz (0.406), Files and Assignment (0.540), Files and Forum (-0.513) and Files and Final Grade (0.144); Folder and Files (0.394), Folder and Quiz (0.869), Folder and Assignment (0.858), Folder and Forum (-0.671), Folder and Final Exam (0.184), Folder and Test2 (0.145) and Folder and Final Grade (0.276); Quiz and Files (0.406), Quiz and Folder (0.869), Quiz and Assignment (0.847), Quiz and Forum (-0.727), Quiz and Final Exam (0.209), Quiz and Test1 (0.144), Quiz and Test2 (0.156) and Quiz and Final Grade (0.286); Assignments and Files (0.540), Assignments and Folder (0.858), Assignments and Quiz (0.847), Assignments and Forum (-0.742), Assignments and Final Exam (0.216), Assignments and Test1 (0.148), Assignments and Test2 (0.174) and Assignments and Final Grade (0.298); Forum and Files (-0.513), Forum and Folder (0.671), Forum and Quiz (-0.727), Forum and Assignment (-0.742), Forum and Final Exam (-0.232), Forum and Test1 (-0.211), Forum and Test2 (-0.202), Forum and Final Grade (-0.331); Final Exam and Folder (0.184), Final Exam and Quiz (0.209), Final Exam and Assignment (0.216), Final Exam and Forum (-0.232), Final Exam and Test1 (0.414), Final Exam and Test2 (0.432), Final Exam and Final Grade (0.731). Therefore, the various components influence final grade. Hence there is significant relationship between use of LMS tools and student's performance. Since Pearson Coefficient value is greater than 0.01, we accept null hypothesis and conclude that there is significant relationship between use of LMS and prevailing students' performance.

VI. CONCLUSION

In conclusion, using our approach of educational data mining and LMS, it was possible to monitor students and staff use of the LMS resources at the university. In addition it was

obvious from correlation analysis that students use of the resources could affect academic performance. In the future, it is suggested that this study be conducted with several courses from different disciplines in order to determine if academic disciplines affect students' performance based on the same tools.

REFERENCES

[1] G. Mariscal, O. Marban and C. Fernandez, " A survey of data mining and knowledge discovery process models and methodologies," The Knowledge Engineering Review, Vol. 25:2, 2010, pp. 137-166.

[2] O Marban, G.Mariscal, and J. Segovia, "A data mining and knowledge discovery process model," Open Access Data Base www.intechweb.org, 2009, pp. 1-16.

[3] H. I. Witten, and E. Frank, Data Mining: Practical Machine Learning Tools and Technology, 2nd ed, New York:Elsevier, 2005.

[4] E. U. Okike, M. Mosanako, and M. Mogorosi, "Academic data Federation, Knowledge Base Construction and Heterogeneous Data Analysis for Improved management at University levels, " Journal of Applied Information Science and Technology, vol. 11, no. 2, 2018, pp. 21-30.

[5] S. N. Hamade, " Students Perceptions of Learning Management Systems in a University Environment : Yahoo Groups Vs BlackBoard, Ninth International Conference on Information Technology, New generations 2012, pp. 594-599.

[6] Y. B. Kurata R. M. L. P. Bano, M. C. T. Marcelo, "Effectiveness of Learning Management System Application in the Learnability of Tertiary Students in an Undergraduate Engineering Program in the Philippines". In: Andre T. (eds) Advances in Human Factors in Training, Education, and Learning Sciences. AHFE 2017. Advances in Intelligent Systems and Computing, vol 596, 2018. Springer, Cham. https://doi.org/10.1007/978-3-319-60018-5_15.

[7] Y. Ghlay, Effectiveness of Learning Management Systems in Higher Education: Views of Lecturers with Different Levels of Activity in LMS, " Journal of Online Higher Education, vol. 3, no. 2, 2019 pp. 29-50.

[8] C. Romero, S. Ventura and E. Garcia, " Data Mining in Course Management Systems:Moodle case Study and Tutorial, " Computers & Education, vol. 51, 2008, pp. 368-384.

[9] E. Garcia, C. Romero, S. Ventura, and C. de Castro, "A collaborative Educational Association Rule Mining tool, " The Internet and Higher Education, vol. 14, no. 2, 2011, pp. 77-88.

[10] A. K. Alhazmi and A. A. Rahman, "Why LMS failed to support student learning in higher education institutions," 2012 IEEE Symposium on E-Learning, E-Management and E-Services, Kuala Lumpur, 2012, pp. 1-5, doi: 10.1109/IS3e.2012.6414943.

[11] R. Babo, and A. Azevedo, Higher Education Instidtutions and Learning Management Systems: Adoption and Standardization, IGI Global, 2012.

[12] J. G. Boticario, and O. C. Santos, " Issues in Developing Adaptive Learning Management Systems for Higher Education Institutions, " . Available https://core.ac.uk/reader/55533720. Retrieved 2 Oct. 2020.

[13] N. Darko-Adjei, "Students Perceptions and Use of the Sakai Learning Management System in the University of Ghana. Available http://ugspace.ug.edu.gh/handle/123456789/26847.

[14] H. Coates, R. James, & G. A. Baldwin, "Critical Examination Of The Effects Of Learning Management Systems On University Teaching And Learning," Tert Educ Manag 11, 19–36 (2005). https://doi.org/10.1007/s11233-004-3567-9

[15] L. V. Ngeze, "Learning Management Systems in Higher Learning Institutions in Tanzania: Analysis of Students' Attitudes and challenges towards the use of UDOM LMS in teaching and learning at the University of Dodoma, " 2016. Available at https://pdfs.semanticscholar.org/2c36/89101b4a64c85c25f3179c5a95e50f8a719a.pdf retrieved 2 Oct. 2020.

[16] M. F. Paulsen, "Experiences with Learning Management Systems in 113 European Institutions, " Journal of Educational Technology & Society, vol. 6. No. 4, 2003, pp.134-148

[17] N. Fathima, D. M. Shannon, and M Ross, Expanding The Technology Acceptance Model (TAM) to Examine Faculty Use of Learning Management Systems (LMSs) In Higher Education Institutions, " Journal of Online Learning and Teaching, vol. 11, no. 2, June, 2015, pp. 210.

[18] V. Mhetre, "Classification based Data mining Algorithms to Predict slow average and fast learners in educational systems using Weka, " ICCM, 2017, pp. 475-479.

[19] J. Talukdar, S. K Kalita, " Detection of Breast Cancer using Data mining tool (Weka), " International Journal of Scietntific & Engineering Research, vol. 6, no. 11, 2015, pp. 1124-1128.

Improved PSO Performance using LSTM based Inertia Weight Estimation

Y.V.R.Naga Pawan[1]
Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation
Green Fields, Vaddeswaram, Guntur District
Andhra Pradesh. India

Kolla Bhanu Prakash[2]
Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation
Green Fields, Vaddeswaram, Guntur District
Andhra Pradesh. India

***Abstract*—Particle Swarm Optimization (PSO) is first introduced in the year 1995. It is mostly an applied population-based meta-heuristic optimization algorithm. PSO is diversely used in the areas of sciences, engineering, technology, medicine, and humanities. Particle Swarm Optimization (PSO) is improved its performance by tuning the inertia weight, topology, velocity clamping. Researchers proposed different Inertia Weight based PSO (IWPSO). Every Inertia Weight based PSO in excelling the existing PSOs. A Long Short Term Memory (LSTM) predicting inertia weight based PSO (LSTMIWPSO) is proposed and its performance is compared with constant, random, and linearly decreasing Inertia Weight PSO. Tests are conducted on swarm sizes 50, 75, and 100 with dimensions 10, 15, and 25. The experimental results show that LSTM based IWPSO supersedes the CIWPSO, RIWPSO, and LDIWPSO.**

***Keywords*—Particle swarm optimization; inertia weight; long short term memory; benchmark functions; convergence**

I. Introduction

Kennedy and Eberhart [1] [2] developed a stochastic population-based optimization algorithm based on the social-behaviour metaphor of a flock of birds or a swarm of bees searching for food. It solves global optimization numerical problems. PSO is applied in every discipline of Science, Engineering, and Technology [3-8]. It is widely applied as an optimization technique in areas like communications, electronics, electrical, manufacturing, grids, cloud computing, algorithms, numerical optimization, etc. [28-40]. PSO can be extended to non-differentiable, non-linear, large search space issues, and provides better performance with decent quality [9].

Since 1995, each year, new PSO variants have been created based on initialization parameters, constriction factor, mutation operator, inertia weight, topologies, parallel processing, fuzzy logic, neural networks, ensemble, etc.,. The new variants mostly supersede established PSO variants. A comprehensive review of PSO variants is discussed in [10] [11].

Many researchers focused their attention on computing inertia weigh for faster convergence of the swarm. Different Inertia Weight Particle Swarm Optimizations (IWPSO) are discussed in [12]. It is observed that every inertia weight computing strategy supersedes the other.

In this work, a new inertia weight computing strategy is proposed. It uses a trained Long Short Term Memory (LSTM) to predict the inertia weight in every iteration, till stopping criteria is met. The predicted IW is used for computation of fitness function. Its performance is compared with Constant Inertia Weight PSO (CIWPSO) [13], Random Inertia Weight PSO (RIWPSO) [14], and Linear Decreasing Inertia Weight PSO (LDIWPSO) [15] using benchmark functions [12].

The remainder of the paper is organized as follows: Particle Swarm Optimization (PSO) and Inertia Weight based PSO is summarized in Section II, the Recurrent Neural Network, LSTM and LSTMIWPSO is briefed in Section III, Experimental Results are discussed in Section IV, and in Section V Conclusion and Future Work is briefed.

II. Particle Swarm Optimization

The formulation of PSO [16] [17] [18] is done based on the objective function given in equation (1). The objective function measures the closeness of the corresponding solution to the optimum.

$$f(x): \mathbb{R}^d \rightarrow \mathbb{R} \quad (1)$$

where d is the number of dimensions of *search space*, *S*. S is a subset of $\mathbb{R}^d$, shown in equation (2) and defined by equation (3). The global optimization problem is shown in equation (4) and equation (5).

$$S \subseteq \mathbb{R}^D \quad (2)$$

$$S = \{px_i \mid px_{min} \leq px_i \leq px_{max}\} \quad (3)$$

$$\min_{x \in S} f(x) \quad (4)$$

The objective function, $f(x)$, needs $px_i \in S$ such that:

$$\forall\, py_i \in S: f(px_i) \leq f(py_i) \quad (5)$$

In the Basic PSO (BPSO), a Swarm, SW, consists of n particles represented as $SW = \{P_1, P_2, P_3, \ldots, P_n\}$. Each Particle P_i has a position in the search space represented by $PX_i = \{px_{i1}, px_{i2}, px_{i3}, \ldots, px_{iD}\}$ where D is D-dimensional search space. In the search space, each particle P_i, moves with a velocity V_i, represented as $PV_i = \{pv_{i1}, pv_{i2}, pv_{i3}, \ldots, pv_{iD}\}$. Each particle, P_i, maintains its best position, Pb_i, represented as $Pb_i = \{pb_{i1}, pb_{i2}, pb_{i3}, \ldots, pb_{iD}\}$. Among the population of all particles, the best particle is determined and represented as

$P_g = \{pg_{i1}, pg_{i2}, pg_{i3}, \ldots, pg_{iD}\}$. The basic equations with the functioning of BPSO are given by (6) and (7).

$$pv_{id} = pv_{id} + c_1 * random(\ \ ) * (pb_i - px_{id}) + c_2 * Random(\ \ ) * (pg_i - px_{id}) \quad (6)$$

$$px_{id} = px_{id} + pv_{id} \quad (7)$$

where c_1 and c_2 are two positive acceleration coefficients, random() and Random() are two random functions in the [0,1]. pv_i s then clamped to a maximum velocity pv_{max}, the parameter given by the user. The first part of the (6) represents the previous velocity, the second part is the cognition part of the particle, and the third part represents the cooperation among the particles [1][17][19].

As particles tends to explore the search space hugely, the velocities of the particles are limited to the constant pv_{max} [16]. The particle velocity is adjusted using.

$$pv_{id} = \begin{cases} pv_{id}\ if\ -pv_{max} \le pv_{id} \le pv_{max} \\ pv_{max}\ if\ pv_{id} > pv_{max} \\ -pv_{max}\ if\ pv_{id} < -pv_{max} \end{cases} \quad (8)$$

The value for pv_{max} is typically chosen as a fraction of the search space dimension shown as (4) [20] [21], where δ is the velocity clamping factor.

$$pv_{max} = \delta * (px_{max} - px_{min}) \text{ where } \delta \in (0, 1) \quad (9)$$

As the search space, S, is bounded by the interval $[px_{min}, px_{max}]$, the velocity clamping [22] of the particle is in the interval [-pv_{max}, pv_{max}] $[pv_{min}, pv_{max}]$,

where $pv_{max} = \delta * {(px_{max} - px_{min})}/{2}$.

A. Inertia Weight based PSO

Shi and Russell Eberhart [13], developed inertia weight based PSO (IWPSO). In IWPSO, exploration and exploitation of swarm particles are controlled. The equation (6) with inertia weight is given by equation (10).

$$pv_{id} = \omega * pv_{id} + c_1 * random(\ \ ) * (pb_i - px_{id}) + c_2 * Random(\ \ ) * (pg_i - px_{id}) \quad (10)$$

III. Recurrent Neural Network

Recurrent Neural Networks (RNN) are time-dynamic discrete systems dealing with input vector sequences [23] [24]. RNNs traditionally propagate information forward in time, forming predictions using only past and present inputs. The basic Recurrent Neural network is shown in Fig. 1. The traditional RNN, for each time step t, the output is computed using equation (11), and the activation function $a^{<t>}$ is computed using equation (12).

$$y^{<t>} = h(W_{ya}\, a^{<t>} + b_y) \quad (11)$$

$$a^{<t>} = g(W_{aa}\, a^{<t-1>} + W_{ax} x^{<t-1>} + b_a) \quad (12)$$

where t represents time, $y^{<t>}$ is the predicted value, W_{ya}, W_{aa}, W_{ax}, b_y, and b_a are the coefficients, and h and g are the activation functions. Generally, activation functions are given in equations (13), (14), and (15).

$$Sigmoid\ function, g(a) = \frac{1}{1+e^{-a}} \quad (13)$$

$$tanh, g(a) = \frac{e^{a} - e^{-a}}{e^{a} + e^{-a}} \quad (14)$$

$$RELU, g(a) = \max(0, a) \quad (15)$$

RNN is observed with vanishing [25] and exploding gradient [26] phenomenon. It is due to multiplicative gradient and resulting in its inability to catch dependencies that can be exponentially decreasing/increasing with respect to the number of layers.

In RNN, the loss function, $\mathcal{L}$, for all time steps is defined based on the loss obtained at every time step.

$$\text{Loss Function}, \mathcal{L}(\hat{y}, y) = \sum_{t=1}^{T_y} \mathcal{L}(\hat{y}^{<t>}, y^{<t>}) \quad (16)$$

$where\ \hat{y}^{<t>}, y^{<t>}$ are predicted and expected outputs.

A. Long Short Term Memory

Long Short Term Memory is special kind of RNN architecture capable in learning long term dependencies. Hochreiter and Schmidhuber [27] introduced the efficient and effective, gradient based the Long Short Term Memory (LSTM). Fig. 2 depicts the dependencies of the memory cell of an LSTM depicting dependencies. In order to deal with vanishing gradient problem, The LSTM has the power to delete or add information to a cell state that is carefully controlled by mechanisms called gates [27]. LSTM uses three gates called update gate (Γ_u), forget gate (Γ_f) and output gate (Γ_o). The computation of $\tilde{c}^{<t>}$, $c^{<t>}$, $a^{<t>}$, Γ_u, Γ_f, Γ_o are shown through equation (17) – equation (22).

$$\tilde{c}^{<t>} = \tanh(w_c[a^{<t-1>}, x^{<t>}] + b_c) > FUnctiontion \quad (17)$$

$$\Gamma_u = \sigma(w_u[a^{<t-1>}, x^{<t>}] + b_u \quad (18)$$

$$\Gamma_f = \sigma(w_f[a^{<t-1>}, x^{<t>}] + b_f \quad (19)$$

$$\Gamma_o = \sigma(w_o[a^{<t-1>}, x^{<t>}] + b_o \quad (20)$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>} \quad (21)$$

$$a^{<t>} = \Gamma_o * \tanh c^{<t>} \quad (22)$$

Let $\hat{y}^{<t>}$ be the predicted output at each time step and $y^{<t>}$ be the actual output at each time step. Then the error at each time step is given by:-

$$E^{<t>} = -y^{<t>} \log(\hat{y}^{<t>}) \quad (23)$$

$$E_{total} = \sum_t E^{<t>} \quad (24)$$

$$E_{total} = \sum_t -y^{<t>} \log(\hat{y}^{<t>}) \quad (25)$$

The value of $\frac{\partial E}{\partial W}$ can be calculated as the summation of the gradients at each step

$$\frac{\partial E}{\partial W} = \frac{\partial E^{<t>}}{\partial \hat{y}^{<t>}} \frac{\partial \hat{y}^{<t>}}{\partial a^{<t>}} \frac{\partial a^{<t>}}{\partial c^{<t>}} \frac{\partial c^{<t>}}{\partial c^{<t-1>}} \ldots \frac{\partial c^{<0>}}{\partial W} \quad (26)$$

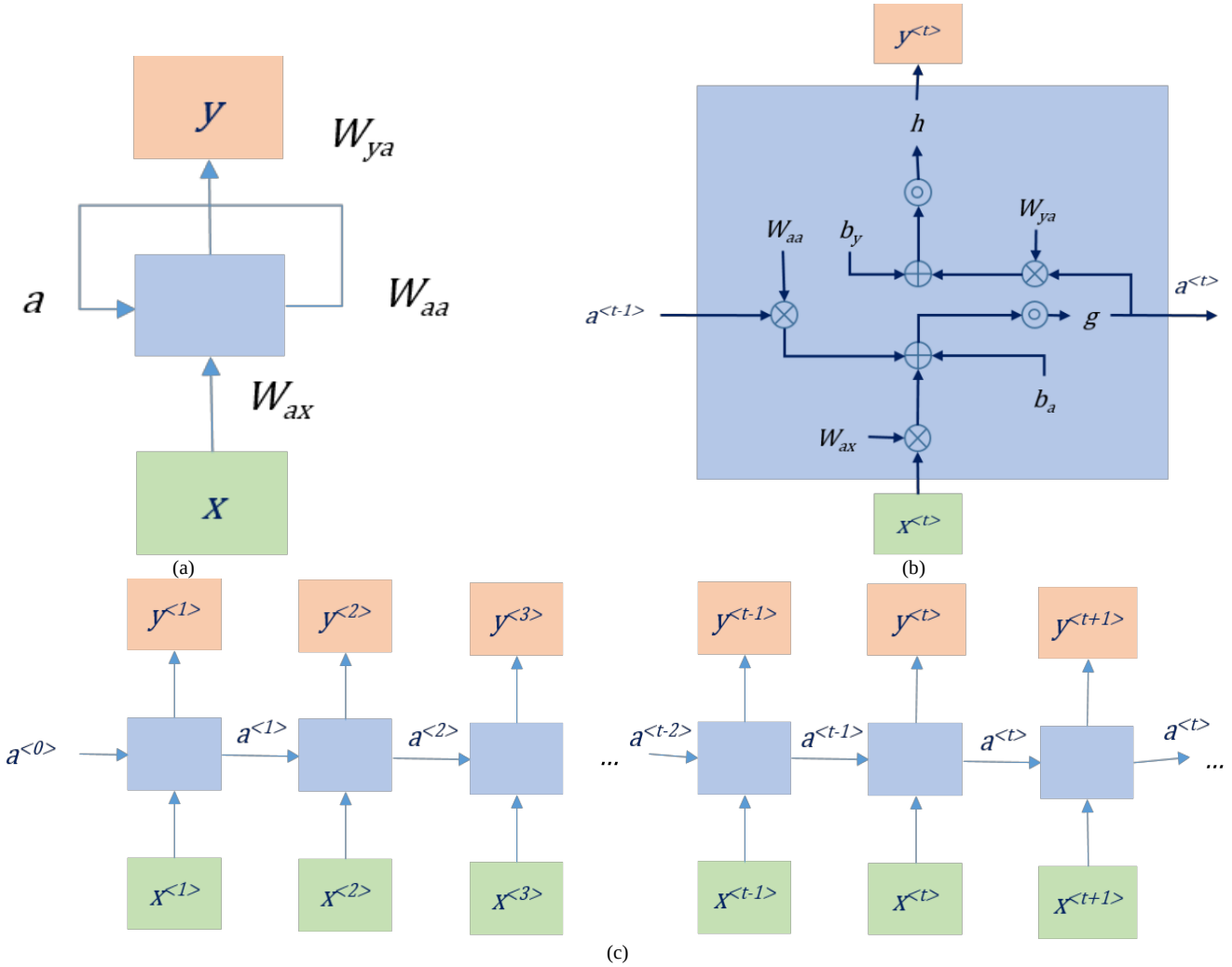


Fig. 1. (a): A Recurrent Neural Network. (b): RNN Cell Handling Dependencies. (c): Unrolled Recurrent Neural Network.

Thus the total error gradient is given by equations (27) and (28):-

$$\frac{\partial E}{\partial W} = \sum_t \frac{\partial E^{<t>}}{\partial W} \tag{27}$$

$$\frac{\partial E}{\partial W} = \sum_t \frac{\partial E^{<t>}}{\partial \hat{y}^{<t>}} \frac{\partial \hat{y}^{<t>}}{\partial a^{<t>}} \frac{\partial a^{<t>}}{\partial c^{<t>}} \frac{\partial c^{<t>}}{\partial c^{<t-1>}} \dots \frac{\partial c^{<0>}}{\partial W} \tag{28}$$

It is to note the gradient equation involves a chain of $\partial c^{<t>}$ for an LSTM Back-Propagation while the gradient equation involves a chain of $\partial a^{<t>}$ for a basic Recurrent Neural Network.

B. LSTM Inertia Weight based PSO

In LSTMIWPSO, the new inertia weight is computed using LSTM. Initially, LSTM is trained with different inertia weights from 0.05 to 1.00. In every iteration, a new IW is predicted using trained LSTM. The predicted IW is used to move the swarm using equations (10) and (7). The process is terminated when the stopping criterion is reached. The pseudocode for LSTMIWPSO is shown in Fig. 3.

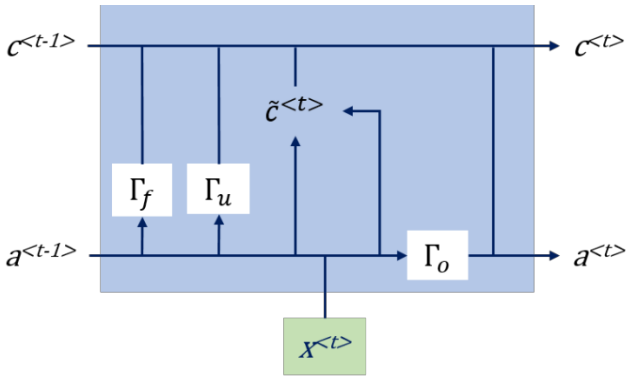


Fig. 2. Memory Cell of an LSTM Showing Dependencies.

The pseudocode for LSTMIWPSO is given below:

Step 1:
Initialization
For each particle, P_i, in the population
Initialize px_i with uniform distribution
Initialize pv_i randomly.
Build and Train LSTM network for Inertia Weight Prediction.
Predict the new Inertia Weight.
Evaluate the objective function of px_i and assigned the value to fitness[i].
Initialize $pbest_i$ with a copy of px_i.
Initialize $pbest\_ftness_i$ with a copy of $fitness_i$.
Initialize pgbest with index of the particle with the least fitness.
Step 2:
Repeat until stopping criterion is reached
For each particle, P_i,:
Update pv_i and px_i according to the equations (10) and (7)
Evaluate $fitness_i$
If $fitness_i < pbest\_fitness_i$ then
$Pbest_i = px_i$
$Pbest\_fitness_i = fitness_i$
Update pgbest by the particle with current least fitness among the population
Predict the new Inertia Weight using trained LSTM

Fig. 3. Pseudocode of LSTMIWPSO.

IV. Experimental Results

Experiments are conducted with different Inertia Weight based PSOs namely, CIWPSO, RIWPSO, LDIWPSO, and LSTMIWPSO over different optimization test problems tabulated in Table I.

Swarm sizes of 50, 75 and 100 particles of different dimensions, 10, 15 and 25, are considered for experiments. A total of 15 simulations are performed to reduce the occurrence of randomness. Along with LSTMIWPSO, LDIWPSO, RIWPSO and CIWPSO are implemented. The results are collected in terms of the best error, mean error, variance, standard deviation, mean square error, root mean square error, mean iteration and mean time taken (in seconds) to evaluate the performance of LSTMIWPSO with CIWPSO, RIWPSO and LDIWPSO.

From Table II and Fig. 4, the performance of LSTMIWPSO, for benchmark functions f1, f3, f4, and f5 as fitness functions, swarm size with the dimension 10, the best error is nearer to CIWPSO, RIWPSO, and LDIWPSO. The best error is moderately higher, in the case of dimensions 15 and 25. For f2 function, the best error for LSTMIWPSO is the same as CIWPSO, RIWPSO, and LDIWPSO.

For swarm sizes 50, 75, and 100 with dimensions 10, 15, and 25 and f1-f5 as fitness functions, the mean error is computed using the CIWPSO, RIWPSO, and LDIWPSO, and LSTMIWPSO. The processed results are collected and tabulated in Table III and graphically shown in Fig. 5. The mean error, except for swarm size 100 and dimension 10, when compared to CIWPSO, RIWPSO and LDIWPSO, for LSTMIWPSO, is smaller.

The variance and standard deviation are computed to access the performance of CIWPSO, RIWPSO, LDIWPSO and LSTMIWPSO. The computed results are tabled in Table IV and V. The same are shown graphically in Fig. 6 and Fig. 7. From Table IV, Table V, Fig. 6, and Fig. 7, it is evident that the performance of LSTMIWPSO in terms variance and standard deviation is flair with swarm sizes 50, 75, and 100, with dimensions 10, 15 and 25 on the benchmark functions f1 - f5.

To access the CIWPSO, RIWPSO, LDIWPSO and LSTMIWPSO performance, the MSE and RMSE are computed. Tables VI and VII show the computed results. The same is seen in Fig. 8 and Fig. 9 graphically. It is evident from Table VI, Table VII, Fig. 8, and Fig. 9 that LSTMIWPSO's output in terms of MSE and RMSE is substantially better for swarm sizes 50, 75, and 100, and for benchmark functions f1 - f5, with dimensions 10, 15 and 25, except for the swarm size 100 and dimension 10.

From Table VIII and Fig. 10, the meantime for LSTMIWPSO is transcending for the swarm sizes 75 and100 with dimension 10. In other scenarios, it is non-paying when compared with other methods for the benchmarks considered.

From Table IX and Fig. 11, the mean iterations for LSTMIWPSO are decent when compared with CIWPSO, RIWPSO, and LDIWPSO, with Swarm size 100 and dimension 10. Similarly, LSTMIWPSO has achieved adequate performance with f2, f3, and f4 benchmark functions.

LSTMIWPSO delivered adequate results over CIWPSO, RIWPSO, and LDIWPSO from the perspective of mean error, variance & standard deviation and, MSE & RMSE. It is good in limited scenarios in terms of best error, Mean Time and Mean Iterations.

TABLE I. BENCHMARK FUNCTIONS (BMF)

Benchmark Function name	Properties	Benchmark Function	Search Space	Best fitness value at
Ackley (f1)	n-dimensional, continuous, multimodal, non-convex, differentiable	$-20\exp(-0.2\sqrt{\frac{1}{n}\sum_{d=1}^{n} pos_d^2} - \exp(\frac{1}{n}\sum_{d=1}^{n}\cos(2\pi\, pos_d)) + 20 + \exp(1)$	[-32, +32]	f(0) = 0
Alpine (f2)	n-dimensional, non-separable, multimodal, non-convex, differentiable	$\sum_{d=1}^{n} \lvert pos_d \,.\sin(pos_d) + 0.1\, pos_d\rvert$	[0, 10]	f(0) = 0
Rastrigin (f3)	n-dimensional, continuous, differentiable, separable, multimodal, convex	$10\,.\,n + \sum_{d=1}^{n}(pos_d^2 - 10\,.\cos(2\pi pos_d))$	[-5.12, +5.12]	f(0) = 0
Rosenbrock (f4)	n-dimensional, continuous, differentiable, non-separable, multimodal, non-convex	$\sum_{d=1}^{n}[100\,.(pos_{d+1} - pos_d^2)^2 + (1 - pos_d)^2]$	[-5, 10]	f(1) = 0
Sphere (f5)	n-dimensional, continuous, convex, differentiable, unimodal, separable	$\sum_{d=1}^{n} pos_d^2$	[-5.12, +5.12]	f(0) = 0

TABLE II. COMPUTED BEST ERROR FOR PSOS WITH RESPECT TO DIFFERENT SWARM SIZES AND DIMENSIONS (FIG. 4)

Swarm Size	Dimension	BMF	PSOs			
			CIWPSO	RIWPSO	LDIWPSO	LSTMIWPSO
50	10	f1	0.000010	**0.000009**	**0.000009**	0.000059
		f2	**0.000000**	**0.000000**	**0.000000**	**0.000000**
		f3	0.000009	0.000010	**0.000007**	0.000009
		f4	0.000008	0.000007	0.000008	**0.000006**
		f5	**0.000007**	0.000009	**0.000007**	0.000009
	15	f1	**0.002602**	0.003873	0.005756	0.015133
		f2	**0.000000**	**0.000000**	**0.000000**	**0.000000**
		f3	0.000052	**0.000040**	0.000093	0.000209
		f4	0.000012	**0.000010**	0.000032	0.000155
		f5	0.000010	0.000283	0.000460	0.001303
	25	f1	**0.031975**	0.056392	0.429402	1.262133
		f2	**0.000000**	**0.000000**	**0.000000**	**0.000000**
		f3	**0.001684**	0.001990	0.007345	0.021180
		f4	**0.000319**	0.000401	0.001488	0.003145
		f5	**0.001131**	0.004371	0.017108	0.035360
75	10	f1	**0.000008**	0.000009	**0.000008**	0.000009
		f2	**0.000000**	**0.000000**	**0.000000**	**0.000000**
		f3	0.000009	0.000009	**0.000004**	0.000008
		f4	**0.000006**	0.000007	0.000007	0.000008
		f5	0.000009	**0.000006**	0.000007	0.000008
	15	f1	0.000519	0.000813	**0.000330**	0.002932
		f2	**0.000000**	**0.000000**	**0.000000**	**0.000000**
		f3	**0.000010**	**0.000010**	0.000022	0.000158
		f4	0.000009	**0.000008**	0.000010	0.000010
		f5	**0.000010**	**0.000010**	**0.000010**	0.000114
	25	f1	**0.027271**	0.044731	0.124196	0.486989

		f2	**0.000000**	**0.000000**	**0.000000**	**0.000000**
		f3	**0.000640**	0.001245	0.006856	0.014690
		f4	0.000125	**0.000074**	0.000165	0.001404
		f5	**0.000529**	0.001319	0.004397	0.008516
100	10	f1	0.000007	0.000007	0.000008	**0.000006**
		f2	**0.000000**	**0.000000**	**0.000000**	**0.000000**
		f3	0.000007	0.000006	**0.000005**	0.000007
		f4	0.000008	0.000007	**0.000006**	0.000007
		f5	**0.000006**	0.000007	0.000007	0.000008
	15	f1	0.000087	0.000278	**0.000035**	0.001478
		f2	**0.000000**	**0.000000**	**0.000000**	**0.000000**
		f3	0.000010	0.000010	**0.000009**	0.000013
		f4	0.000009	**0.000007**	0.000008	0.000010
		f5	0.000010	**0.000009**	**0.000009**	0.000012
	25	f1	**0.014225**	0.030603	0.083952	0.133495
		f2	**0.000000**	**0.000000**	**0.000000**	**0.000000**
		f3	**0.000387**	0.001223	0.000928	0.002193
		f4	0.000076	**0.000030**	0.000340	0.000644
		f5	0.000279	**0.000026**	0.001968	0.004294

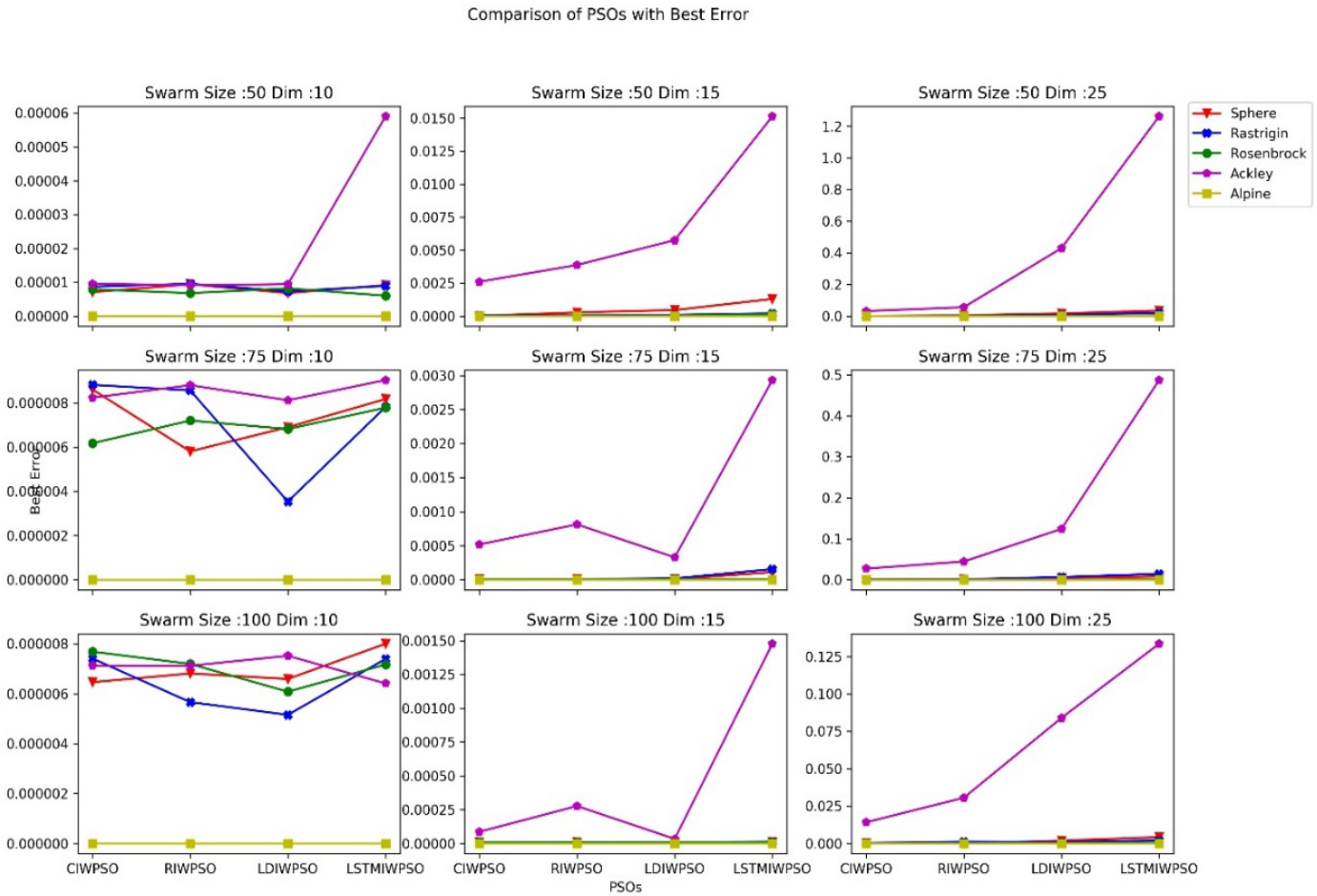


Fig. 4. Best Error Computed for the Swarm Size of 50, 75, and 100 with Dimensions 10, 15 and 25.

TABLE III. COMPUTED MEAN ERROR FOR PSOS WITH RESPECT TO DIFFERENT SWARM SIZES AND DIMENSIONS. (FIG 5)

Swarm Size	Dimension	BMF	PSOs			
			CIWPSO	RIWPSO	LDIWPSO	LSTMIWPSO
50	10	f1	3.829570	5.472203	8.170220	**2.726947**
		f2	54.932726	54.983211	62.611015	**41.404392**
		f3	0.127753	0.150475	0.328519	**0.098769**
		f4	0.052547	0.028840	0.061509	**0.030330**
		f5	0.225241	0.148772	0.319764	**0.102929**
	15	f1	10.028167	14.396357	19.456177	**6.658518**
		f2	85.250766	90.343807	91.788437	**76.181956**
		f3	0.244939	0.261647	0.445211	**0.182114**
		f4	0.039496	0.053941	0.073551	**0.027963**
		f5	0.268360	0.376879	0.456456	**0.178421**
	25	f1	33.171864	39.722826	52.601215	**25.140189**
		f2	105.036487	104.440098	150.532868	**102.950521**
		f3	0.779991	1.220720	1.224990	**0.636701**
		f4	0.117500	0.181210	0.204600	**0.110673**
		f5	0.839036	0.981923	1.341617	**0.669297**
75	10	f1	4.463292	4.308925	10.524850	**3.454499**
		f2	57.500383	58.948269	64.581158	**44.469230**
		f3	0.217088	0.141495	0.297972	**0.219178**
		f4	0.050142	**0.030269**	0.057245	0.055129
		f5	0.210561	**0.142058**	0.341972	0.294266
	15	f1	8.007375	10.336619	14.119257	**5.558153**
		f2	87.973719	85.348446	101.576149	**69.612380**
		f3	0.204710	0.265032	0.370679	**0.128782**
		f4	0.045569	0.045599	0.065754	**0.021016**
		f5	0.199728	0.283209	0.368008	**0.138094**
	25	f1	23.154072	27.528130	41.546209	**17.278593**
		f2	158.914098	154.805922	159.278378	**105.304169**
		f3	0.602166	0.707678	1.080641	**0.460648**
		f4	0.097639	0.141208	0.176523	**0.074067**
		f5	0.557023	0.795652	1.061555	**0.464447**
100	10	f1	**4.888757**	5.141855	9.926288	8.860705
		f2	53.176830	50.619999	59.964870	**34.532425**
		f3	0.214487	**0.131158**	0.275377	0.304044
		f4	0.051496	**0.030658**	0.050414	0.070342
		f5	0.232056	**0.155269**	0.306427	0.364919
	15	f1	6.210106	6.730709	12.998519	**4.498284**
		f2	94.473432	82.665049	92.999693	**69.658835**
		f3	0.213848	0.267667	0.361411	**0.103306**
		f4	0.049426	0.041846	0.079216	**0.024861**
		f5	0.197348	0.222021	0.349822	**0.104669**
	25	f1	20.104113	22.065833	35.406524	**14.607688**
		f2	158.586419	104.118385	180.423557	**102.108097**
		f3	0.491150	0.705747	0.947726	**0.383817**
		f4	0.078310	0.113074	0.150317	**0.059783**
		f5	0.505790	0.651178	0.972611	**0.386508**

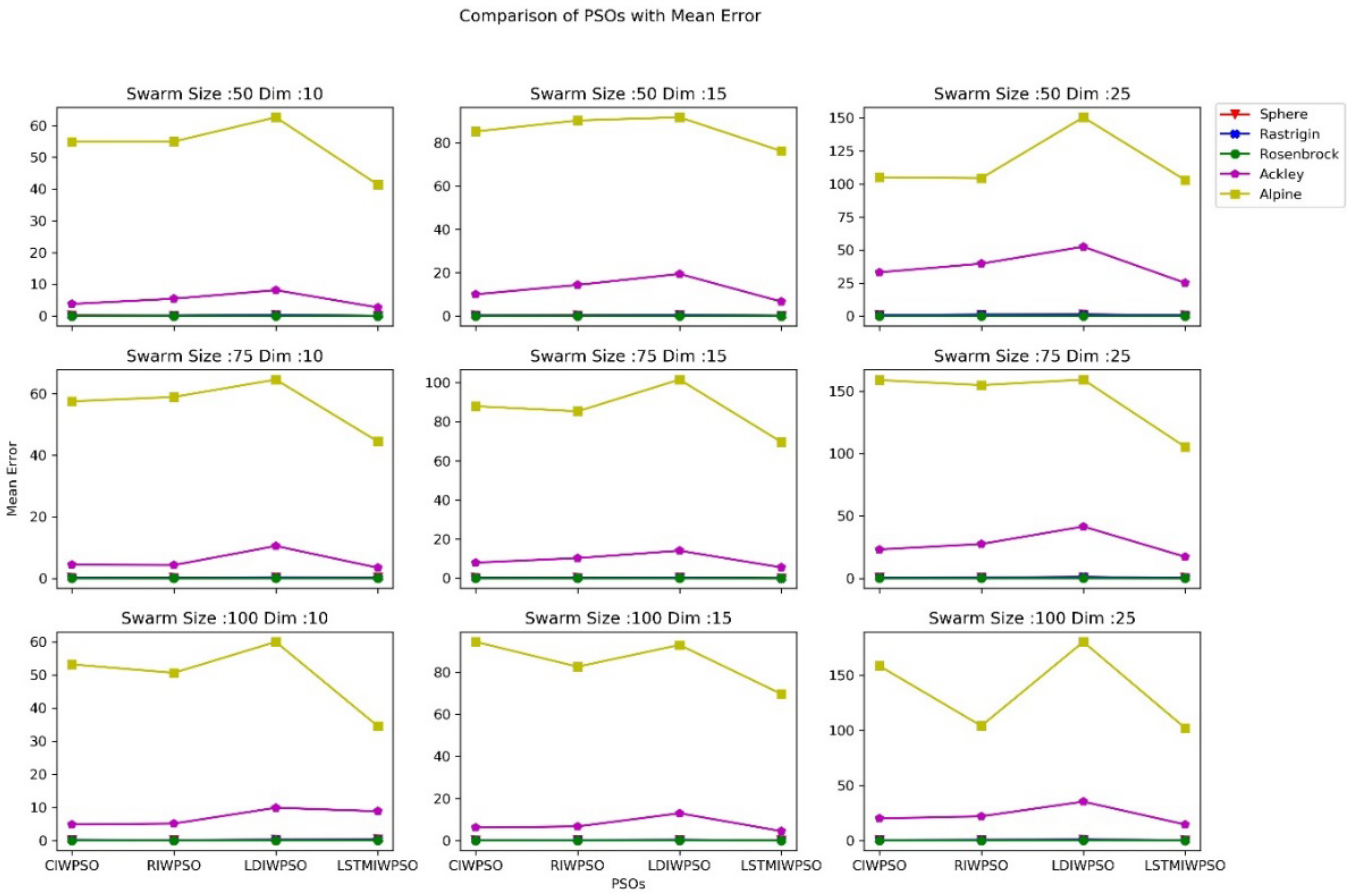


Fig. 5. Mean Error Computed for the Swarm Size of 50, 75, and 100 with Dimensions 10, 15 and 25.

TABLE IV. COMPUTED VARIANCE FOR PSOS WITH RESPECT TO DIFFERENT SWARM SIZES AND DIMENSIONS. (FIG. 6)

Swarm Size	Dimension	BMF	PSOs			
			CIWPSO	RIWPSO	LDIWPSO	LSTMIWPSO
50	10	f1	2.24626E+03	3.38851E+03	4.67975E+03	**9.63606E+02**
		f2	3.70666E+03	3.69348E+03	4.13007E+03	**2.80457E+03**
		f3	1.62064E+00	2.07616E+00	5.02929E+00	**9.41880E-01**
		f4	1.20428E-01	6.00913E-02	1.40112E-01	**4.69751E-02**
		f5	3.80843E+00	2.25555E+00	5.23459E+00	**9.06364E-01**
	15	f1	1.03989E+04	1.26799E+04	1.72068E+04	**3.34864E+03**
		f2	1.07078E+04	1.22065E+04	1.20338E+04	**8.28701E+03**
		f3	6.12127E+00	5.95316E+00	1.12496E+01	**2.54517E+00**
		f4	1.53739E-01	1.85184E-01	2.84031E-01	**5.52154E-02**
		f5	6.13349E+00	7.48563E+00	9.61024E+00	**2.78097E+00**
	25	f1	5.01816E+04	4.97107E+04	7.48609E+04	**1.94431E+04**
		f2	3.92425E+03	4.12939E+03	3.72553E+04	**2.44186E+03**
		f3	2.79965E+01	2.78540E+01	4.39840E+01	**1.18317E+01**
		f4	6.95780E-01	8.72313E-01	1.18338E+00	**3.70447E-01**
		f5	3.28743E+01	3.26392E+01	4.72444E+01	**1.44918E+01**
75	10	f1	2.09251E+03	2.13522E+03	5.23097E+03	**1.20472E+03**
		f2	3.88104E+03	3.57230E+03	4.06253E+03	**2.88180E+03**
		f3	3.10357E+00	1.88301E+00	3.61849E+00	**1.57710E+00**
		f4	1.11167E-01	6.93797E-02	1.36452E-01	**6.08377E-02**
		f5	3.25187E+00	**1.78081E+00**	4.39770E+00	2.39049E+00
	15	f1	7.28370E+03	8.33270E+03	1.39262E+04	**3.36059E+03**

		f2	1.02799E+04	9.87329E+03	1.21953E+04	**8.51558E+03**
		f3	5.22510E+00	6.01068E+00	8.47051E+00	**1.76031E+00**
		f4	1.86551E-01	1.57682E-01	2.49966E-01	**4.57045E-02**
		f5	4.41418E+00	5.74053E+00	8.20269E+00	**2.02216E+00**
	25	f1	3.80660E+04	3.68687E+04	5.74084E+04	**1.39583E+04**
		f2	3.84260E+04	3.80308E+04	3.66169E+04	**4.34064E+03**
		f3	2.33483E+01	1.93322E+01	3.67910E+01	**9.85238E+00**
		f4	6.57021E-01	6.94181E-01	1.00068E+00	**2.66049E-01**
		f5	2.15837E+01	2.26723E+01	3.76093E+01	**1.01227E+01**
100	10	f1	**2.16477E+03**	2.46721E+03	5.04084E+03	2.69153E+03
		f2	2.82820E+03	3.13702E+03	3.16738E+03	**2.24332E+03**
		f3	2.69688E+00	**1.48920E+00**	3.48140E+00	2.32310E+00
		f4	1.10163E-01	**5.92556E-02**	1.15670E-01	8.74897E-02
		f5	3.43568E+00	**2.10168E+00**	3.99214E+00	2.99379E+00
	15	f1	5.89145E+03	6.23226E+03	1.23640E+04	**2.40498E+03**
		f2	1.05402E+04	8.93997E+03	1.02147E+04	**8.05012E+03**
		f3	5.26484E+00	6.10921E+00	8.28945E+00	**1.41493E+00**
		f4	1.94574E-01	1.36136E-01	3.09916E-01	**5.09724E-02**
		f5	5.05269E+00	4.93674E+00	7.31503E+00	**1.41872E+00**
	25	f1	2.96978E+04	3.21262E+04	4.78574E+04	**1.25404E+04**
		f2	3.95934E+04	3.14574E+03	4.17462E+04	**1.55817E+03**
		f3	1.95162E+01	2.34042E+01	3.57999E+01	**7.60381E+00**
		f4	5.05917E-01	6.94567E-01	8.81406E-01	**2.23919E-01**
		f5	2.10448E+01	2.21517E+01	3.04020E+01	**8.26956E+00**

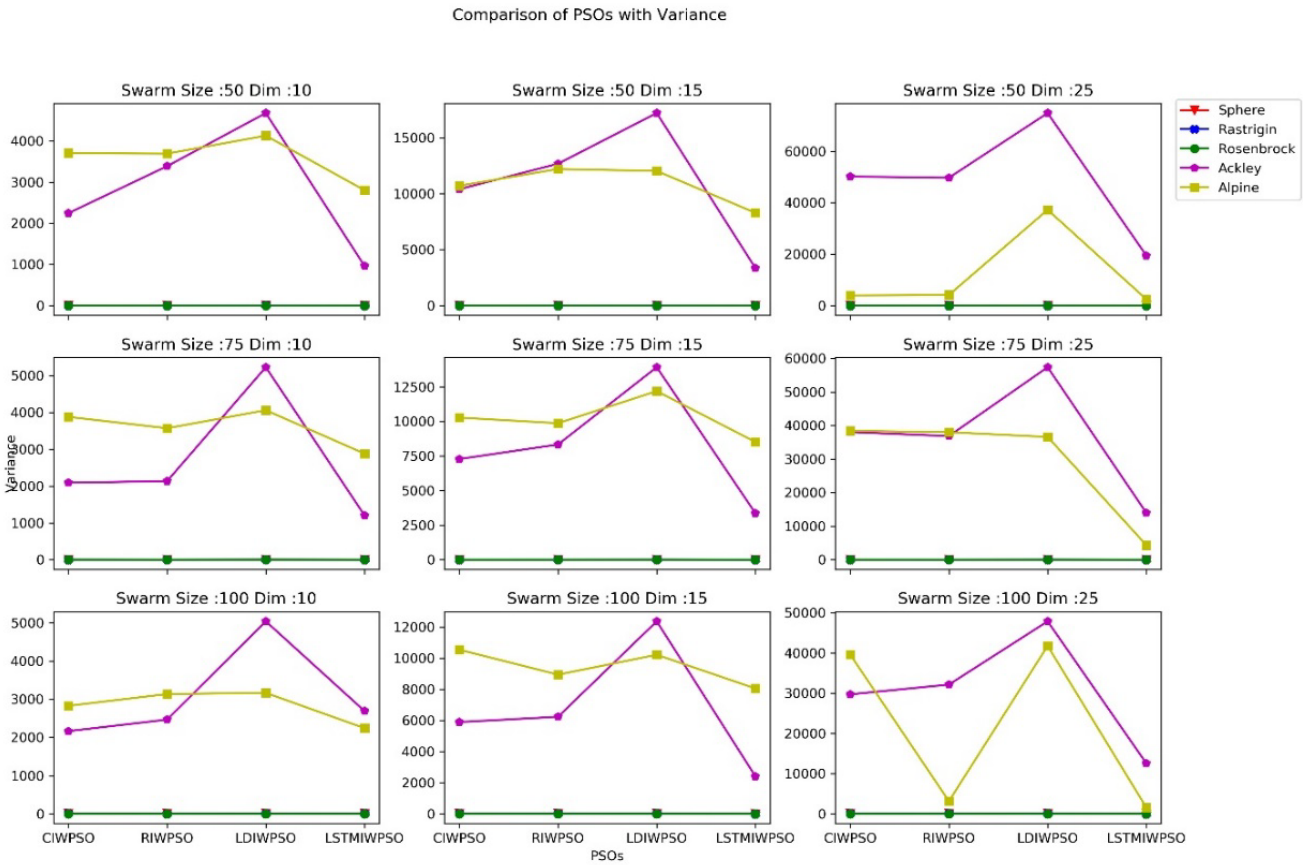


Fig. 6. Variance Computed for the Swarm size of 50, 75, and 100 with Dimensions 10, 15 and 25.

TABLE V. COMPUTED STANDARD DEVIATION FOR PSOS WITH RESPECT TO DIFFERENT SWARM SIZES AND DIMENSIONS. (FIG. 7)

Swarm Size	Dimension	BMF	PSOs			
			CIWPSO	RIWPSO	LDIWPSO	LSTMIWPSO
50	10	f1	47.394690	58.210892	68.408725	**31.042005**
		f2	60.882317	60.774044	64.265634	**52.958210**
		f3	1.273044	1.440890	2.242608	**0.970505**
		f4	0.347028	0.245135	0.374315	**0.216737**
		f5	1.951520	1.501847	2.287923	**0.952031**
	15	f1	101.974922	112.604830	131.174709	**57.867444**
		f2	103.478490	110.482926	109.698532	**91.033028**
		f3	2.474120	2.439909	3.354034	**1.595360**
		f4	0.392095	0.430330	0.532945	**0.234980**
		f5	2.476588	2.735988	3.100039	**1.667625**
	25	f1	224.012548	222.959056	273.607132	**139.438498**
		f2	62.643794	64.260301	193.016188	**49.415159**
		f3	5.291170	5.277686	6.632043	**3.439728**
		f4	0.834135	0.933977	1.087831	**0.608644**
		f5	5.733607	5.713069	6.873457	**3.806805**
75	10	f1	45.743919	46.208445	72.325443	**34.709071**
		f2	62.297977	59.768716	63.737955	**53.682396**
		f3	1.761696	1.372227	1.902232	**1.255827**
		f4	0.333417	0.263400	0.369394	**0.246653**
		f5	1.803295	**1.334471**	2.097070	1.546120
	15	f1	85.344604	91.283633	118.009469	**57.970561**
		f2	101.389839	99.364403	110.432280	**92.279921**
		f3	2.285847	2.451669	2.910414	**1.326765**
		f4	0.431915	0.397092	0.499966	**0.213786**
		f5	2.100994	2.395941	2.864034	**1.422026**
	25	f1	195.105057	192.012109	239.600461	**118.145195**
		f2	196.025427	195.014739	191.355322	**65.883501**
		f3	4.832010	4.396835	6.065558	**3.138851**
		f4	0.810568	0.833175	1.000340	**0.515799**
		f5	4.645820	4.761545	6.132645	**3.181614**
100	10	f1	46.527046	49.671001	70.998898	**51.879999**
		f2	53.180790	56.009131	56.279433	**47.363728**
		f3	1.642218	1.220329	1.865850	**1.524173**
		f4	0.331908	0.243425	0.340103	**0.295787**
		f5	1.853558	1.449718	1.998035	**1.730258**
	15	f1	76.755788	78.944668	111.193549	**49.040620**
		f2	102.665394	94.551437	101.067981	**89.722486**
		f3	2.294524	2.471681	2.879141	**1.189510**
		f4	0.441106	0.368965	0.556701	**0.225771**
		f5	2.247819	2.221878	2.704631	**1.191100**
	25	f1	172.330480	179.237775	218.763244	**111.983985**
		f2	198.980864	56.086860	204.318741	**39.473677**
		f3	4.417719	4.837791	5.983300	**2.757501**
		f4	0.711278	0.833407	0.938832	**0.473201**
		f5	4.587458	4.706557	5.513798	**2.875684**

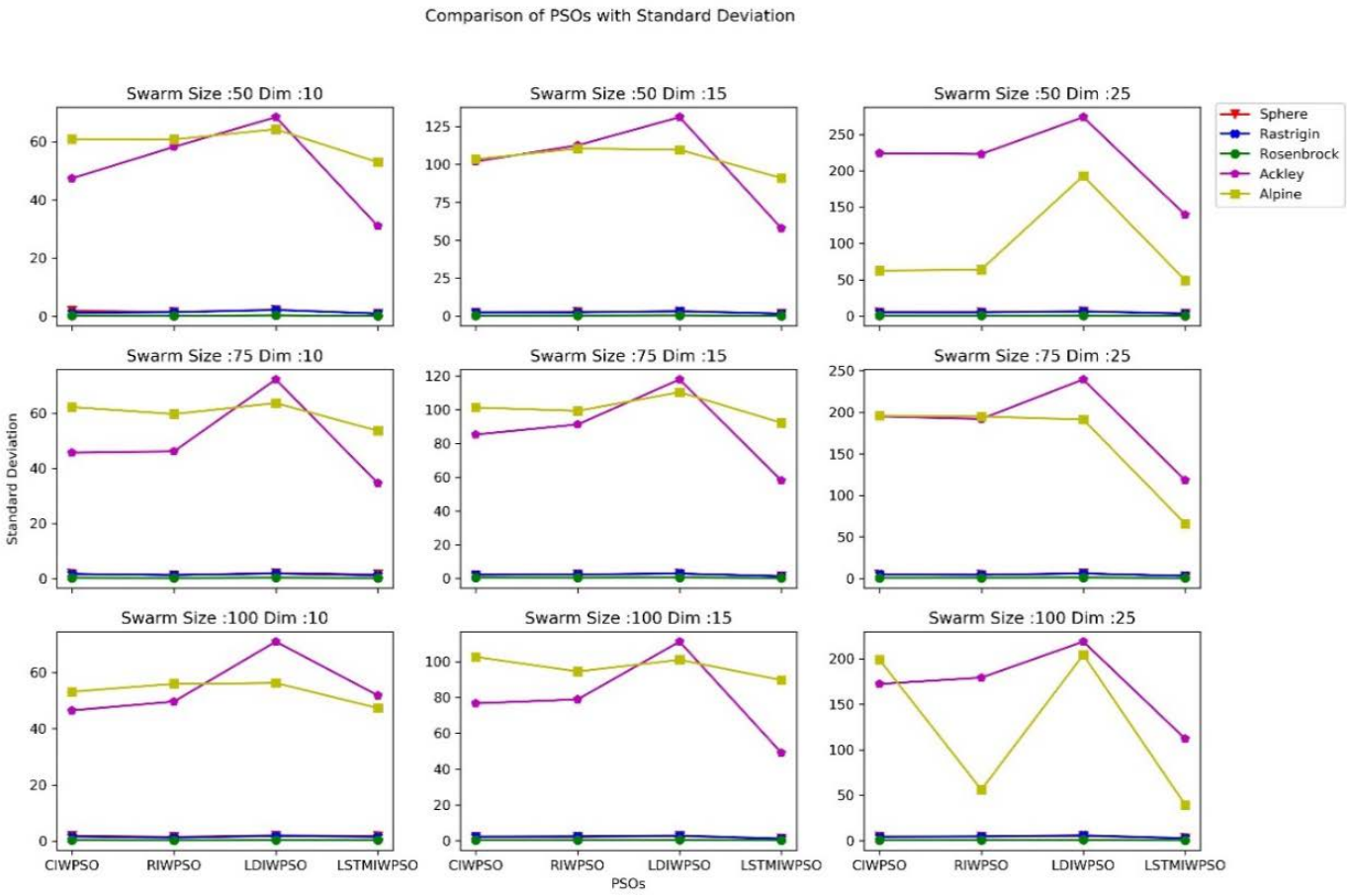


Fig. 7. Standard Deviation Computed for the Swarm size of 50, 75, and 100 with Dimensions 10, 15 and 25.

TABLE VI. COMPUTED MEAN SQUARED ERROR (MSE) FOR PSOS WITH RESPECT TO DIFFERENT SWARM SIZES AND DIMENSIONS. (FIG. 8)

Swarm Size	Dimension	BMF	PSOs			
			CIWPSO	RIWPSO	LDIWPSO	LSTMIWPSO
50	10	f1	2.26077E+03	3.41820E+03	4.74615E+03	**9.70978E+02**
		f2	6.67734E+03	6.66534E+03	7.98667E+03	**4.49193E+03**
		f3	1.63680E+00	2.09866E+00	5.13660E+00	**9.51544E-01**
		f4	1.23163E-01	6.09165E-02	1.43872E-01	**4.78862E-02**
		f5	3.85862E+00	2.27750E+00	5.33624E+00	**9.16860E-01**
	15	f1	1.04988E+04	1.28863E+04	1.75842E+04	**3.39275E+03**
		f2	1.78673E+04	2.02623E+04	2.03206E+04	**1.40315E+04**
		f3	6.18086E+00	6.02122E+00	1.14470E+01	**2.57817E+00**
		f4	1.55289E-01	1.88081E-01	2.89422E-01	**5.59937E-02**
		f5	6.20509E+00	7.62717E+00	9.81795E+00	**2.81262E+00**
	25	f1	5.12787E+04	5.12853E+04	7.76228E+04	**2.00738E+04**
		f2	1.49534E+04	1.50335E+04	5.96243E+04	**1.30396E+04**
		f3	2.86030E+01	2.93423E+01	4.54817E+01	**1.22363E+01**
		f4	7.09540E-01	9.05091E-01	1.22516E+00	**3.82671E-01**
		f5	3.35760E+01	3.36012E+01	4.90412E+01	**1.49388E+01**
75	10	f1	2.11220E+03	2.15361E+03	5.34109E+03	**1.21651E+03**
		f2	7.12371E+03	6.99224E+03	8.16321E+03	**4.82459E+03**
		f3	3.15013E+00	1.90286E+00	3.70678E+00	**1.62461E+00**
		f4	1.13649E-01	7.02846E-02	1.39706E-01	**6.38427E-02**
		f5	3.29564E+00	1.80078E+00	4.51394E+00	**2.47614E+00**

	15	f1	7.34733E+03	8.43899E+03	1.41247E+04	**3.39126E+03**
		f2	1.78875E+04	1.70372E+04	2.23337E+04	**1.32933E+04**
		f3	5.26664E+00	6.08052E+00	8.60735E+00	**1.77677E+00**
		f4	1.88610E-01	1.59751E-01	2.54270E-01	**4.61430E-02**
		f5	4.45377E+00	5.82035E+00	8.33757E+00	**2.04109E+00**
	25	f1	3.85996E+04	3.76240E+04	5.91306E+04	**1.42559E+04**
		f2	6.33172E+04	6.16264E+04	6.16239E+04	**1.54260E+04**
		f3	2.37094E+01	1.98317E+01	3.79563E+01	**1.00639E+01**
		f4	6.66511E-01	7.14074E-01	1.03177E+00	**2.71517E-01**
		f5	2.18925E+01	2.33039E+01	3.87337E+01	**1.03377E+01**
100	10	f1	**2.18834E+03**	2.49337E+03	5.13868E+03	2.76904E+03
		f2	5.61246E+03	5.65115E+03	6.70557E+03	**3.40020E+03**
		f3	2.74221E+00	**1.50622E+00**	3.55665E+00	2.41426E+00
		f4	1.12781E-01	**6.01845E-02**	1.18192E-01	9.23756E-02
		f5	3.48875E+00	**2.12554E+00**	4.08539E+00	3.12512E+00
	15	f1	5.92962E+03	6.27715E+03	1.25321E+04	**2.42506E+03**
		f2	1.93249E+04	1.56617E+04	1.87256E+04	**1.28186E+04**
		f3	5.31015E+00	6.18042E+00	8.41942E+00	**1.42551E+00**
		f4	1.96993E-01	1.37877E-01	3.16160E-01	**5.15856E-02**
		f5	5.09126E+00	4.98564E+00	7.43683E+00	**1.42958E+00**
	25	f1	3.01000E+04	3.26109E+04	4.91078E+04	**1.27530E+04**
		f2	6.43586E+04	1.39835E+04	7.37897E+04	**1.19835E+04**
		f3	1.97562E+01	2.39007E+01	3.66957E+01	**7.75062E+00**
		f4	5.12015E-01	7.07306E-01	9.03942E-01	**2.27478E-01**
		f5	2.12992E+01	2.25742E+01	3.13459E+01	**8.41840E+00**

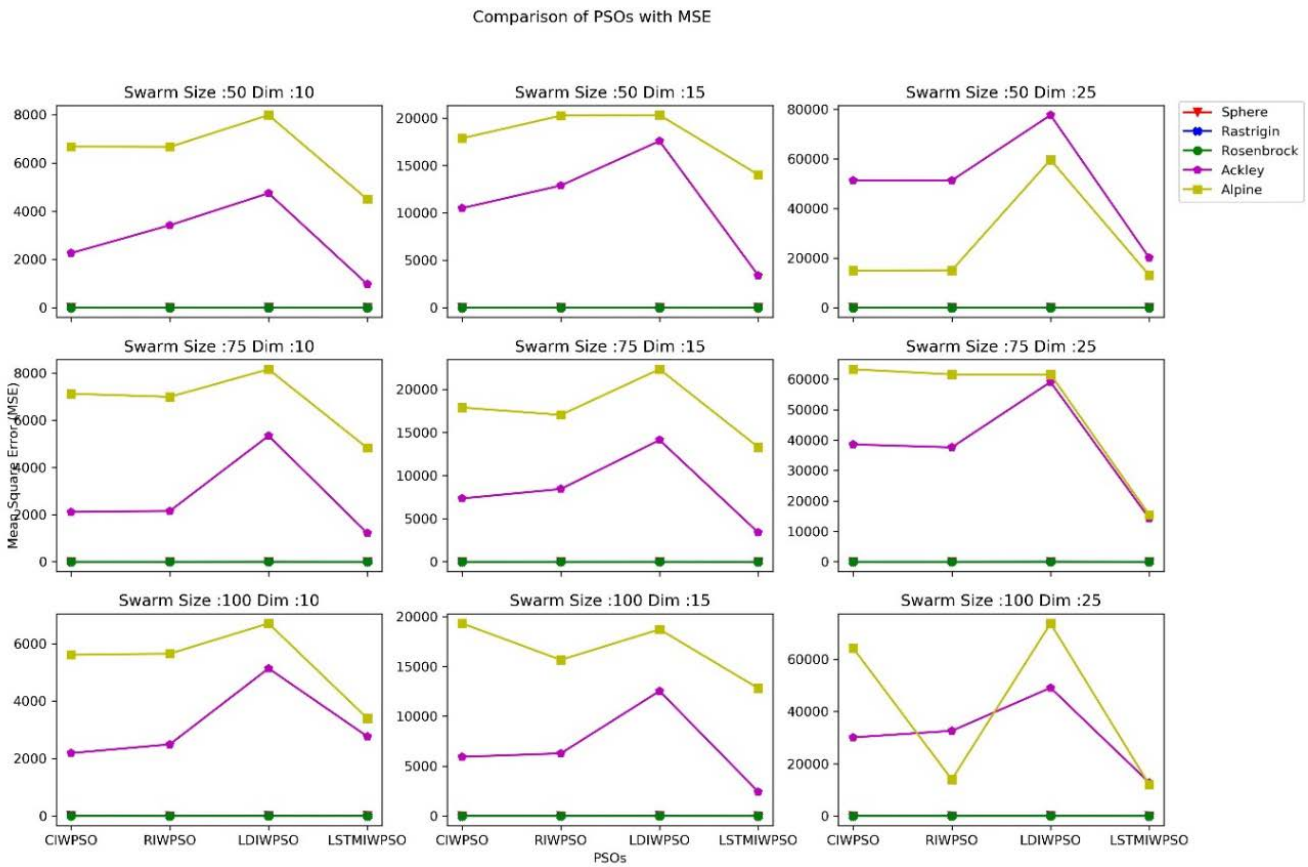


Fig. 8. MSE Computed for the Swarm Size of 50, 75, and 100 with Dimensions 10, 15 and 25.

TABLE VII. COMPUTED ROOT MEAN SQUARED ERROR (RMSE) FOR PSOS WITH RESPECT TO DIFFERENT SWARM SIZES AND DIMENSIONS. (FIG. 9)

Swarm Size	Dimension	BMF	PSOs			
			CIWPSO	RIWPSO	LDIWPSO	LSTMIWPSO
50	10	f1	47.547507	58.465408	68.892273	**31.160521**
		f2	81.715000	81.641530	89.368178	**67.021852**
		f3	1.279373	1.448675	2.266406	**0.975471**
		f4	0.350946	0.246813	0.379305	**0.218829**
		f5	1.964338	1.509139	2.310031	**0.957528**
	15	f1	102.463436	113.517652	132.605429	**58.247349**
		f2	133.668738	142.345836	142.550222	**118.454673**
		f3	2.486133	2.453817	3.383343	**1.605668**
		f4	0.394067	0.433683	0.537979	**0.236630**
		f5	2.491002	2.761733	3.133362	**1.677088**
	25	f1	226.447895	226.462645	278.608614	**141.682136**
		f2	122.283946	122.611075	244.180950	**114.190940**
		f3	5.348177	5.416850	6.744009	**3.498047**
		f4	0.842342	0.951363	1.106869	**0.618604**
		f5	5.794484	5.796650	7.002942	**3.865068**
75	10	f1	45.958712	46.406975	73.082779	**34.878568**
		f2	84.402064	83.619612	90.350480	**69.459280**
		f3	1.774860	1.379441	1.925301	**1.274603**
		f4	0.337119	0.265112	0.373773	**0.252671**
		f5	1.815388	**1.341929**	2.124604	1.573574
	15	f1	85.716591	91.863986	118.847212	**58.234483**
		f2	133.744088	130.526763	149.444504	**115.296759**
		f3	2.294917	2.465871	2.933828	**1.332957**
		f4	0.434293	0.399688	0.504252	**0.214809**
		f5	2.110395	2.412541	2.887486	**1.428668**
	25	f1	196.467698	193.969044	243.167928	**119.398100**
		f2	251.628992	248.246634	248.241652	**124.201480**
		f3	4.869227	4.453277	6.160870	**3.172369**
		f4	0.816401	0.845029	1.015762	**0.521073**
		f5	4.678940	4.827408	6.223642	**3.215230**
100	10	f1	**46.779695**	49.933663	71.684569	52.621656
		f2	74.916360	75.174099	81.887555	**58.311259**
		f3	1.655961	**1.227283**	1.885909	1.553789
		f4	0.335829	**0.245325**	0.343790	0.303934
		f5	1.867821	**1.457922**	2.021234	1.767802
	15	f1	77.004050	79.228452	111.947053	**49.244864**
		f2	139.013944	125.146853	136.841670	**113.219355**
		f3	2.304377	2.486044	2.901624	**1.193948**
		f4	0.443839	0.371318	0.562281	**0.227125**
		f5	2.256382	2.232855	2.727056	**1.195651**
	25	f1	173.493486	180.584992	221.602772	**112.929011**
		f2	253.690037	118.251843	271.642611	**109.469148**
		f3	4.444791	4.888839	6.057695	**2.783994**
		f4	0.715552	0.841015	0.950759	**0.476947**
		f5	4.615105	4.751235	5.598742	**2.901447**

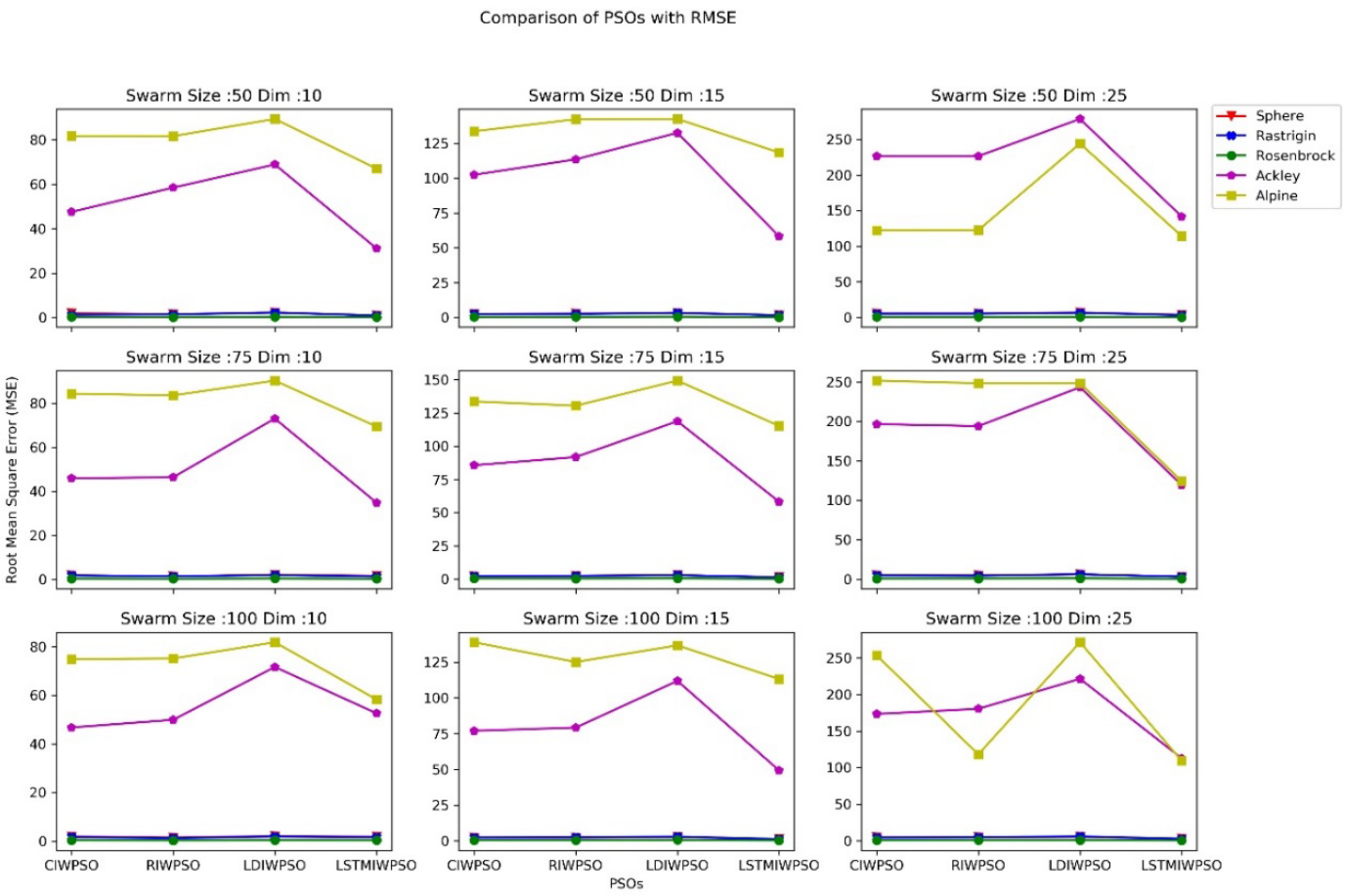


Fig. 9. RMSE Computed for the Swarm size of 50, 75, and 100 with Dimensions 10, 15 and 25.

TABLE VIII. COMPUTED MEAN TIME (IN SECONDS) FOR PSOS WITH RESPECT TO DIFFERENT SWARM SIZES AND DIMENSIONS. (FIG. 10)

Swarm Size	Dimension	BMF	PSOs			
			CIWPSO	RIWPSO	LDIWPSO	LSTMIWPSO
50	10	f1	3.312812	**3.183287**	3.589538	7.459693
		f2	0.021386	0.025384	**0.020334**	0.053375
		f3	2.257866	3.293823	1.**886251**	5.159045
		f4	1.**138227**	2.282051	1.398866	2.658974
		f5	1.**624259**	3.076825	2.082976	4.621895
	15	f1	4.306329	**4.294003**	4.869144	9.409708
		f2	0.032513	0.035312	**0.026717**	0.089734
		f3	4.**308127**	4.373820	5.000313	9.357706
		f4	4.**368624**	4.424788	4.412930	9.501312
		f5	4.**314457**	4.375219	4.511282	9.428390
	25	f1	6.178701	**6.145023**	6.274644	13.369280
		f2	0.461714	0.442326	**0.055366**	1.930295
		f3	6.135661	6.**082227**	6.220433	13.182448
		f4	6.**006940**	6.069369	6.115273	13.603162
		f5	6.129612	6.086691	6.**078426**	13.568970
75	10	f1	3.273303	4.050883	2.**842036**	5.982660
		f2	0.026450	0.026450	0.**022253**	0.059991
		f3	1.**855718**	3.775724	2.897390	2.032672
		f4	1.**208450**	2.102096	2.027942	1.292319
		f5	1.956389	2.852364	2.153064	**1.722074**

	15	f1	6.526218	6.**518706**	6.866313	13.257469
		f2	0.041508	0.038976	**0.030914**	0.117080
		f3	6.**276307**	6.611565	6.669305	13.423912
		f4	4.**794559**	6.383440	5.537492	12.628298
		f5	6.**380909**	6.439272	6.652887	13.195447
	25	f1	8.**987696**	9.227687	9.441670	19.331790
		f2	0.073887	0.077486	**0.063161**	1.549432
		f3	9.**008012**	9.178240	9.021073	19.068662
		f4	9.058448	9.075037	8.**982564**	19.274107
		f5	9.163049	9.168713	9.**048472**	19.067931
100	10	f1	3.016330	4.171748	3.530148	2.**414135**
		f2	0.032180	0.033646	0.**028782**	0.058631
		f3	1.819471	3.799044	2.733838	1.**586756**
		f4	1.574290	2.531497	2.621043	1.**241292**
		f5	2.020080	3.923700	2.951757	1.**423464**
	15	f1	9.002749	8.**977202**	9.125075	17.114905
		f2	0.**046437**	0.049836	0.046704	0.112377
		f3	7.**197202**	8.082389	7.425860	18.320117
		f4	4.**608342**	8.081904	5.849772	13.539670
		f5	7.656653	**7.275687**	8.531872	17.034604
	25	f1	12.429958	**12.134209**	12.552006	24.936847
		f2	0.091344	0.841345	**0.070490**	3.426473
		f3	12.091566	12.138803	**12.043132**	25.225323
		f4	12.340614	13.969420	**12.167084**	27.915432
		f5	**11.938530**	12.319092	12.264439	24.890808

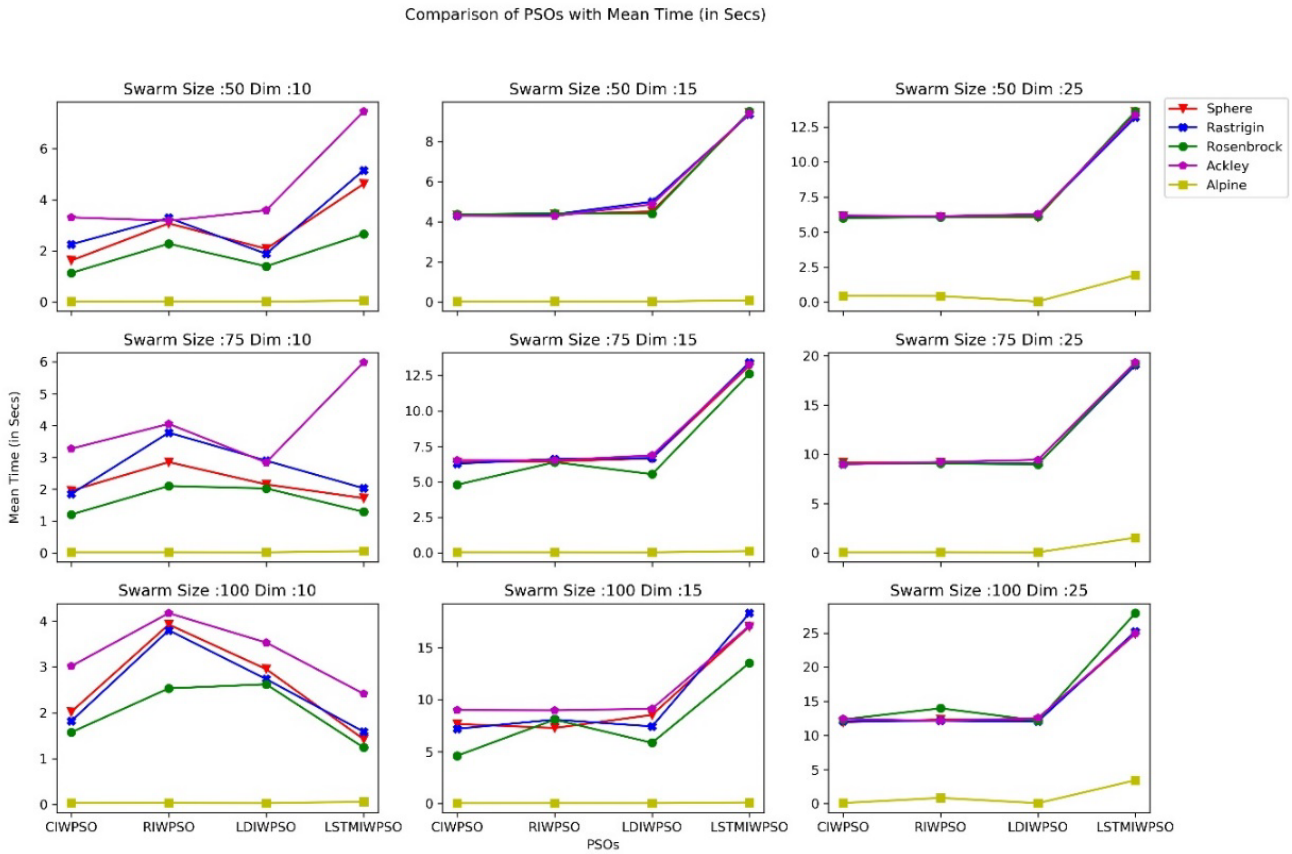


Fig. 10. Mean Time (In Secs) Computed for the Swarm Size of 50, 75, and 100 with Dimensions 10, 15 and 25.

TABLE IX. MEAN ITERATIONS FOR PSOS WITH RESPECT TO DIFFERENT SWARM SIZES AND DIMENSIONS. (FIG. 11)

Swarm Size	Dimension	BMF	PSOs			
			CIWPSO	RIWPSO	LDIWPSO	LSTMIWPSO
50	10	f1	954.67	907.13	**864.67**	1000.00
		f2	5.27	4.80	**4.33**	6.93
		f3	652.13	942.40	**543.67**	694.47
		f4	**306.33**	617.27	411.20	353.20
		f5	468.20	842.73	584.67	615.20
	15	f1	**1000.00**	**1000.00**	**1000.00**	**1000.00**
		f2	6.60	7.67	**5.80**	9.33
		f3	**1000.00**	**1000.00**	**1000.00**	**1000.00**
		f4	1000.00	**989.87**	1000.00	1000.00
		f5	**983.93**	**1000.00**	**1000.00**	**1000.00**
	25	f1	**1000.00**	**1000.00**	**1000.00**	**1000.00**
		f2	73.80	75.53	**8.53**	148.40
		f3	**1000.00**	**1000.00**	**1000.00**	**1000.00**
		f4	**1000.00**	**1000.00**	**1000.00**	**1000.00**
		f5	**1000.00**	**1000.00**	**1000.00**	**1000.00**
75	10	f1	623.13	791.27	**536.87**	579.20
		f2	4.07	4.33	**3.87**	5.53
		f3	363.33	731.40	491.60	**198.40**
		f4	231.93	409.73	395.33	**118.47**
		f5	378.07	545.87	417.47	**168.73**
	15	f1	**1000.00**	**1000.00**	**1000.00**	**1000.00**
		f2	5.20	5.47	**4.53**	8.33
		f3	**974.93**	991.07	1000.00	1000.00
		f4	**729.13**	982.33	851.73	957.67
		f5	**981.60**	992.33	1000.00	1000.00
	25	f1	1000.00	1000.00	1000.00	1000.00
		f2	7.07	6.87	**6.73**	80.47
		f3	**1000.00**	**1000.00**	**1000.00**	**1000.00**
		f4	**1000.00**	**1000.00**	**1000.00**	**1000.00**
		f5	**1000.00**	**1000.00**	**1000.00**	**1000.00**
100	10	f1	442.47	595.20	481.87	**178.07**
		f2	4.33	4.33	3.67	4.20
		f3	265.33	545.87	402.27	120.67
		f4	217.93	360.47	379.67	94.00
		f5	296.67	551.93	406.20	**108.67**
	15	f1	**1000.00**	**1000.00**	**1000.00**	**1000.00**
		f2	5.00	5.33	**4.93**	6.40
		f3	**835.07**	932.47	854.53	1000.00
		f4	**533.40**	937.13	667.60	698.20
		f5	892.40	**836.73**	856.60	1000.00
	25	f1	**1000.00**	**1000.00**	**1000.00**	**1000.00**
		f2	**6.87**	72.93	5.47	140.33
		f3	**1000.00**	**1000.00**	**1000.00**	**1000.00**
		f4	**1000.00**	**1000.00**	**1000.00**	**1000.00**
		f5	**1000.00**	**1000.00**	**1000.00**	**1000.00**

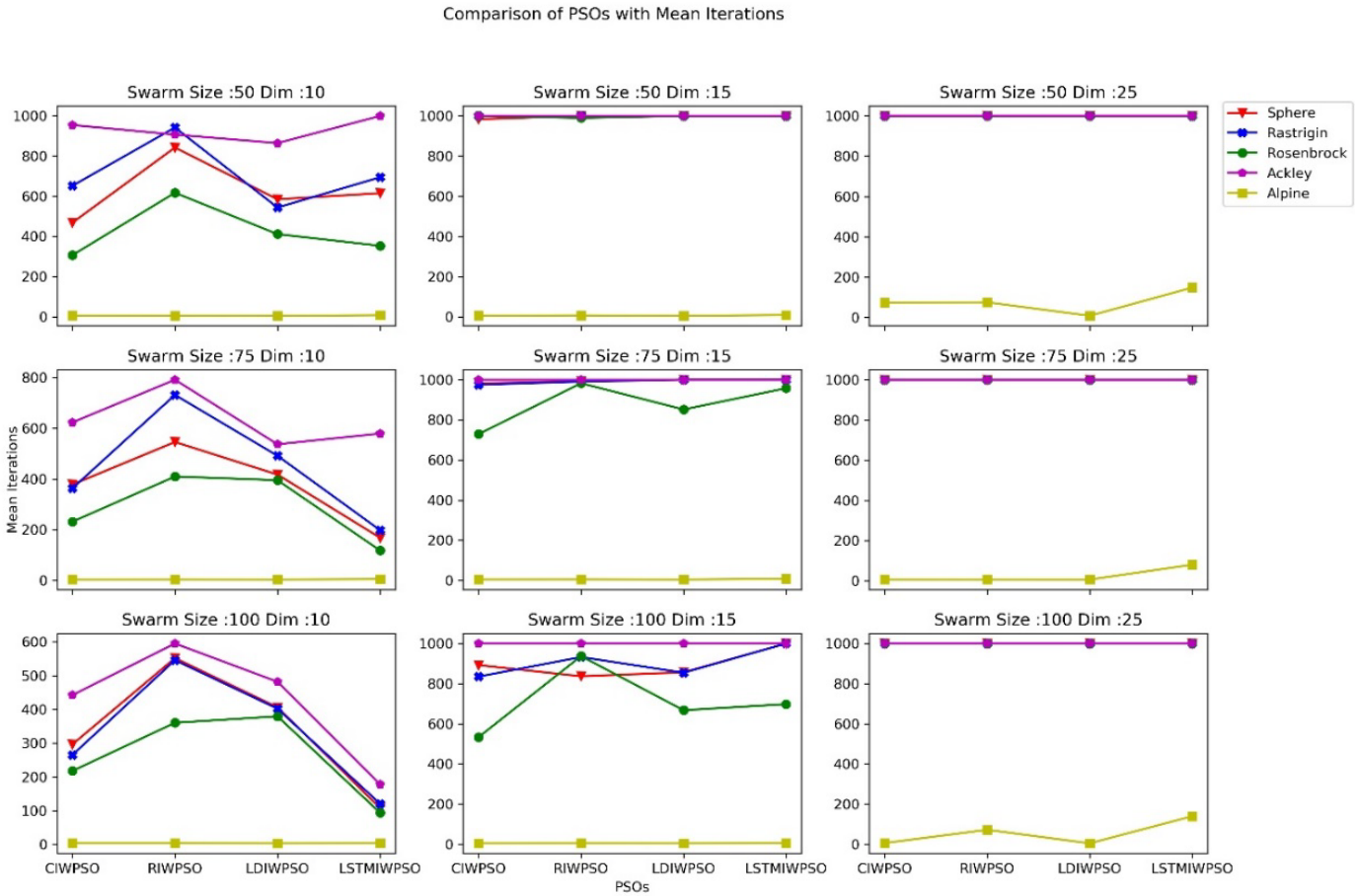


Fig. 11. Mean Iterations Computed for the Swarm size of 50, 75, and 100 with Dimensions 10, 15 and 25.

V. Conclusion and Future Work

In this paper, a new inertia weight based PSO using LSTM (LSTMIWPSO) is presented. A set of 5 most common optimization test problems and eight criteria are considered to assess the performance of LSTMIWPSO against CIWPSO, RIWPSO, and LDIWPSO. The overall outcome shows that LSTMIWPSO is progressive with CIWPSO, RIWPSO, and LDIWPSO. In the future, the parameters of LSTM are tuned to enhance efficiency. Also, more experiments with larger swarm sizes and dimensions are conducted to evaluate LSTMIWPSO performance with other existing inertia weight based PSO. There is a scope for the use of LSTMIWPSO in the optimization of the different optimization applications without any restriction of the domains specified.

References

[1] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in International Conference on Neural Networks, vol. 4. IEEE, 1995, pp. 1942–1948.

[2] R. Eberhart and J. Kennedy, “A new optimizer using particle swarm theory,” in International Symposium on Micro Machine and Human Science. IEEE, 1995, pp. 39–43.

[3] C. M. Huang, C. J. Huang and M. L. Wang, A Particle Swarm Optimization to Identifying the ARMAX Model for Short-Term Load Forecasting. IEEE Tran. on Power Systems, 20(2): 1126-1133, May 2005.

[4] X. H. Hu, R. C. Eberhart, Y. H. Shi, Engineering optimization. with particle swarm. Proc. of the IEEE Swarm Intelligence Symposium, pages 53-57, 2003.

[5] X. H. Hu, Y. H. Shi, R. Eberhart, Recent advances in particle swarm. Congress on Evolutionary Computation, pages 90-97, 2004.

[6] M. Clerc, Particle Swarm Optimization, ISTE Publishing Company, 2006.

[7] N. Nedjah, L. D. M. Mourelle, Systems Engineering Using Particle Swarm Optimization, Nova Science Publishers, 2007.

[8] Yudong Zhang, Shuihua Wang, and Genlin Ji, A Comprehensive Particle Swarm Optimization Algorithm and Its Applications, Mathematical Problems in Engineering, Volume-2015, pp 1-38.

[9] Ashok Kumar, Brajesh Kumar Singh, B.D.K.Patro, Particle Swarm Optimization: A Study of Variants and their applications, International Journal of Computer Applications, Volume 135 (5), November 2016, PP 24-30.

[10] Y. Zhang, S. Wang, G. Ji, A comprehensive survey on particle swarm optimization algorithm and its applications, Math. Probl. Eng. (2015).

[11] M. R. Bonyadi and Z. Michalewicz, "Particle swarm optimization for single-objective continuous space problems: a review," Evolutionary computation, 2016.

[12] J.C. Bansal, P.K. Singh, M. Saraswat, A. Verma, S.S. Jadon, A. Abraham, Inertia weight strategies in particle swarm optimization, in: 2011 Third World Congress on Nature and Biologically Inspired Computing (NaBIC), 2011, pp.633–640.

[13] Y. Shi, R.C. Eberhart, A modified particle swarm optimizer, in: Proceeding of the IEEE Congress on Evolutionary Computation (CEC), IEEE, Piscataway, NJ, USA, 1998, pp. 69–73. Anchorange, AC, USA.

[14] R.C. Eberhart and Y. Shi., “Tracking and optimizing dynamic systems with particle swarms”, In Evolutionary Computation, 2001. Proceedings of the 2001 Congress on, volume 1, pages 94–100. IEEE, 2002.

[15] J. Xin, G. Chen, and Y. Hai., “A Particle Swarm Optimizer with Multistage Linearly-Decreasing Inertia Weight”, In Computational Sciences and Optimization, 2009. CSO 2009. International Joint Conference on, volume 1, pages 505–508. IEEE, 2009.

[16] Serani, Andrea & Leotardi, Cecilia & Iemma, Umberto & Campana, Emilio & Fasano, Giovanni & Diez, Matteo. (2016). Parameter selection in synchronous and asynchronous deterministic particle swarm optimization for ship hydrodynamics problems. Applied Soft Computing. 49. 313-334. 10.1016/j.asoc.2016.08.028.

[17] M.R. Bonyadi, Z. Michalewicz, Impacts of coefficients on movement patterns in the Particle Swarm Optimization algorithm, IEEE Trans. Evol. Comput. 21 (3) (2017) 378–390.

[18] Andries P. Engelbrecht, Computational Intelligence: An Introduction, 2nd Edition, Wiley Publications, West Sussex, England, 2007.

[19] Riccardo Poli, "Analysis of the Publications on the Applications of Particle Swarm Optimisation", Journal of Artificial Evolution and Applications, Vol.2008, January, 2018, 10 Pages.

[20] E. T. Oldewage, A. P. Engelbrecht and C. W. Cleghorn, The merits of velocity clamping particle swarm optimisation in high dimensional spaces, 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, 2017, pp. 1-8, doi: 10.1109/SSCI.2017.8280887.

[21] Frans van den Bergh. An Analysis of Particle Swarm Optimizers. PhD thesis, University of Pretoria, 2001.

[22] Federico Marini, Beata Walczak, Particle swarm optimization (PSO). A tutorial, Chemometrics and Intelligent Laboratory Systems, Volume 149, Part B, 2015, Pages 153-165, ISSN 0169-7439, https://doi.org/10.1016/j.chemolab.2015.08.020.

[23] Barbara Hammer, On the approximation capability of recurrent neural networks, Neurocomputing, Volume 31, Issues 1–4, 2000, Pages 107-123, ISSN 0925-2312, https://doi.org/10.1016/S0925-2312(99)00174-5.

[24] Afshine Amidi, Shervine Amidi, Recurrent Neural Network Cheatsheet, https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks#architecture, Last accessed on 17.10.2020.

[25] Y. Bengio, P. Simard and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," in IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157-166, March 1994, doi: 10.1109/72.279181.

[26] Mikolov, Tomas & Joulin, Armand & Chopra, Sumit & Mathieu, Michael & Ranzato, Marc'Aurelio. (2014). Learning Longer Memory in Recurrent Neural Networks.

[27] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

[28] Bewoor, L.A., Chandra Prakash, V., Sapkal, S.U., Evolutionary hybrid particle swarm optimization algorithm for solving NP-hard no-wait flow shop scheduling problems, Algorithms, Vol: 10, Issue: 4, pp: - , 2017, DoI :10.3390/a10040121.

[29] Lakshmi Prasad, M., Sastry, J.K.R., Building test cases by particle swarm optimization (PSO) for multi output domain embedded systems using combinatorial techniques, Journal of Advanced Research in Dynamical and Control Systems, Vol: 10, Issue: 6, pp: 1221 - 1229, 2018, DoI.

[30] Bewoor, L.A., Prakash, V.C., Sapkal, S.U., Production scheduling optimization in foundry using hybrid Particle Swarm Optimization algorithm, Procedia Manufacturing, Vol: 22, Issue: , pp: 57 - 64, 2018, DoI :10.1016/j.promfg.2018.03.010.

[31] Bewoor, L.A., Chandra Prakash, V., Sapkal, S.U., An improved evolutionary hybrid particle swarm optimization algorithm to minimize makespan for no wait flow shop scheduling, Journal of Theoretical and Applied Information Technology, Vol: 96, Issue: 9, pp: 2534 - 2545, 2018, DoI :.

[32] Prasanna Kumar, K., Kishore, M.G.V., Hemanth, K.V., Sreekar, L., Synthesis of antenna array using modified particle swarm optimization technique, International Journal of Innovative Technology and Exploring Engineering, Vol: 8, Issue: 5, pp: 1 - 5, 2019, DoI :.

[33] Ahilan, A., Manogaran, G., Raja, C., Kadry, S., Kumar, S.N., Agees Kumar, C., Jarin, T., Krishnamoorthy, S., Malarvizhi Kumar, P., Chandra Babu, G., Senthil Murugan, N., Parthasarathy, Segmentation by Fractional Order Darwinian Particle Swarm Optimization Based Multilevel Thresholding and Improved Lossless Prediction Based Compression Algorithm for Medical Images, IEEE Access, Vol: 7, Issue: , pp: 89570 - 89580, 2019, DoI :10.1109/ACCESS.2019.2891632.

[34] Senthil Kumar, A.M., Parthiban, K., Siva Shankar, S., An efficient task scheduling in a cloud computing environment using hybrid Genetic Algorithm - Particle Swarm Optimization (GA-PSO) algorithm, Proceedings of the International Conference on Intelligent Sustainable Systems, ICISS 2019, Vol: , Issue: , pp: 29 - 34, 2019, DoI :10.1109/ISS1.2019.8908041.

[35] Bewoor, L.A., Chandraprakash, V., Sapkal, S.U., Evolutionary hybrid particle swarm optimization algorithm to minimize makespan toschedule a flow shop with no wait, Journal of Engineering Science and Technology, Vol: 14, Issue: 2, pp: 609 - 628, 2019, DoI :.

[36] Kumar, V.H., Varma, P.S., Kumar, T.B., Sreelatha, E., Economic and emission dispatch problem using particle swarm optimization, International Journal of Innovative Technology and Exploring Engineering, Vol: 8, Issue: 6, pp: 939 - 944, 2019, DoI :.

[37] Patro, P., Kumar, K., Suresh Kumar, G., Neuro fuzzy system with hybrid ant colony particle swarm optimization (Haso) and robust activation, Journal of Advanced Research in Dynamical and Control Systems, Vol: 12, Issue: 3 Special Issue, pp: 741 - 750, 2020, DoI :10.5373/JARDCS/V12SP3/20201312.v

[38] Mehrotra, S., Sharan, A., Comparative Analysis of K-Means Algorithm and Particle Swarm Optimization for Search Result Clustering, Smart Innovation, Systems and Technologies, Vol: 165, Issue: , pp: 109 - 114, 2020, DoI :10.1007/978-981-15-0077-0_12.

[39] Nagesh, P., Srinivasu, N., Ranganadh, N.S.R., Jayanth, J., Naveen, A., Optimal vm placement using particle swarm optimization, International Journal of Scientific and Technology Research, Vol: 9, Issue: 4, pp: 611 - 615, 2020, DoI :.

[40] Syamala Rao, P., Parthasaradhi Varma, G., Durga Prasad, Ch., Identification of linear and non linear curve fitting models using particle swarm optimization algorithm, AIP Conference Proceedings, Vol: 2269, Issue: , pp: - , 2020, DoI :10.1063/5.0019657.

Identifying the Impacts of Active and Passive Attacks on Network Layer in a Mobile Ad-hoc Network: A Simulation Perspective

Uthumansa Ahamed[1]
Department of Physical Sciences
Faculty of Applied Sciences
Rajarata University of Sri Lanka
Mihintale, Sri Lanka

Shantha Fernando[2]
Department of Computer Science
& Engineering, Faculty of Engineering
University of Moratuwa
Moratuwa, Sri Lanka

***Abstract*—In this research, we attempted to investigate about the features and behaviors of network layer based active and passive attacks in Ad-hoc On-Demand Vector (AODV) routing protocol in Mobile Ad-hoc Networks (MANET). Through the literature survey, we try to understand the features of each attacks and examine the behaviors of these attacks through simulations via Network Simulator 2 (NS2). Blackhole, Grayhole and Wormhole attacks are used in this simulation study. Each attacks are introduced independently into the network to find the impacts on network performances that are evaluated through Packet Delivery Ratio (PDR), Average End-to-End Delay (AEED), Throughput, Average Data Dropping Rate (ADDR) and Simulation Processing Time at Intermediate Nodes (SPTIN). To obtain more accurate results, simulation parameters are maintained same in each simulation. A controller network is simulated to compare with each attack simulation. Simulations are repeated by changing the number of connected intermediate nodes (hops) in the network. We observed at collected data analysis, the lowest SPTIN in the network that contained a Blackhole or Grayhole attack out of these three attacks. The network which is affected by a Blackhole attack shows higher amount of ADDR than controller network. Furthermore data forwarding rate is higher in the network which is affected by a Wormhole attack. Finally, according to the simulation studies, we are able to understand that Blackhole and Grayhole attacks cause more damage to the network performances than Wormhole attacks.**

***Keywords*—Active attack; network layer; passive attack; performance matrices; simulation study**

I. Introduction

4G technology allows devices to communicate each other through wireless medium even at the absence of predefined infrastructure networks. These wireless devices are capable of making connections between themselves when they are capable to listen to one-another. This type of network is called an Ad Hoc Network [1]. MANET is a type of wireless ad hoc network. In MANET, each device is defined as a node and fundamental feature of a MANET is node mobility. Mainly three different types of nodes are available in MANET. Those are source node, destination node and routing (intermediate) node. In pure MANET paradigm, there is no fixed infrastructure. Usually nodes can join and leave from the network without any constrain. Therefore, network topology changes eventually. Another important feature of a node in MANET is limited radio range which leads to depend on the help of neighbouring nodes (multi-hop) to communicate with destination node when source and destination nodes are not in their radio range. Furthermore, nodes use Open Systems Interconnection (OSI) model standards to communicate among them. Because of multi-hop nature nodes often act not only as hosts but also as routers. In a multi-hop network one or more intermediate nodes are possible to connect in a route between source and destination [1]–[3]. This opportunistic nature of MANET made attention to use on military and rescue agencies particularly under disorganized or hostile environments where services of infrastructure networks are unavailable because of disaster situations [2]. Relatively low cost on network deployment made MANET a more common and smart alternative even for commercial uses such as virtual classrooms [3].

Open network boundary, infrastructure-less nature and dynamic network topology are some fundamental characters of a MANET. These characters expose MANET into different types of security attacks on each OSI layers and routing protocols are operated on Network layer of OSI model. Routing protocols in the MANET can be categorized into Two: Proactive and Reactive. Proactive routing protocols need to maintain routing information even if there are no demand for a communication [4]–[6]. These types of protocols are only suitable for stationary and fewer number of nodes [7]. Reactive routing protocols start to find a route to destination only for a demand. There are no any routing details at the beginning [8]. Hybrid routing protocols are combination of proactive and reactive routing protocols. Still it has the limitation of Proactive routing protocols [7], [9].

Furthermore, most of the routing protocols (e.g.: AODV) are working based on the trustworthiness of each node. “All nodes are reliable” is the main assumption of pure routing protocols [8]. Therefore, attacking node can be a part of a MANET easily. Most routing protocols perform well, but fail to address the network security. Therefore security attacks are needed to be addressed to protect the network performances during the data communication through the network. This research is aimed to identify the impacts of active and passive attacks on network layer in a MANET. Outcomes of this research will be help to re-design routing protocol with an adaptive model to handle network layer attacks in MANET.

The rest of this paper is organized as follows. State of

the art of the network layer attacks is presented on Section 2. Different network layer attacks on AODV protocol are described in Section 3. Our research methodology explains in Section 4. In Section 5, we present our simulation results and discussion. Summary of analysis and discussion is included in Section 6. Finally, Section 7 explains about the conclusion and future works.

II. State-of-the-Art

Authors in [10] proposed classifications of different security attacks on a MANET. It is helpful for better understanding of each attack. The affects of Wormhole attack on few routing protocols were summarized by authors in [11]. Furthermore, detailed comparative analysis on detection and prevention techniques of Wormhole attack is presented in the same research, though this study does not carryout any simulation study. Authors in [12] presented a study on Wormhole attack prevention techniques and a simulation study on Wormhole attack on AODV and DSR routing protocols and few simulation results are unclear about the Wormhole attack which is applied on AODV or DSR. Authors in [13] discussed about state of the art on prevention mechanisms of Blackhole attack. In [14] surveyed some of the existing solutions for Blackhole, Grayhole and Wormhole attacks. Authors in [15] simulated four different types of routing attacks (Active attacks) but their simulation results contradict with their Blackhole attack definition. According to the definition in their research paper, Blackhole attack drops all the packet what it receives except Routing Request (RREQ) packets even though their simulation results show considerable amount of data transaction in the presence of Blackhole attack. Authors in [16] conducted a study on few network layer attacks and routing protocols. Finally, they suggested some solutions for routing protocols to overcome network layer attacks through the literature survey. In [17] they presented a survey of significant network layer attacks and review of intrusion detection mechanisms that have been proposed in the literature. Authors in [18] reviewed mitigation of various routing attacks and prevention on these attacks. Authors in [19] presented a study on Blackhole attack through inducting malicious node activity in AODV under different scenarios. In [20] they investigated some security issues in MANET as well as countermeasures against such attacks in existing MANET protocols.

Most of these studies rely only on the theoretical findings. Therefore, it is difficult to identify the impact of active and passive attacks on MANET. Furthermore, it is difficult to understand the impact of each attack separately on the network performances. Therefore it is important to conduct a simulation study on attacks in order to identify the impacts of active and passive attack to propose a better countermeasure on network layer attacks.

III. Network Layer Attacks on AODV

The primary function of the network layer is routing [21]. In MANET most of the attacks are delivered after accessing the routing information [22]. The followings are some examples for network layer attacks [20], [23].

1) Blackhole Attack
2) Grayhole Attack
3) Wormhole Attack
4) Routing Table Overflow
5) Byzantine Attack
6) Link Spoofing Attack

In Blackhole attack malicious node which is the originator of the attack sends reply having destination sequence number in maximum possible value and hop count in minimum value during route discovery. Then source establishes a path including malicious node as router in it. In this path, destination may be found or may not. Then all the traffics will be redirected by the malicious node. Moreover, the route established by the malicious node starts to drop rather than deliver or retransmit when it receives data. It is possible to appear one or more malicious nodes in a route [20], [24].

Grayhole attack can be described as an extension of Blackhole attack. Attacking node behaves as a genuine node as well as a malicious node. During a communication between source node and destination node, attacking node acts as a genuine node by delivering or retransmitting what is received. To some period it drops all packets that it receives. In some other cases, attacker node drops data packets from a specific node and forward or retransmit data packets from other nodes.

Minimum two or more nodes are involved in a Wormhole attack. A private link (called as Wormhole tunnel) is established in-between these malicious nodes. These nodes may get themselves involved in more routes. Imitate with shortest path to source node during route discovery. When they become a router in a route, start to exploit data packet that they received. Wormhole nodes can drop, modify or send data packets to third party for malicious purpose. It is difficult to detect because of its cooperative nature [25], [26].

Furthermore, Blackhole and Grayhole attacks are categorized as Active attacks which disrupt the network performances and collapse the network. Wormhole attack is categorized as Passive attack which do not harm the network performances but collects or steals data which are needed to form an Active attack from the network. Rather than most of the network layer attacks listed above, Blackhole, Grayhole and Wormhole attacks (Hole attacks) behave relatively in similar manner. Following are some similarities between these Hole attacks.

- All of these attacks are network layer oriented [16], [20].
- Deliver false details during the routing discovery [16], [20], [24].
- Each attack intentionally drops packet during the session [16], [20], [23].
- During the session each attack shows misbehaving activity [14], [16], [20].
- During the attack, single node involves to alter data packets [14], [20].
- Routing protocol mislead by each attack [14], [16], [20], [24].
- Each attack advertises fake route during routing discovery [16], [20].

TABLE I. DIFFERENCES BETWEEN ATTACKS

Feature	Blackhole Attack	Grayhole Attack	Wormhole Attack
No of nodes need to form an attack	One node [20], [24]	One node [14]–[17], [25]	Two nodes [12], [20] [23], [25]
Attack type	Active [13], [23], [24]	Active [14]–[17], [25]	Passive [25]
Ability to communicate with destination node	Can not [13], [24]	Can [14], [16] [25], [27]	Can [12], [20], [25]
Attacker position	Part of the network [13], [20]	Part of the network [14]–[17], [25]	Both in the same network or different networks [25]
Data in RREP packets	False data [13], [20] [24], [26]	True data [14], [25], [26]	False or true data [28]
Data packet forwarding and transmission	Drops all data packets that that it receives [13], [23], [24]	Drops only selected data packets or drop only data packets from precise node [14], [16], [25]	Eventually drop data packets or forward or retransmit as normal node [28]
Network performance	Entire network will collapse [13], [20], [24]	will be reduced or network collapse after some period [14]–[16] [24], [25]	will be reduced but network will not collapse [28]

Table I shows differences and unique features between these attacks. Because of these unique features of each attack, they differ from one another.

IV. METHODOLOGY

During the research, NS2 is used as the test bed to simulate different scenarios. Table II shows simulation parameters maintained in NS2 during the simulations. For more accuracy, readings are recorded by changing the number of connected intermediate nodes (10, 15, 20, 25 and 30) at each simulation and different attacks are introduced into the network to check the impacts on the Performance Matrixes (PM): PDR, AEED and Throughput. In addition, performance on ADDR and SPTIN also have studied to understand more about each attacks. The attacks are injected by modifying the AODV routing protocol.

Each attack is introduced to the network individually to check the impacts separately. Furthermore, a network without any attack is simulated as the controller. Impact on each attacks are analyzed with respect to the controller. Recorded data are analyzed in Tracegraph 2.02 and visualized in Microsoft Office Excel 2007. During the simulations following assumptions are considered.

- All nodes were considered to be identical in software and hardware configurations.
- All the nodes except malicious nodes show no any malicious behavior during the data communication.

TABLE II. NS2 SIMULATION PARAMETERS

Simulation parameter	Value
Simulator	NS2 (v.2.34)
Number of nodes	10, 15, 20, 25, 30
Transmitter range	250 m
Bandwidth	2.0 x 10^6 bps
Frequency	9.14 x 10^8 Hz
Antenna/OmniAntenna X, Y, Z	0, 0, 1.5 m
Traffic type	Constant bit rate (CBR)
Radio-propagation model	TwoRayGround
Network interface type	Phy/WirelessPhy
Routing protocol	AODV
Max packets in Interface Queue	50
Time of simulation	5 s
Mobility model	None

V. SIMULATION RESULTS AND DISCUSSION

In order to investigate on the impacts of each Hole attacks, network performances evaluate with PDR, EED and Throughput. Furthermore, test result analysis with another two different parameters: ADDR; SPTIN.

A. PDR

PDR is a ratio between successfully received packets and total number of packets sent by the sender. PDR then is multiplied by 100 to obtain as a percentage [29]. A graph is plotted by using simulation results for PDR vs. number of connected nodes in the network. It is illustrated in Fig. 1.

$$PDR = \frac{Successfully\ received\ packets}{Total\ number\ of\ sent\ packets} \times 100 \quad (1)$$

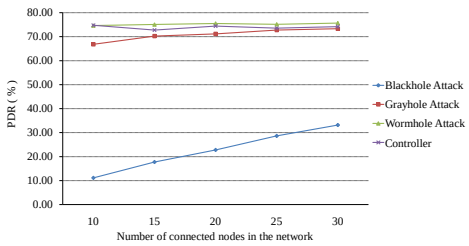


Fig. 1. Graph of PDR vs. Number of Connected Nodes in the Network.

According to the simulated results, controller network shows 73.75% of average PDR rate. The network with Wormhole attack shows a higher average PDR value, which is 75.19%. The reason for this is the communication between the attacker nodes via a Wormhole tunnel. This shows higher data flow through Wormhole tunnel. The network which is affected by Grayhole attack shows 70.85% of PDR value. The network which is affected by Blackhole attack shows lower PDR. The average PDR value is 22.69%. This is because attacker node drops the entire packets that it receives except AODV routing packets. But when the number of connected nodes in the

network increases, PDR value also increases. This is because when the number of connected nodes in the network increases, routing protocol is sending to find the route by sending more RREQ packets.

B. AEED

$$AEED = \frac{Total\ no\ of\ packets\ reached\ by\ destination\ node}{Total\ time\ taken\ to\ receive\ all\ packets\ by\ destination\ node} \quad (2)$$

End-to-End Delay is an amount of the time which is taken by a packet to reach destination node from the source node. Unit is seconds [20], [29]. Fig. 2 shows the graph which is plotted for AEED vs. number of connected nodes in the network. Controller network shows higher AEED value and when the number of nodes increases the value gradually increases. This is because when the number of nodes increases in the network, data packets need to pass through more intermediate nodes to reach the destination node. Average value of a given number of nodes in the network is used to plot the graph in Fig. 2. Furthermore, mean value of each AEED is 0.096634435696 seconds.

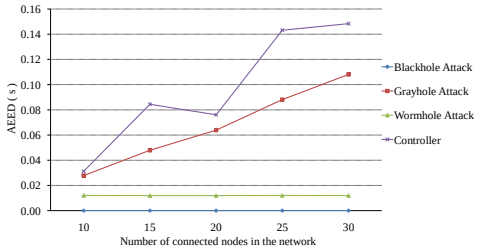


Fig. 2. Graph of AEED vs. Number of Connected Nodes in the Network.

The lowest AEED value is shown in the presence of Blackhole attack because there are no any data packets available to reach at the destination node. All the data packets are dropped by the attacker node in the network. Therefore, AEED value is not available. This is assumed as 0 for each number of nodes in the network for the graphing purpose. AEED value of the network which is affected by Grayhole attacks is lower than the value of the controller network. When the number of nodes increases in the network, AEED value increases gradually. However, it is still lower than the AEED value of controller network. Data packets communication speed through this network is 1.43 times faster than the speed in controller network. The mean value of all AEED of the network which is affected by Grayhole attack is 0.067165942022 seconds.

In the presence of a Wormhole attack data forwarding speed is abnormally faster than the controller. It is because of the faster connection in between two Wormhole nodes. Therefore, this network shows the lowest AEED value. According to the simulation results, the mean value of all AEED value is 0.012056294606 seconds. Therefore, data transferring speed is 8 times faster than the controller network. However, during a Wormhole attack, increment of the number of connected nodes in the network is not affected by AEED value. Therefore, the value is quiet similar in each scenario.

C. Throughput

$$Throughput = \frac{Total\ number\ of\ packets\ received\ by\ destination\ node}{Total\ time\ taken\ to\ receive\ packets} \quad (3)$$

This is an important measure to check the performance of the network. This value is a ratio which is calculated as Eq 3 by total number of packets received by destination node over total time taken to receive all packets. Units are bytes per second (bps).

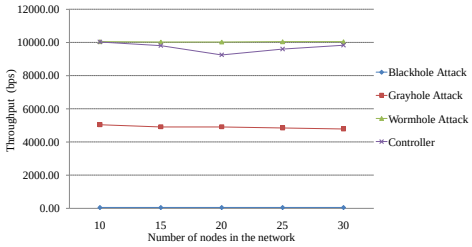


Fig. 3. Graph of Throughput vs. Number of Connected Nodes in the Network.

Fig. 3 shows a graph that is plotted between throughput vs. number of connected nodes in the network. The average throughput is 9706 bps. This is the 80.95 % of total data sent by source node. The network which is affected by Wormhole attack shows higher amount of throughput than controller network. The average throughput value is 10040 bps which is 83.92% of total data sent by source node. The lowest throughput value is recorded as 45 bps when a Blackhole attack is affected on a network. These amounts of routing data allow through Blackhole node. When Grayhole attack is affected on a network then the throughput amount will be lower than controller network and higher than Blackhole attack affected network. The average value is 4897 bps. The 41.03% of sent data from source node is received by destination node. Each network which is affected by any attack including controller network shows throughput decreasing while increasing the number of connected nodes in the network.

D. ADDR

$$ADDR = \frac{Total\ average\ data\ dropping\ rate\ of\ all\ nodes}{Simulation\ time} \quad (4)$$

Fig. 4 is a graph that shows average data dropping rate of a network. Graph plotted between ADDR vs. number of connected nodes in the network. Controller network shows 1091.34 bps as average ADDR. The ADDR value of the controller network increases by increasing the number of connected nodes in the network. The lowest average of ADDR value is recorded when Wormhole attack is affected on a network. It is 116.52 bps. This is 1/10 lower than controller networks' value. Furthermore, when number of connected nodes in the network increases, the ADDR decreases, because of the affect of a Wormhole attack. The higher average of

ADDR values is observed when a network is affected by a Blackhole attack. It is 8674.83 bps. This is 8 times higher than ADDN value of controller network. These affected networks do not show remarkable increase or decrease in ADDR during the change of number of connected nodes in the network. When a Grayhole attack is affected on a network, it shows 4350.73 bps. This is 4 times higher than ADDN value of controller network. Furthermore, when number of connected node increases the ADDN value also increases. The highest ADDR values are shown in the presence of Blackhole attack and the mean value of ADDR values is 0.106 bps. This means when a Blackhole attack is presented in a network, it drops data 106 times than a network without a Blackhole attack. The networks which are affected by a Grayhole attack show 0.008 bps of mean value for all ADDR value on the network.

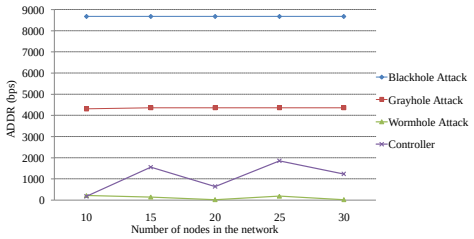


Fig. 4. Graph of ADDR vs. Number of Connected Nodes in the Network.

E. SPTIN

$$SPTIN = \frac{Total\ processing\ times\ of\ connected\ nodes\ in\ the\ network}{No\ of\ connected\ nodes\ in\ the\ network} \quad (5)$$

Fig. 5 is a graph which is plotted between SPTIN and number of connected nodes in the network. According to the graph, the SPTIN value of the controller network is 0.0137 seconds. The network which is affected by Balckhole attack shows higher amount of SPTIN. The average SPTIN value is 0.0570 seconds. It is nearly 4 times higher than SPTIN value of the controller network. Lower average SPTIN value is observed at the network which is affected by Wormhole attack. It is 0.0006 seconds. This is 20 times lower than average SPTIN value of the controller network. Average SPTIN value is 0.0020 seconds when a network is affected by Grayhole attack. This value is 7 times lower than average SPTIN value of controller network. All networks show relatively similar SPTIN value at the lowest number of connected nodes in the network except the network which is affected by a Blackhole attack. Furthermore, the linear trendline for all SPTIN value of the controller networks, Blackhole and Grayhole attack affected networks show gradual increase of SPTIN value with increase of the number of connected nodes in the network. However, when Wormhole attack is affected on a network, the SPTIN values of the networks are relatively same while the number of connected nodes in the network is increasing.

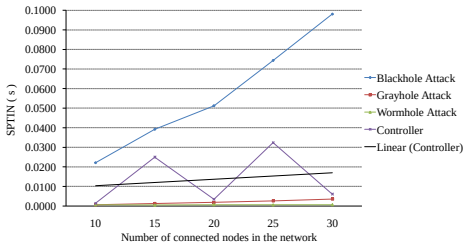


Fig. 5. Graph of SPTIN vs. Number of Connected Nodes in the Network.

VI. Summary of Analysis and Discussion

Table III shows summary of analyzed results. All values are converted into a percentage value with respect to the value of controller network. In Active attacks, Blackhole attacks causes severe damage to a network performances than Grayhole attack. Although performances of a network which is affected by a Grayhole attack varies in between the performances of the network which is affected by Blackhole attack and performances of the controller network. Furthermore, there are significant amount of enhancement of the network performances on the network by the affect of a Wormhole attack. Among the reactive routing protocols, AODV routing protocol performs well [5], [30]–[32]. Therefore, AODV routing protocol is used for network simulations in this research.

TABLE III. Summary of Analyzed Data

Network parameter	Controller (%)	Blackhole (%)	Grayhole (%)	Wormhole (%)
PDR	73.75	22.69	70.85	75.19
AEED	100.00	∞	69.51	12.48
Throughput	80.95	0.38	41.03	83.92
ADDR	100.00	10.68	398.66	794.88
SPTIN	100.00	416.75	14.74	4.68

VII. Conclusion and Future Works

According to the simulation results, we can conclude that active attacks are more destructive than passive attacks. In an Active attack, malicious node drops data packets. Though in a passive attack, malicious nodes provide better performances than as usual to become a part of the network. The reason for higher performance is Wormhole tunnel.

AODV routing protocol is more suitable for MANET. Pure AODV protocol is only considering about data communication but not data security. In AODV routing protocol, most of the malicious nodes become a part in the network at initial route discovery process. Therefore, node selection process for a network must be more qualitative and precise by concerning on data security. Furthermore, in MANET a routing protocol can be able to identify affects of Hole attacks therefore it is possible to apply suitable mechanisms to prevent these attacks through the routing protocol. Therefore, routing protocol should be equipped with an adaptive model which includes different suitable security mechanisms to prevent and handle the Hole attacks. In our future work, we intend to modify AODV protocol with an adaptive model to prevent these Hole attacks.

REFERENCES

[1] H. Xu, Y. Zhao, L. Zhang, J. Wang, "A Bio-Inspired Gateway Selection Scheme for Hybrid Mobile Ad Hoc Networks," *IEEE Access,* vol. 7, pp. 61997-62010, 2019, doi: 10.1109/ACCESS.2019.2916189

[2] L. Khoukhi, H. Badis, L. Boulahia, M. Esseghir, "Admission control in wireless ad hoc networks: a survey," *EURASIP Journal on Wireless Communications and Networking,* vol. 109(2013), pp. 1-13, 2013, doi: 10.1186/1687-1499-2013-109

[3] H. Deng, W. Li, D. Agarwal, "Routing security in wireless ad-hoc networks," *IEEE Communication Magazine,* vol. 40 (10), pp. 70-75, 2002, doi: 10.1109/MCOM.2002.1039859

[4] N. Brahmi, M. Boussedjra, J.Mouzna, "Routing in vehicular ad hoc networks: towards road-connectivity based routing," in *Mobile Ad-hoc Networks: Applications,* X. Wang, Ed. Janeza Trdine, Rijeka, Croatia: InTech, 2011. pp. 89-106.

[5] T. Nakashima, "Theory and applications of ad hoc networks," in *Mobile Ad-Hoc Networks: Protocol Design*, X. Wang, Ed. Janeza Trdine, Rijeka, Croatia: InTech, 2011. pp. 615-638.

[6] A. Ade, P. Tijare, "Performance comparison of AODV, DSDV, OLSR and DSR routing protocols in mobile ad hoc networks," *International Journal of Information Technology and Knowledge Management 2010,* vol. 2 (2), pp. 545-548, 2010, doi:10.1109/PDGC.2016.7913218

[7] M. Kang, D. Kum, J. Bae, Y. Cho, A. Le, "Mobility aware routing protocol for mobile ad hoc network" in *2012 International Conference on Information Networking,* Dresden, Germany: IEEE, 2012. pp. 410-414, doi: 10.1109/ICCW.2009.5208062

[8] C. Perking, E. Royer, "Ad-hoc on-demand distance vector routing," in *2nd IEEE Workshop on Mobile Computing Systems and Applications,* New Orleans, LA, USA: IEEE, 1999. pp. 90-100.

[9] L. He, J. Huang, F. Yang, "A noval hybrid wireless routing protocol for WMNs," *2010 International Conference on Electronics and Information Engineering,* Kyoto: IEEE, 2010. pp. 281-285, doi: 10.1109/ICEIE.2010.5559874

[10] R. Meddeb, B. Triki, F. Jemili, O. Korbaa, "A survey of attacks in mobile ad hoc networks," *2017 International Conference on Engineering & MIS (ICEMIS),* Monastir: IEEE, 2017, pp. 1-7, doi: 10.1109/ICEMIS.2017.8273007

[11] S. Banerjee, K. Majumder, "A Comparative Study on Wormhole Attack Prevention Schemes in Mobile Ad-Hoc Network," in *Recent Trends in Computer Networks and Distributed Systems Security. SNDS 2012. Communications in Computer and Information Science,* M. Thampi, Y. Zomaya, T, Strufe, A. Calero, T. Thomas, Eds. Berlin, Heidelberg: Springer, 2012, pp. 372–384, doi: 10.1007/978-3-642-34135-9_37

[12] R. Maulik, N. Chaki, "A comprehensive review on wormhole attacks in MANET," in *2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM),* Krackow: IEEE, 2010, pp. 233-238, doi: 10.1109/CISIM.2010.5643657

[13] F. Tseng, L. Chou, H. Chao, "A survey of black hole attacks in wireless mobile ad hoc networks," in *Human-centric Computing and Information Sciences 1,* Thaiwan: Springer, 2011, vol. 4 (2011), doi: 10.1186/2192-1962-1-4

[14] H. Jhaveri, J. Patel, C. Jinwala, "DoS Attacks in Mobile Ad Hoc Networks: A Survey," in *2012 Second International Conference on Advanced Computing & Communication Technologies,* Rohtak, Haryana: IEEE, 2012, pp. 535-541, doi: 10.1109/ACCT.2012.48

[15] A. Abdelshafy, B. King, "Analysis of security attacks on AODV routing," in *8th International Conference for Internet Technology and Secured Transactions (ICITST-2013),* London: IEEE, 2013, pp. 290-295, doi: 10.1109/ICITST.2013.6750209

[16] A. Saeed, A. Raza, H. Abbas, "A Survey on Network Layer Attacks and AODV Defense in Mobile Ad Hoc Networks," in *2014 IEEE Eighth International Conference on Software Security and Reliability-Companion,* San Francisco, CA: IEEE, 2014, pp. 185-191, doi: 10.1109/SERE-C.2014.37

[17] A. Nadeem, P. Howarth,"A Survey of MANET Intrusion Detection & Prevention Approaches for Network Layer Attacks," in *IEEE Communications Surveys & Tutorials,* Guildford, UK: IEEE, 2013, vol. 15, no. 4, pp. 2027-2045, doi: 10.1109/SURV.2013.030713.00201

[18] M. Karthigha, L. Latha, K. Sripriyan, "A Comprehensive Survey of Routing Attacks in Wireless Mobile Ad hoc Networks," in *2020 International Conference on Inventive Computation Technologies (ICICT),* Coimbatore, India: IEEE, 2020, pp. 396-402, doi: 10.1109/ICICT48043.2020.9112588

[19] A. Aggarwal, N. Chaubey, A. Jani, "A simulation study of malicious activities under various scenarios in Mobile Ad hoc Networks (MANETs)," in *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, Kottayam, India: IEEE, 2013, pp. 827-834, doi: 10.1109/iMac4s.2013.6526521

[20] B. Kannhavong, H. Nakayama, Y. Nemoto, N. Kaito, A. Jamalipour, "A survey of routing attacks in mobile ad hoc networks," in *IEEE Wireless Communications,*: IEEE, 2007, vol. 14 (5), pp. 85-91, doi: 10.1109/MWC.2007.4396947

[21] M. Rmayti, Y. Begriche, R. Khatoun, L. Khoukhi, D. Gaiti, "Flooding attacks detection in MANETs," in *2015 International Conference on Cyber Security of Smart cities (SSIC),* Shanghai, China: IEEE, 2015. pp. 1-6, doi: 10.1109/SSIC.2015.7245675

[22] S. Djahel, F. Abdesselam, A. Khokhar, "A cross layer framework to mitigate a joint MAC and routing attack in multihop wireless networks," in *2009 IEEE 34th Conference on Local Computer Networks,* Zurich, Switzerland: IEEE, 2009. pp. 730-737, doi: 10.1109/LCN.2009.5355066

[23] J, Karlsson, L. Dooley, G. Pulkkis, "Secure routing for MANET connected internet of things systems," in *IEEE International Conference on Future Internet of Things and Cloud,* Barcelona, Spain; IEEE, 2018. pp. 114-119.

[24] G. Li, Z. Yan, Y. Fu, "A study and simulation research of blackhole attack on mobile adhoc network" in *2018 IEEE Conference on Communications and Network Security (CNS),* Beijing, China; IEEE, 2018. pp. 1-6.

[25] R. Mehta, M. Parmar, "Trust based mechanism for Securing IoT Routing Protocol RPL against Wormhole &Grayhole Attacks," in *2018 3rd International Conference for Convergence in Technology (I2CT)* Pune, India: IEEE, 2018, pp. 1-6.

[26] P. Roshani, A. Patel, "Techniquesto mitigate grayhole attack in MANET: a survey," in *2017 International Conference on Innovations in information Embedded and Communication Systems (ICIIECS),* Coimbatore, India: IEEE, 2017, pp. 1-4.

[27] K. Ullah, P. Das, "Trust-based routing for mitigating grayhole attack in MANET," in *Proceedings of the International Conference on Computing and Communication Systems, Lecture Notes in Networks and Systems* J. Mandal, G. Saha, D. Kandar, A. Maji, Eds. Singapore: Springer Singapore, 2018. pp. 713-721, doi: 10.1007/978-981-10-6890-4_68

[28] M. Patel, A. Aggarwal, N. Chaubey, "Analysis of Wormhole Attacks in Wireless Sensor Networks," in *Recent Findings in Intelligent Computing Techniques. Advances in Intelligent Systems and Computing,* P. Sa, S. Bakshi, I, Hatzilygeroudis, M. Sahoo, Eds. Springer-Singapore: Springer, 2018, vol. 708, doi: 10.1007/978-981-10-8636-6_4

[29] M. Sohail, L.Wang, B.Yamin, "Trust mechanism based AODV routing protocol for forward node authentication in mobile ad hoc network" in *Mobile Ad-hoc and Sensor Networks,* L. Zhu, S. Zhong, Eds. Beijing, China: Springer Singapore, 2018. pp. 338–349.

[30] C. Lee, U. Lee, M. Gerla, "Survey of Routing Protocols in Vehicular Ad Hoc Networks," in *Advances in Vehicular Ad-Hoc Networks: Developments and Challenges,* M. Watfa Ed. Hershey PA 17033: IGI Global, 2015, vol. 9 (7), ch. 8, pp. 149-170.

[31] Y. Bai, Y. Mai, N. Wang, "Performance comparison and evaluation of the proactive and reactive routing protocols for MANETs," in *2017 Wireless Telecommunications Symposium (WTS),* Chicago, IL, USA: IEEE, 2017, pp. 1-5.

[32] A. Kanthe, D. Simunic, R. Prasad, "Comparison of AODV and DSR on-demand routing protocols in mobile ad hoc networks," in *2012 1st International Conference on Emerging Technology Trends in Electronics, Communication & Networking,* Gujarat, India: IEEE, 2012, pp. 1-5, doi: 10.1109/ET2ECN.2012.6470118

Hybrid Solution for Container Placement and Load Balancing based on ACO and Bin Packing

Oussama SMIMITE[1]
LabSIV, Department of Computer Science
Faculty of Science, Ibn Zohr University
BP 8106, 80000 Agadir, Morocco

Karim AFDEL[2]
LabSIV, Department of Computer Science
Faculty of Science, Ibn Zohr University
BP 8106, 80000 Agadir, Morocco

***Abstract*—Currently, data centers energy consumption in the cloud is attracting a lot of interest. One of the most approaches to optimize energy and cost in data centers is virtualization. Recently, a new type of container-based virtualization has appeared, containers are considered very light and modular virtual machines, they offer great flexibility and the possibility of migration from one environment to another, which allows optimizing applications for the cloud. Another approach to saving energy is to consolidate the workload, which is the amount of processing that the computer has to perform at any given time. In this article, we will study the container placement algorithm that takes into account the QoS requirements of different users in order to minimize energy consumption. Thus, we proposed a Hybrid approach for managing resources and workload based on ant colony optimization (ACO) and the first-fit decreasing (FFD) algorithm to avoid unnecessary power consumption. The results of the experiment indicate that using the first-fit decreasing algorithm (FFD) for container placement is better than ant colony optimization especially in a homogeneous systems. On the other hand the ant colony optimization shows very satisfying results in the case of workload management.**

***Keywords*—Cloud; virtualization; container; placement; Green IT; containerization**

I. Introduction

Recently, Cloud computing is considered as a new model that offered virtually immense resources. Customers can allocate resources as needed and pay as much as they have used. Thus, resources are managed by the cloud provider according to customer demand. In 2017, data centers in the United States consumed more than 90 billion kilowatt hours of electricity. Globally, data center power consumption was approximately 416 terawatts, or about 3% of all electricity produced on the planet.In a sense, the energy consumption of data centers worldwide was 40% more than all the energy consumed by the United Kingdom, an industrialized country with more than 65 million inhabitants. And this consumption will double every four years. [1, 2, 3, 4] From a business perspective, reducing energy consumption can lead to massive cost reductions. Moreover, in addition to the huge energy costs, heat dissipation inevitably increases with increasing energy consumption and doubles the probability of hardware failure [5, 6]. Therefore, reducing energy consumption not only saves a large amount of money and improves system reliability, but also helps protect our environment. According to [7], data centers emit CO2 like Argentina entirely, and their emissions are likely to exponentially increase in the coming years.

II. Background

There are different approaches to energy conservation. Besides the possibility of using more energy-efficient Hardware, reducing energy wasted due to the overuse of hardware is very important. The existing data center infrastructure is generally over-provisioned to maintain the availability of service during periods of high demand. However, the average use is low in datacenter due to tot he fort demand for resources in existing data centers.

Consequently, stopping or suspending unnecessary servers can impact a large part of the resources, which can influence performance constraints on clients.

Virtualization technologies such VMware, Xen, and Hyper V [8, 9, 10] are widely used in cloud datacenter due to their ease of use, flexibility of resource ,cost efficiency and the simplicity of enabling the high availability. More precisely, virtualization technologies offer the possibility of fine-tuning the resources allocation by associating processors, RAM, disk space and network bandwidth to a specific Virtual Machine [11, 12]. This approach has allowed the development of solutions such as Software as a Service (SaaS) and Platform as a Service (PaaS), on top of the usual Infrastructure as a Service (IaaS), where services providers can quickly make available virtual machines with the required resources to their customers almost in no time, and not burden them with the pain of infrastructure management.

Unlike traditional IT systems, Virtualization makes cloud computing more suitable for marketing, it provides promising approach to divide the resources of one or more physical servers into various parts and each part runs in an isolated environment [13]. To better manage resources, we can create isolated virtual machines (VMs) for each application, which allows us to parameterize the size of the resource such as the memory and the size of the processor according to the variable demand of the customers.

As virtualization allows us to create virtual instances, whether virtual machines (VM) or a container, the problem of virtual instances placement has become an important research subject in cloud computing. placement involves finding an optimal method for placing virtual instances on physical servers in order to efficiently use cloud resources [14, 15]. To maintain the servers in the data center, a lot of energy is consumed and the cost of cooling the installations is very high, which can translate to a very high cost [16]. Therefore, the goal of placement of virtual instances is to efficiently use physical

resources to host virtual resources, in order to reduce the number of running physical servers. In a cloud environment, good placement of virtual instances means placing the VMs in a way that the service level agreement (SLA) is guaranteed without signaling losses at the provider level. Within a data center, it is possible to place VMs with the main objective of reducing energy consumption and cost. Several studies have been developed in this context. We will present them in the following.

Besides the importance of the placement of virtual instances in a cloud environment, load management is a crucial issue to be resolved in order to maintain the system stability and improve the reliability of the cloud environment. [17, 18, 19] Load balancing ensures that all system instance do roughly the same amount of work at all times.

The organization of the paper is as follows. in Section III ,Containerization technologies, the Placement problem ,and load balancing problem are discussed . Section IV covers the Proposed System Architecture and the two Proposed types of algorithms: First Fit Decreasing (FFD)algorithm as a classic algorithm and Ant colony optimization (ACO) algorithm as a metaheuristic algorithm. Section V presents a analysis of experimental results and evaluation.In the last section, conclusions and future work are discussed.

III. Formal Problem Definition

A. Containerization

Containerization, or virtualization that uses containers, is a technology that virtualizes hardware resources in a container and ships applications and their dependencies across multiple operating systems. at the time of migration, the containers guarantee that their content is identical, and that it is secure, thanks to the isolation.

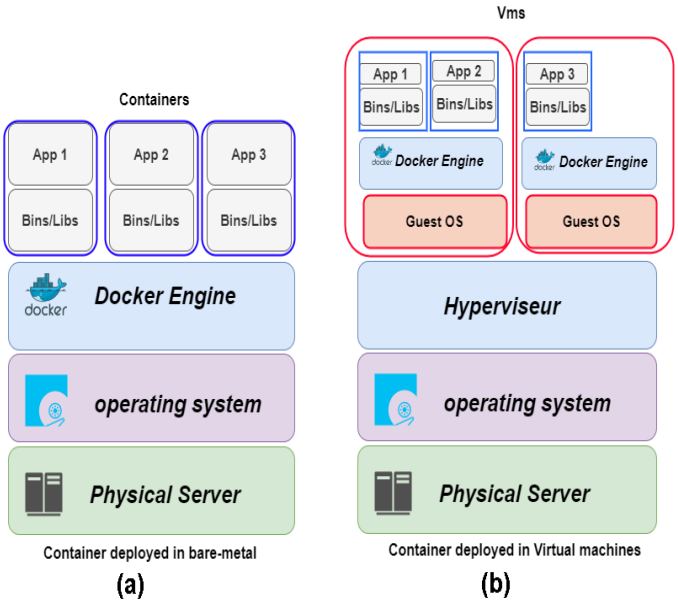


Fig. 1. Container Deployed in VMs VS Container Deployed in Bare-Metal.

1) Container deployed in Bare-Metal: Running containers on bare metal have many advantages, such as makes it possible to deploy applications with high performance in environments that can easily switch from the host server to another because there is no hardware emulation layer separating the containers from a host server [20, 21]. Furthermore, containerization allows for application isolation. Even if containers may not offer the same level of isolation as VMs, they set strict limits on the privileges and accessibility of resources associated with each container. But there are considerable difficulties that prevent the direct deployment of containers on the bare-metal host's server, such as the problem of updating physical servers. In fact, to replace a bare metal server, you must recreate the container environment from scratch while using virtualization makes it easy to migrate VMs to a new server. Another problem is that the containers depend on the type of the operating system, for example, Linux containers run in Linux hosts and Windows containers run on Windows hosts, also there are a few hosts that offer bare-metal solutions,most cloud platforms require VMs (see Fig. 1(a)).

2) Container deployed in virtual machines: deploying containers on Vms, offer advantages such as applications can be easily moved from one host to another by transferring images from one server to another. from a security point of view, applications that run in different VMs are isolated What makes management easier [22]. Also the possibility of grouping the same types of containers in a VM which allows creating a more coherent system. But virtual machines also have disadvantages such as under utilization of resources because of the pre-allocation of these resources even if not used. also, the VMs cannot directly access the physical hardware in order to unload it in the event of an overload (see Fig. 1(b)).

B. Container Placement Problem

Virtualization is a technique used to take better advantage of hardware resources, so it is useful for deploying more test environments thanks to the use of virtual machines. However, the appearance of containers has been evaluated as an improvement of virtualization. For this, we can consider the container placement as an improved version of virtual machine placement, to better managing resources in a cloud environment.The placement of containers is an important operation that has a direct effect on resource utilization, energy consumption, and Resource utilization cost. an efficient placement optimizes the use of material resources by minimizing the number of physical machines active in a data center, which allows both to minimize the cost of resources utilization and reduces energy consumption by stopping the inactive physical machine.

C. Load Balancing Problem

load balancing is an approach of distributing workloads across various computing resources to improve response time and resource utilization. load balancing is used to balance the load between the different resources of the system to avoid having idle resources and overused resources. [23] in a homogeneous environment where all resources are identical, the load must be distributed equally.but in heterogeneous environments, resources that have more capacity should be used more than other resources.

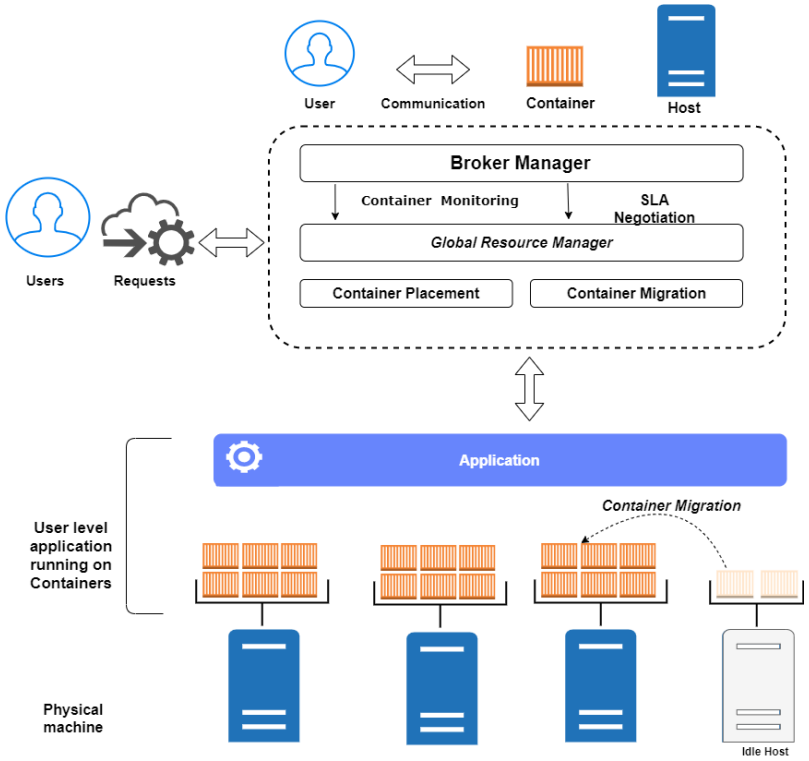


Fig. 2. Proposed System Architecture.

IV. Proposed System and Algorithms

A. Proposed System Architecture

Fig. 2 shows the proposed architecture of the system, which consists of:

1) The Brokers: The Broker is the main component of the System, it serves as an agent between service providers and customers. he may also sign SLA terms with cloud providers in place of the client. When the customer request has been sent, the broker initiates the process and submits those requests to the other modules. after that, based on information provided by the monitor module he makes the decision whether to approve or reject the request. It provides an interface for multiple clouds and sharing resources.

2) The Monitor of containers: This module's key role is to control resource state, RAM use, processor utilization, SLA violation, and power consumption. Once an exception in resource usage is triggered, The controller must send a warning to The Global Resources Manager to take necessary action.

3) The Global Resource Manager: The main function of this module is to analyze user requests and check the QoS requirements before choosing to accept or refuse the request. To guarantee that no SLA violation persists the module requests updated information from the Containers Monitor to reallocate resources efficiently.

4) The Physical Machines (PM): Also known as the "bare-metal server" is the physical machine, which is the support of hardware to create and host Container. The PM can host several Containers depending on their capacity.

5) The Containers (CNT): Containers are the layer of virtualization that runs within the operating system. Therefore containers are relatively light and take only seconds to get started, unlike VM. The speed, flexibility, and portability of containers permit them to Help optimizing software development. Process of transferring Containers from one PM to another called Container migration.

B. Proposed Algorithms

C. First Fit Decreasing algorithm (FFD)

The bin packing In operational research is an algorithmic problem that involves storing objects with a minimum number of boxes. It can be applied in IT such as the storage of files on IT support.[24] To solve the bin packing problem, we often use simple algorithms like first-fit decreasing (FFD) which works as follows: we sort the list of articles in decreasing order of size, then we put each article in order. In first-fit, we put the current article in the first box that can contain it (see Fig. 3) . This algorithm allow to obtain very good results in practice. In our case, we used the FFD algorithm to allocate the containers in the hosts as the Container placement is considered a Bin Packing problem [25, 26, 27]. The bin represents the physical machine and the items are the containers to be assigned to the Bin. The containers are sorted first in descending order of their Ram memory capacity. The pseudo-code of containers placement is presented in algorithm 1.

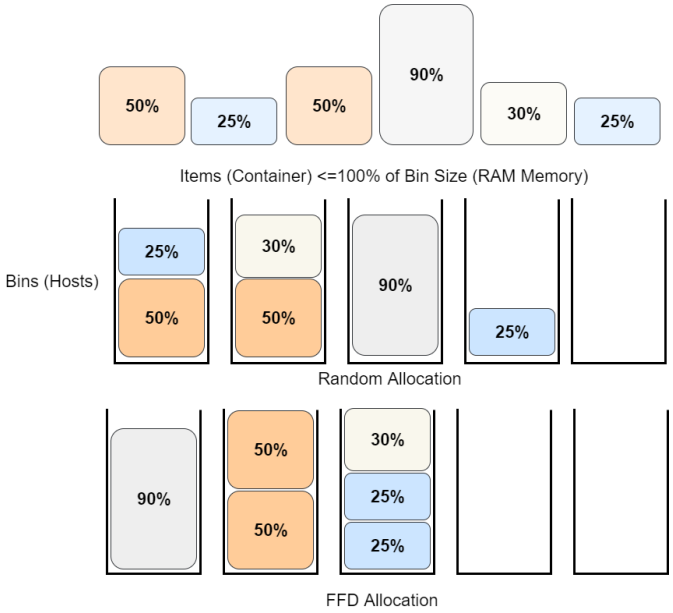


Fig. 3. FFD illustration example.

Algorithm 1 Container placement (FFD).

Require: ContainerList ,HostListOutput /: ContainerPlacement
Sort the containers in descending order according to the RAM memory
while container in containerList **do**
if $container.Ram <$ container.Ram.next **then**
Exchange (container, container.next)
end if
end while
max=VmRAM *threshold
while Host in HostList **do**
while container in containerList **do**
if $VmRamEstimate(Vm, container) <$ max **then**
Allocate (Host ,Container)
end if
end while
end while

D. Ant Colony Optimization algorithm (ACO)

Ant Colony Optimization (ACO) is a meta-heuristic inspired from the natural food-discovery behavior of real ants. due to the limited memory of the ants they have developed a system of communication based on chemical substance called pheromone, this last is used by each ant to mark their tracks .Other ants can smell the concentration of this substance and chose the paths probabilistically according to the quantity of the pheromone. after a while, the entire ant colony move towards the shortest path to the food source. at first, the algorithm was developed to solve the Traveling Salesman Problem (TSP). after that, it has been successfully adapted to solve many other complex combinatorial optimization problems.in our case, we try to use this algorithm to solve the placement problem, and also the load balancing issues [28]. The pseudo-codes of containers placement and the load balancing are presented consecutively in algorithm 2 and algorithm 3. The containers placement is represented as a graph g= (N, E) where N is the Set of Containers and the physical machines, and E represent the connections between Containers and physical machines as mentioned in Fig. 4(A) we can also represent the load balancing issue as graph G1= (N1, E1) where N1 is the set of Tasks and containers and E1 the connections between the task and containers as shown in 4(B)

1) ACO Container Placement: In virtualization, once the virtual machine starts, the RAM memory (R) allocated by the VM becomes unavailable for the physical machine.for that in our approach, we focus on Ram memory as being an important parameter for placing containers in the host.

In the proposed algorithm for containers placement, each ant receives all the containers and try to assign them to the host using the probabilistic decision (Prob) rule mentioned equation 1

$$Prob_{Host} = PH(Host)^{alpha} * H(Host)^{beta} \quad (1)$$

the probabilistic calculation (Prob) is based on the present concentration of pheromone (PH), and a heuristic (H) which help ants to chose the most optimizing hosts.Besides, two parameters $alpha, beta \geq 0$ are used to point out more the heuristic information or the pheromone.

for each container(CNT) we calculated the possible ram allocation (RA) for every host(PM) using the equation 2 we can represented the results as a graph $G1 = (C, ((H1, A1), (H2, A1), .., (Hn, An)))$ whereby, C is the container,H is the host and A is the possible allocation for every host.

$$RA^{PM}_{CNT} = CNT_R / PM_R \quad (2)$$

the heuristic (H) is calculated based on 2 as mentioned in equation 3 we can also represented that as a graph $G2 = ((H1, \sum A/A1), (H2, \sum A/A2), .., (Hn, \sum A/An))$ whereby, H is the host and $\sum$ A is a is the sum of the possible allocation of each container .

$$H^{PM}_{CNT} += \sum RA(CNT)/RA(CNT)_{PM} \quad (3)$$

the pheromone (PH) concentration of the host is initialized by the RAM memory capacity of each host .after a host has been chosen by an ant the pheromone concentration is updated according to the equation 4

$$PH^{BestPM} = PH(BestPM) * (1 - rho) + Q/RA(BestPM) \quad (4)$$

as, the constant rho,$0 \leqslant rho \leqslant 1$ is used to simulate pheromone evaporation.and Q is an adaptive parameter. in the end, we compare all ants allocation proposition for each container and choose the best solution.

2) ACO load balancing: As mentioned before, the objective of this approach is to distribute the workloads on the cloud resources in a balanced way.to do this we consider the workload as a list of cloudlets that must be run on a set of containers. In the proposed algorithm for load balancing, each ant receives all the cloudlets (Cl) the and try to execute them in the appropriate container(CNT) using the probabilistic (Prob) decision rule mentioned equation 5

$$Prob_{CNT} = PH(CNT)^{alpha} * ET(CNT)^{beta} \quad (5)$$

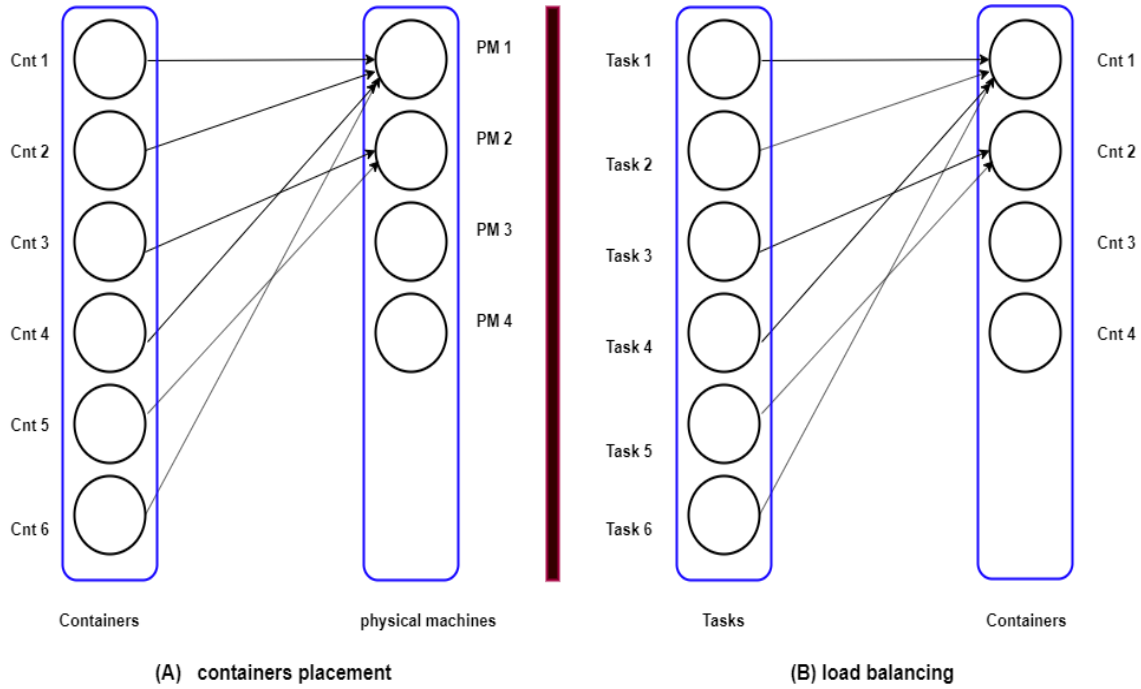


Fig. 4. Problem representation Based on ACO.

Algorithm 2 ACO-based Container placement

Require: ContainerList ,HostList Output /: ContainerPlacement
Initialize parameters, Set pheromone value
for all ContainerList **do**
  for all HostList **do**
    T= container.getRam()/Host.getRam()
  end for
  initializeRamAllocation(T);
end for
for all HostList **do**
  C = host.getRam();
  initializePheromone(C);
end for
for t=1 **to** t=tmax **do**
  for all ContainerList **do**
    calculate allocation percentage of each container according to Eq 3
    for all HostList **do**
      Calculate probability according to Eq 1
      initializeProbability(P);
    end for
    Compare ants solutions and vote for best solution
    for i=1 **to** k **do**
      vote(Hosts,probab)
    end for
    for all HostList **do**
      if $MaxVote$ **then**
        Allocate (BestHost ,Container)
        UpdatePheromones according to Eq 4
      end if
    end for
  end for
end for

the probabilistic calculation (Prob) is based on the present concentration of pheromone (PH), and a heuristic (ET) which help ants to chose the most optimizing container. Besides, two parameters $alpha, beta \geq 0$ are used to point out more the heuristic information or the pheromone.

for each cloudlet, we calculate the estimate Execution time for every container using the equation 6 we can represented the results as a graph $G3 = (Cl, ((C1, T1), (C2, T1), .., (Cn, Tn)))$ whereby, Cl is the cloudlet ,C is the container and T is the estimate Execution time for every container.

$$ET_{Cl}^{CNT} = Cl_{Length}/(CNT_{NbrPes} * CNT_{Mips} \quad (6)$$

the heuristic (H) is calculated based on 6 as mentioned in equation 7 we can also represented that as a graph $G4 = ((C1, \sum T/T1), (C2, \sum T/T2), .., (Cn, \sum T/Tn))$ C is the container and $\sum$ T is a is the sum of the estimate Execution time of each cloudlet.

$$H_{Cl}^{CNT} += \sum ET(Cl)/ET(Cl)_{CNT} \quad (7)$$

the pheromone concentration of the host is initialized by the Number of MIPS of each container. After a container has been chosen by an ant the pheromone concentration is updated according to the equation 8

$$PH^{BestCNT} = PH * (1 - rho) + Q/ET(BestCNT); \quad (8)$$

as, the constant rho,$0 \leqslant rho \leqslant 1$ is used to simulate pheromone evaporation, and Q is an adaptive parameter. in the end, we compare all ants proposition for each Cloudlet and choose the best solution.

Algorithm 3 ACO-based Load Balancing

```
Require: CloudletList ,HostList Output /: Load Balancing
  Initialize parameters, Set pheromone value
  for all CloudletList do
    for all ContainerList do
      T=Cloudlet.Length()/(Container.NbrPes()*Container.Mips())
    end for
    initializeExecTimes(T);
  end for
  for all ContainerList do
    C = Container.NbrPes()*Container.Mips()
    initializePheromone(C);
  end for
  for t=1 to t=tmax  do
    for all CloudletList do
      calculate estimate Execution time of each Cloudlet
      according to Eq 6
      for all ContainerList do
        Calculate probability according to Eq 5
        initializeProbability(P);
      end forCompare ants solutions and vote for best solu-
      tion
      for i=1 to k do
        vote(Container,probab)
      end for
      for all ContainerList do
        if MaxVote then
          Allocate (BestContainer,Cloudlet)
          UpdatePheromones according to Eq 8
        end if
      end for
    end for
  end for
```

V. Experimental Setup and Results

To evaluate our proposed method, simulation experiments are implemented on CloudSim[29, 30] to study the effects.

We tried to apply the two proposed algorithms (ACO and FFD) on the Container placement and load balancing.

For the Container placement we adapt two approach , in The first we use a homogeneous system, where hosts has the same characteristic (CPU, RAM, Bandwidth), in the other approach we use hosts with different characteristics for the load balancing, in the heterogeneous system, we use a different types of containers, and in the other scenario we use many identical containers .

A. Containers Placement

1) Homogeneous system: In this experiment, our configuration consists in using 3 identical hosts, 90 containers as shown in the Table I.

a) Scenario 1: In this Scenario,we apply the FFD algorithm to assign all container in the hosts based on RAM memory. Container placement can be considered a Bin Packing problem,The bin represents the Host and the items being the containers to be assigned to the Bin. The containers are sorted first in descending order of their Ram memory capacity Before applying the FFD algorithm to the allocation problem, we should first make sure that the total memory capacity of the containers does not exceed the host's available memory capacity. in order to avoid a probable performance degradation because of the overloading of ram memory, we define a maximum threshold, 80% of host RAM memory, which can reduce the risk of Service Level Agreement (SLA) violation.

TABLE I. Characteristics of hosts and containers in experiment 1.

	Number	MIPS	RAM	BW
Host -type 1-	3	37274/8	32768	1000000
CONTAINER -type 1-	21	2358	128	2500
CONTAINER -type 2-	23	4658	256	2500
CONTAINER -type 3-	23	9320	512	2500
CONTAINER -type 4-	23	18636	1024	2500

b) Scenario 2: In this Scenario, we repeat the previous scenario but we sorted first all hosts in descending order of their Ram memory capacity before applying the FFD algorithm.

c) Scenario 3: In this Scenario,we apply the ACO algorithm and inspected their efficiency by experimentation. The parameters (alpha, Beta,rho, tmax, m the number of ants and Q) considered here are those that affect directly or indirectly the computation of the algorithm Table II. showing the selected ACO parameter. in order to avoid a probable performance degradation because of the overloading of ram memory, we define a maximum threshold, 80% of host RAM memory, which can reduce the risk of SLA violation.

TABLE II. Selected Parameters of ACO

Parameter	alpha	beta	rho	Q	m	Tmax
Value	1	2	0.7	100	10	100

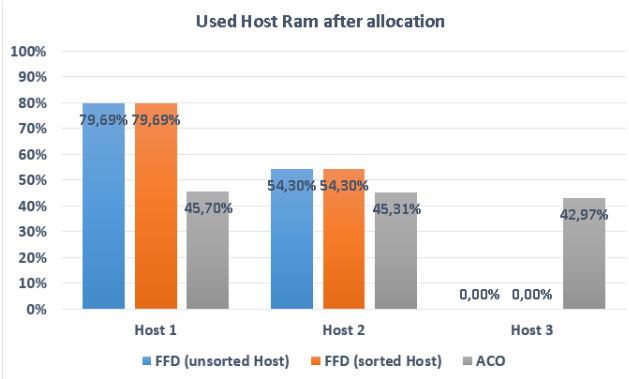


Fig. 5. Used Host Ram after Allocation (Homogeneous System).

As it is presented in Fig. 5 using the FFD algorithm allows us to use the totality of RAM memory of host 1 and reduce the total number of Hosts used. on the other hand, when using ACO algorithm the three hosts are used and the Ram memory used of each host does not exceed 46

Regarding the distribution of RAM memory of the containers on the hosts, we notice that in the case of using the FFD

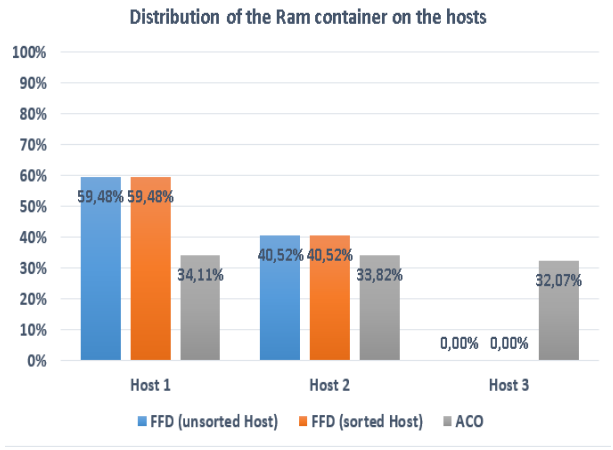


Fig. 6. Distribution of the Ram Container on the Hosts (Homogeneous System).

algorithm the first host has favored compared to the others hosts. on the other hand, when we use the ACO algorithm the distribution is done in a balanced way. (see Fig. 6)

2) Heterogeneous system: In this experiment, our configuration consists in using three different hosts, 80 containers as shown in Table III.

TABLE III. CHARACTERISTICS OF HOSTS AND CONTAINERS IN EXPERIMENT 2 .

	Number	MIPS	RAM	BW
Host -type 1-	1	37274/8	32768	1000000
Host -type 2-	1	37274/4	16384	1000000
Host -type 3-	1	37274/2	8162	1000000
CONTAINER -type 1-	20	2358	128	2500
CONTAINER -type 2-	20	4658	256	2500
CONTAINER -type 3-	20	9320	512	2500
CONTAINER -type 4-	20	18636	1024	2500

In this approach, we repeat the three preceding scenarios (FFD sorted ,FFD unsorted and ACO) and compare the results.

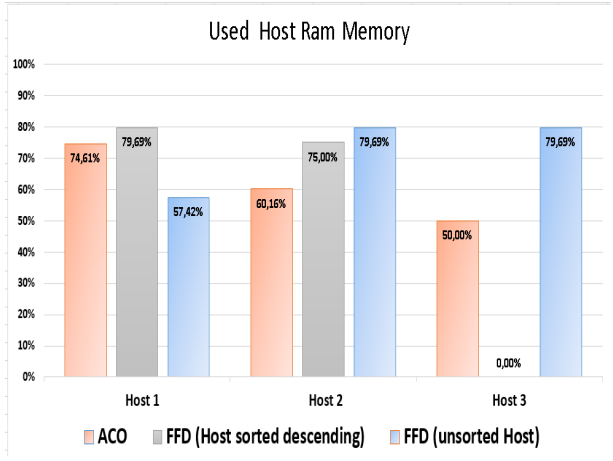


Fig. 7. Used Host Ram after Allocation (Heterogeneous System).

As it is shown in Fig. 7 when we use the FFD algorithm in heterogeneous unsorted system all hosts are used and the Ram memory used of Host 1 does not exceed 58%. but sorting host before applying the FFD algorithm allows us to use the totality of RAM memory of host 1 and reduce the total number of Hosts used. on the other hand, when using ACO algorithm the three hosts are used and the Ram memory used of each host depends on the characteristics of each host.

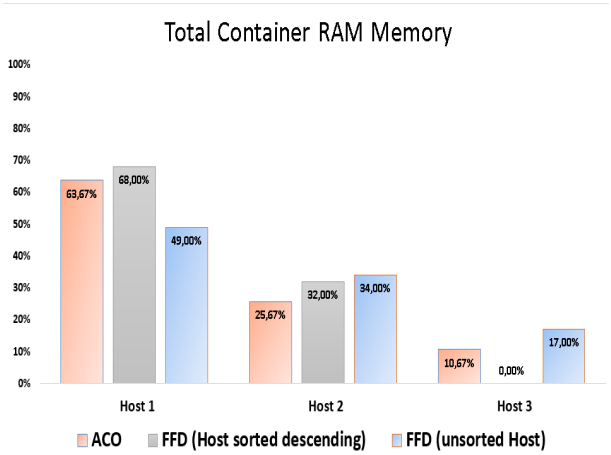


Fig. 8. Distribution of the Ram Container on the Hosts (Heterogeneous System).

Regarding the distribution of RAM memory of the containers on the hosts, we notice that in the case of using the FFD algorithm in an unsorted system the load which can handle by two hosts is distributed into three hosts, what is shown in the sorted system. on the other hand, when we use the ACO algorithm the distribution is done in a balanced way based on the percentage of each host in Ram's total memory (see Fig. 8).

To better evaluate the algorithms proposed or to propose a scenario where we combined between the two preceding systems ,we use several group of hosts as shown in Table IV.

TABLE IV. CHARACTERISTICS OF HOSTS AND CONTAINERS IN EXPERIMENT 3 .

	Number	MIPS	RAM	BW
Host -type 1-	10	37274/8	32768	1000000
Host -type 2-	10	37274/4	16384	1000000
Host -type 3-	10	37274/2	8162	1000000
CONTAINER -type 1-	75	2358	128	2500
CONTAINER -type 2-	75	4658	256	2500
CONTAINER -type 3-	75	9320	512	2500
CONTAINER -type 4-	75	18636	1024	2500

The third experiment gives us a general idea of the usefulness of each algorithm. for container placement, the most important thing is to allocate all containers using the minimum host taking under consideration the maximum threshold for RAM usage Fig. 9 shows that the use of the FFD algorithm reduced the total number of hosts used from 30 hosts in the case of ACO to 17 hosts in an unsorted system and to 6 in a sorted system. which allowed us to optimize nearly 40% to 80% of the host used.

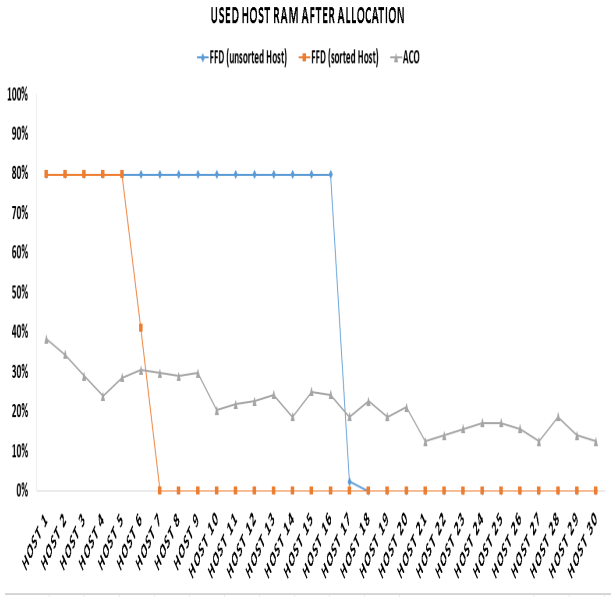


Fig. 9. Used Host Ram after Allocation.

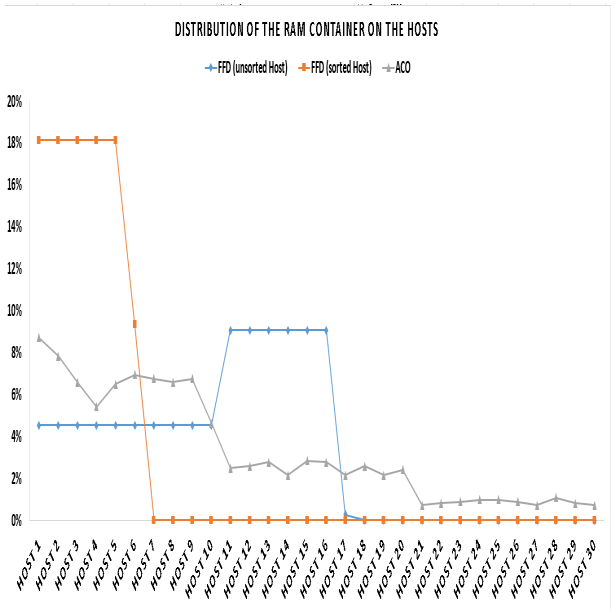


Fig. 10. Distribution of the Ram Container on the Hosts.

As shown in Fig. 10 the ACO curve can be divided into three parts according to the characteristics of the host. each part can be considered as a homogeneous system where the hosts are identical, this is why in each party the load is distributed in a balanced way.

To take advantage of the two algorithms, we propose a hybrid solution (ACO-FFD) where we apply first the FFD algorithm to optimize the number of hosts after that apply the ACO algorithm on the chosen hosts to balance the load between them. Fig. 11 and 12 show the results.

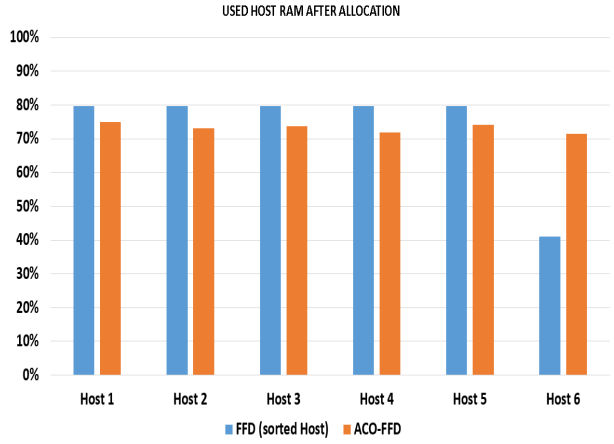


Fig. 11. Used Host Ram after Allocation (Hybrid Approach).

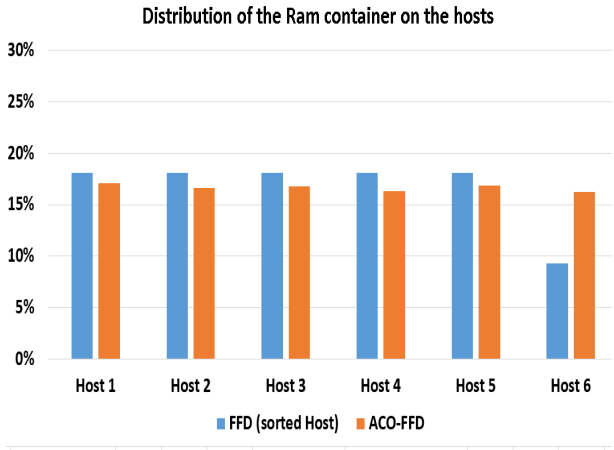


Fig. 12. Distribution of the Ram Container on the Hosts (Hybrid Approach).

B. Task Scheduling and Load Balancing

To assess the impact of our algorithm proposed on the task Scheduling and Load Balancing, we consider each task as cloudlet identified by its length must be performed in containers according to the MIPS number of each container.

1) Heterogeneous system: In this experiment, our configuration consists in using four different containers and 100 Cloudlet where their size varies between 100 and 400 mips as shown in Table V.

TABLE V. CHARACTERISTICS OF CONTAINERS AND CLOUDLETS IN EXPERIMENT 4.

	Number	MIPS	RAM	Lenght
CONTAINER -type 1-	1	2358	128	
CONTAINER -type 2-	1	4658	256	
CONTAINER -type 3-	1	9320	512	
CONTAINER -type 4-	1	18636	1024	
Cloudlet	100			100 /400

a) Scenario 1: In this Scenario, we apply the FFD algorithm to allocate all Cloudlet in Containers based on the

length of each cloudlet and the Mips of each container. in order to avoid a probable performance degradation because of the overloading of container, we define a maximum threshold, 80% of container Mips, which can reduce the risk of degradation of the quality of service.

b) Scenario 2: In this Scenario, we apply the ACO algorithm and inspected their efficiency by experimentation. The parameters (alpha, Beta, rho, tmax, m the number of ants and Q) considered here are those that affect directly or indirectly the computation of the algorithm. Table VI showing the selected ACO parameter.

TABLE VI. SELECTED PARAMETERS OF ACO

Parameter	alpha	beta	rho	Q	m	Tmax
Value	1	2	0.7	100	10	100

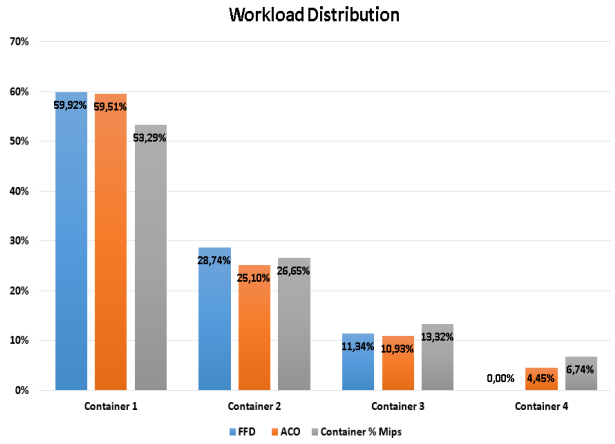


Fig. 13. Workload Distribution on Containers (Heterogeneous System).

Regarding the Workload Distribution on containers, we notice that in the case of using the FFD algorithm until we reach the maximum threshold, the first containers are favored compared to the others hosts and container 4 is not used. On the other hand, when we use the ACO algorithm the distribution is done in a balanced way based on the percentage of each host in MIPS's total (see Fig. 13).

2) Homogeneous system: In this experiment, our configuration consists using 100 Cloudlet where their size varies between 100 and 400 mips and four identical containers as shown in Table VII.

TABLE VII. CHARACTERISTICS OF HOSTS AND CONTAINERS IN EXPERIMENT 2 .

	Number	MIPS	RAM	Lenght
CONTAINER -type 3-	4	9320	512	
Cloudlet	100			100/400

In this approach, we repeat the two preceding scenarios (FFD and ACO) and compare the results.

As it is shown in Fig. 14 using the FFD algorithm we notice that almost 90% of the workload is managed by the first three containers and that just 11% is managed by container 4.

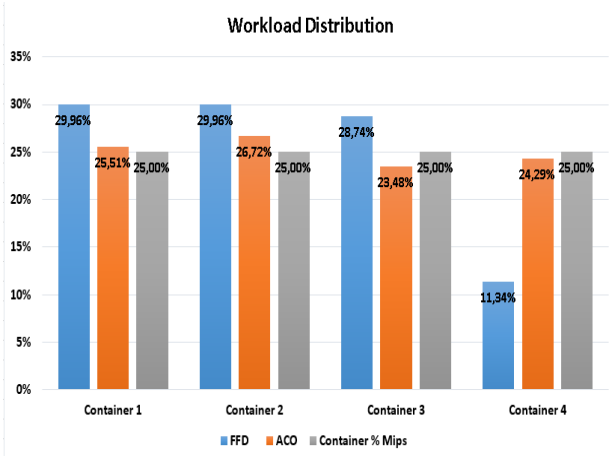


Fig. 14. Workload Distribution on Containers (Homogeneous System).

on the other hand, when we use the ACO algorithm the distribution is done in a balanced way. almost 25 % of workload for each container.

VI. CONCLUSION AND FUTURE WORK

In this work, we try to present Containerization technologies, the Placement problem, load balancing problem and there impact in a cloud environment. In addition, we provide a container allocation approach based on the ACO and FFD algorithm, taking into account QoS requirements and service level agreement. The use of the FFD algorithm has allowed us to better manage the placement of containers using a minimum number of hosts, which reduces power consumption. The use of ACO shows very acceptable results for a balanced workload management. At the end, a hybrid approach was proposed between the two methods in order to benefit from the advantages of each of these algorithms. As a perspective of our research work, we plan to track under-utilized Hosts by proposing solutions based on the Metaheuristic algorithm to optimize our architecture based on containerization.

REFERENCES

[1] A. Shehabi, S. Smith, D. Sartor, R. Brown, M. Herrlin, J. Koomey, E. Masanet, N. Horner, I. Azevedo, and W. Lintner, "United states data center energy usage report," Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), Tech. Rep., 2016.

[2] X. Zhang, T. Lindberg, N. Xiong, V. Vyatkin, and A. Mousavi, "Cooling energy consumption investigation of data center it room with vertical placed server," *Energy procedia*, vol. 105, pp. 2047–2052, 2017.

[3] W. Jiye, Z. Biyu, Z. Fa, S. Xiang, Z. Nan, and L. Zhiyong, "Data center energy consumption models and energy efficient algorithms," *Journal of Computer Research and Development*, vol. 56, no. 8, p. 1587, 2019.

[4] S. Pang, K. Xu, S. Wang, M. Wang, and S. Wang, "Energy-saving virtual machine placement method for user experience in cloud environment," *Mathematical Problems in Engineering*, vol. 2020, 2020.

[5] M. I. Green, "Cloud computing and its contribution to climate change," *Greenpeace International*, vol. 83, 2010.

[6] S. K. Mishra, B. Sahoo, and P. P. Parida, "Load balancing in cloud computing: a big picture," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 2, pp. 149–158, 2020.

[7] J. M. Kaplan, W. Forrest, and N. Kindler, "Revolutionizing data center energy efficiency," Technical report, McKinsey & Company, Tech. Rep., 2008.

[8] P. K. Das, "Comparative study on xen, kvm, vsphere, and hyper-v," in *Emerging Research Surrounding Power Consumption and Performance Issues in Utility Computing*. IGI Global, 2016, pp. 233–261.

[9] F. Rodríguez-Haro, F. Freitag, L. Navarro, E. Hernánchez-sánchez, N. Farías-Mendoza, J. A. Guerrero-Ibáñez, and A. González-Potes, "A summary of virtualization techniques," *Procedia Technology*, vol. 3, pp. 267–272, 2012.

[10] S. S. Kolahi, V. S. Hora, A. P. Singh, S. Bhatti, and S. R. Yeeda, "Performance comparison of cloud computing/iot virtualization software, hyper-v vs vsphere," in *2020 Advances in Science and Engineering Technology International Conferences (ASET)*. IEEE, 2020, pp. 1–6.

[11] S. J. Vaughan-Nichols, "New approach to virtualization is a lightweight," *Computer*, vol. 39, no. 11, pp. 12–14, 2006.

[12] R. Bachu, "A framework to migrate and replicate vmware virtual machines to amazon elastic compute cloud: Performance comparison between on premise and the migrated virtual machine," 2015.

[13] A. Abohamama and E. Hamouda, "A hybrid energy–aware virtual machine placement algorithm for cloud environments," *Expert Systems with Applications*, vol. 150, p. 113306, 2020.

[14] Y. Gao, H. Guan, Z. Qi, Y. Hou, and L. Liu, "A multi-objective ant colony system algorithm for virtual machine placement in cloud computing," *Journal of Computer and System Sciences*, vol. 79, no. 8, pp. 1230–1242, 2013.

[15] C. Vijaya and P. Srinivasan, "A hybrid technique for server consolidation in cloud computing environment," *Cybernetics and Information Technologies*, vol. 20, no. 1, pp. 36–52, 2020.

[16] M. Uddin, Y. Darabidarabkhani, A. Shah, and J. Memon, "Evaluating power efficient algorithms for efficiency and carbon emissions in cloud data centers: A review," *Renewable and Sustainable Energy Reviews*, vol. 51, pp. 1553–1563, 2015.

[17] A. Hota, S. Mohapatra, and S. Mohanty, "Survey of different load balancing approach-based algorithms in cloud computing: a comprehensive review," in *Computational Intelligence in Data Mining*. Springer, 2019, pp. 99–110.

[18] V. Priya, C. S. Kumar, and R. Kannan, "Resource scheduling algorithm with load balancing for cloud service provisioning," *Applied Soft Computing*, vol. 76, pp. 416–424, 2019.

[19] P. Kumar and R. Kumar, "Issues and challenges of load balancing techniques in cloud computing: a survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–35, 2019.

[20] I. Odun-Ayo, V. Geteloma, I. Eweoya, and R. Ahuja, "Virtualization, containerization, composition, and orchestration of cloud computing services," in *International Conference on Computational Science and Its Applications*. Springer, 2019, pp. 403–417.

[21] A. Abuabdo and Z. A. Al-Sharif, "Virtualization vs. containerization: Towards a multithreaded performance evaluation approach," in *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*. IEEE, 2019, pp. 1–6.

[22] S. C. Mondesire, A. Angelopoulou, S. Sirigampola, and B. Goldiez, "Combining virtualization and containerization to support interactive games and simulations on the cloud," *Simulation Modelling Practice and Theory*, vol. 93, pp. 233–244, 2019.

[23] R. Mishra and A. Jaiswal, "Ant colony optimization: A solution of load balancing in cloud," *International Journal of Web & Semantic Technology*, vol. 3, no. 2, p. 33, 2012.

[24] L. Tadic, P. Afric, L. Sikic, A. S. Kurdija, V. Klemo, G. Delac, and M. Silic, "Analysis and comparison of exact and approximate bin packing algorithms," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2019, pp. 919–924.

[25] X. Tang, Y. Li, R. Ren, and W. Cai, "On first fit bin packing for online cloud server allocation," in *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2016, pp. 323–332.

[26] H. I. Christensen, A. Khan, S. Pokutta, and P. Tetali, "Approximation and online algorithms for multidimensional bin packing: A survey," *Computer Science Review*, vol. 24, pp. 63–79, 2017.

[27] G. Dósa and L. Epstein, "The tight asymptotic approximation ratio of first fit for bin packing with cardinality constraints," *Journal of Computer and System Sciences*, vol. 96, pp. 33–49, 2018.

[28] E. Feller, L. Rilling, and C. Morin, "Energy-aware ant colony based workload placement in clouds," in *2011 IEEE/ACM 12th International Conference on Grid Computing*. IEEE, 2011, pp. 26–33.

[29] M. C. Silva Filho, R. L. Oliveira, C. C. Monteiro, P. R. Inácio, and M. M. Freire, "Cloudsim plus: a cloud computing simulation framework pursuing software engineering principles for improved modularity, extensibility and correctness," in *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*. IEEE, 2017, pp. 400–406.

[30] D. A. A. G. Singh, R. Priyadharshini, and E. J. Leavline, "Analysis of cloud environment using cloudsim," in *Artificial Intelligence and Evolutionary Computations in Engineering Systems*. Springer, 2018, pp. 325–333.

Speaker-Independent Speech Recognition using Visual Features

Pooventhiran G.[1], Sandeep A.[2], Manthiravalli K.[3], Harish D.[4], Karthika Renuka D.[5]
Undergraduate, Dept. of Information Technology
PSG College of Technology
Coimbatore, Tamil Nadu, India - 641004[1,2,3,4]
Associate Professor, Dept. of Information Technology
PSG College of Technology
Coimbatore, Tamil Nadu, India - 641004[5]

***Abstract*—Visual Speech Recognition aims at transcribing lip movements into readable text. There have been many strides in automatic speech recognition systems that can recognize words with audio and visual speech features, even under noisy conditions. This paper focuses only on the visual features, while a robust system uses visual features to support acoustic features. We propose the concatenation of visemes (lip movements) for text classification rather than a classic individual viseme mapping. The result shows that this approach achieves a significant improvement over the state-of-the-art models. The system has two modules; the first one extracts lip features from the input video, while the next is a neural network system trained to process the viseme sequence and classify it as text.**

Keywords*—Visual speech recognition; audio speech recognition; visemes; lip reading system; Convolutional Neural Network (CNN)

I. Introduction

Visual Speech Recognition (VSR) is the process of extracting textual or speech data from facial features through image processing techniques. It plays a vital role in human-computer interaction; mostly in noisy environments, it complements Automatic Speech Recognition systems to improve performance [1][2]. Like speech recognition systems, lip reading (LR) systems also face problems due to variances in skin tone, speaking speed, pronunciation, and facial features. A standalone lip reading system may not be very efficient. Several factors, such as skin tone, accents, duration of utterances, limit this efficiency. The LR systems can be synchronized with an Audio Speech Recognition system to improve the confidence of classification by using both model's advantages [3]. Many systems limit the datasets to contain only a few words and phrases rather than all possible sentences to simplify this problem. Speech recognition systems are of two types: Speaker-dependent and Speaker-independent systems. Speaker-dependent systems train on data from a single speaker and are suitable for speech and speaker verification applications [4]. Speaker-independent systems train on data from several speakers to generalize and are suitable for text transcription and voice-activated applications. Our project is a speaker-independent system trained on data from lip movements (lip features or visemes) extracted from the input video file. The input will have many parameters like height, width, and frame rate. Our system emphasizes the same frame rate. It extracts lip features from each frame and stores them. A problem found is that there will not be any perceivable difference between the two frames. Also, a training dataset cannot provide apt text matches when trained with a different number of frames. Thus, we go by concatenating a fixed number of frames and classify a sequence of visemes directly to text rather than to phonemes [5]. The system comprises two segments: one being the feature extraction system that extracts lip features and makes it into a visual feature cube, while the other being a Convolutional Neural Network trained on a rich dataset, which matches visemes to the corresponding text.

The paper is organized as follows: Section 2 explores the related literature; Section 3 describes the dataset used in the experiment; the proposed technique is explained in Section 4, followed by an analysis of results in Section 5 while Section 6 concludes the paper.

II. Existing Models

In VSR systems, only the lip movements provide a significant contribution to knowledge retrieval. Many approaches are used in the literature to extract different features for LR systems.

A. Lip Feature Extraction in YIQ domain

This method proposed by [6] converts the video sequence in the Red Green Blue (RGB) domain to the Luminance Inphase Quadrature (YIQ) domain. The 'Y' component represents the luminance, while 'I' and 'Q' represent the chrominance information. Using the YIQ format helps localize lip features as human lips are usually brighter in the 'Q' space while the overall face is brighter in the 'I' space. A solid model can exploit this contrast for lip localization and lip tracking by segmenting the image in 'I' space.

B. Segmentation Method

In this method, [7] uses two approaches: edge detection and region segmentation. These methods detect the contour of the outer lip, and their results are combined using AND or OR fusion. They first found the mouth Region of Interest (ROI), which is then given to edge detection and region segmentation methods. The combination of results from these two methods provides the final outer lip contour.

C. Zernike Features

The model proposed by [8] aims to improve audio-visual recognition accuracy. The proposed solution includes extracting visual features using Zernike moments and audio features using Mel frequency cepstral coefficients on the visual vocabulary of independent standard words dataset on a series on the visual utterance. 'Viola-Jones' detector based on 'AdaBoost' method, used for face recognition and mouth portion, is calculated from the ROI bounding box's median coordinates. Zernike movements for ROI are computed for each frame resulting in 9x1 columns. One visual utterance is captured for two seconds forming 52 frames; therefore, the Zernike features for one visual utterance result in 468x1 for a single word. Further, Principal Component Analysis (PCA) is used to convert original features to independent linear variables possessing the most information. The performance, which was based on visual-only and audio-only features, resulted in 63.88% and 100% accuracy, respectively, which is relatively higher.

D. Deep Neural Networks

This Speaker-independent lip reading system by [9] uses techniques such as Linear Discriminant Analysis (LDA), Maximum Likelihood Linear Transform (MLLT), and Speaker Adaptive Training (SAT). The visual features are extracted in the following pipeline: the features are mean-normalized on a per-speaker basis and are decorrelated and reduced to a dimension of 40 using LDA and MLLT, and then, SAT is applied to normalize the variation in acoustic features of different speakers.

DNN is experimented as promising for speaker-independent lip reading even with limited training data and without a pre-training stage. The best-known result for a speaker-independent lip reading system is to use a hybrid system that uses MLLT followed by SAT.

III. Dataset

In the first module, viseme extraction is done using the DLIB module functions, which uses a pre-trained dataset, while the second module employs CNN that uses MIRACL-VC1 dataset [10].

A. Shape Predictor

MIRACL-VC1 is a trained dataset for dlib used for matching visemes, called "shape predictor 68 face landmarks". It provides the means to match facial features. The interface is provided through predictor and detector classes from the dlib package. The face detector used is made using the classic Histogram of Oriented Gradients (HOG) feature combined with a linear classifier, an image pyramid, and a sliding window detection scheme.

The landmark points from 48 to 68, shown in Fig. 1, are assumed to approximate the lip portion. So, those landmarks are considered as edge points while cropping.

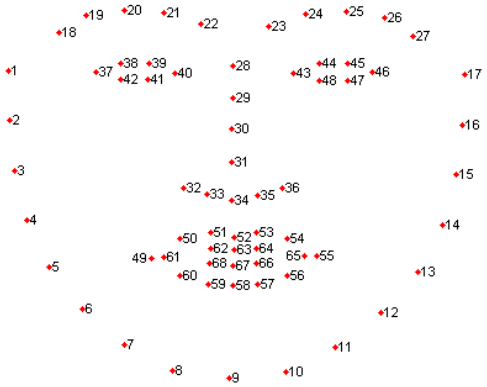


Fig. 1. Template 68-Point Facial Feature Image [11]

TABLE I. Words and Phrases available in MIRACL-VC1 dataset

ID	Word	ID	Phrase
01	Begin	01	Stop navigation
02	Choose	02	Excuse me
03	Connection	03	I am sorry
04	Navigation	04	Thank you
05	Next	05	Good bye
06	Previous	06	I love this game
07	Start	07	Nice to meet you
08	Stop	08	You are welcome
09	Hello	09	How are you
10	Web	10	Have a good time

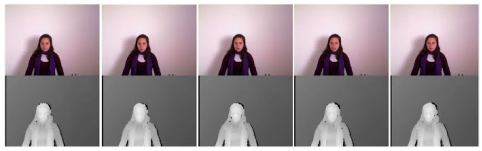


Fig. 2. Sample Color and Depth Image Frames [10]

B. MIRACL-VC1

MIRACL-VC1 is a lip reading database that includes both depth and color images as features. It facilitates multiple research areas such as speech recognition, face detection, and biometrics. Fifteen speakers (five men and ten women) who are positioned in the view of a Microsoft Kinect sensor utter ten times a set of ten words and ten phrases as shown in Table I. Each example in the dataset comprises color and depth images, both of size 640x480, synchronized. The sample color and depth images are shown in Fig. 2. The dataset contains a total number of 3000 examples (15 x 10 x 10 = 1500 images - color and depth images each). Our system utilizes only color images of words.

IV. Design and Implementation

The first step is extracting lip features from the video. The features are further given to a 3D CNN [4] that can classify visemes to the corresponding text. These two functionalities are separated into two modules: the pre-processing module and the CNN module. The input video file is pre-processed to extract lip features from the facial features and fed to CNN to

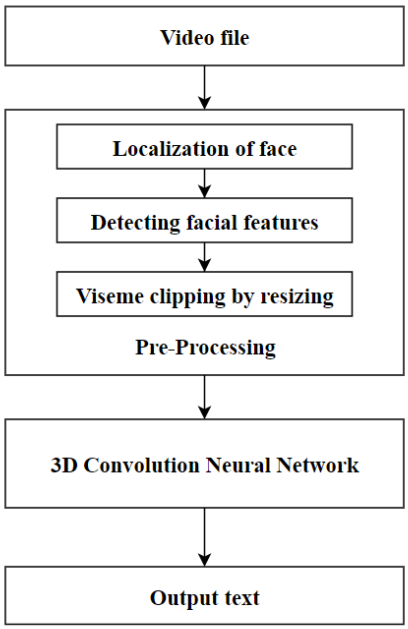


Fig. 3. Flow Chart of the Proposed Design

classify the visemes to text.

The flowchart shown in Fig. 3 is explained as follows.

A. Pre-Processing

The visemes need to be extracted from each frame. For this, the video is broken into individual frames first. Since the frame rate differs from one video to another, they need to be equalized to have the same frame rate (30fps). The processed frames are passed to the face tracking module.

B. Face Tracking

Facial tracking obtains data about still images and video sequences by automatically tracking the facial landmarks. Specific facial landmarks mapped, such as 48 face landmarks, 68 face landmarks, are available. It involves two steps: Localization of face and Detection of key facial structures. Since we do not need all the points in the frame, only the facial region is tracked first and using 68 facial landmarks, the key features are detected. We use the frontal face detector and shape predictor modules of the dlib package to achieve this.

1) Localization of Face: A pre-trained Histogram of Oriented Gradients (HOG) with Linear SVM Object Detector or deep learning-based algorithms can be applied to localize the face. The aim is to obtain the (x, y) coordinates of the face (formed as a bounding box) through these methods.

2) Detection of Key Facial Structures: A variety of facial landmark detectors are available that try to localize and label the following facial regions effectively:

- Mouth
- Right eyebrow
- Left eyebrow
- Right eye
- Left eye
- Nose
- Jaw

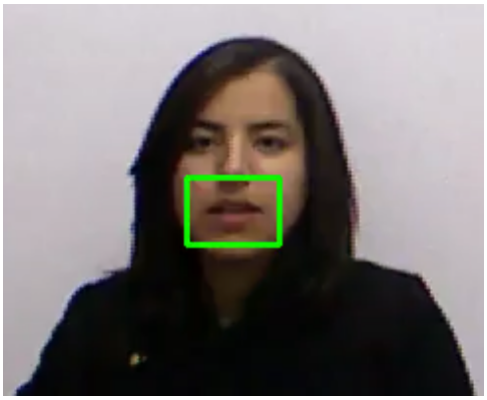


Fig. 4. Lip Region Bordered Input Video

The dlib library uses One Millisecond Face Alignment with an Ensemble of Regression Trees face detector from Kazemi and Sullivan [12]. The method works using the following:

1) The image is projected into a normalized coordinate system from which features are extracted. This process is repeated until convergence.
2) Prior probabilities on the distance between pairs of input pixels to boost the algorithm to work efficiently on a large number of relevant features.

The method builds an ensemble of regression trees on the training data to estimate the facial landmark positions by identifying pixel intensities that correspond to these landmarks themselves. This library, coupled with OpenCV, can provide a detector that can capture the necessary points, in our case, the lip visemes coordinates, as shown in Fig. 4.

C. Resizing

The tracked facial images can be of any angle of view. The speaker would have spoken the phrases either by looking straight into the camera or while looking somewhere. Since this poses a difficulty in detecting facial features, as some features may be lost, we can either restrict the speakers where to look or cropping and resize the image only to contain the lip region. The face images may be in different sizes. So, the detected faces are clipped and resized into the same size (30x48), which helps the CNN process them efficiently. It can be done by finding out the lip region edge points and cropping the image's desired portion. The resized images are shown in Fig. 5.

D. Convolutional Neural Network

Convolutional Neural Networks are most commonly used for image processing tasks [13]. The CNN architecture used is shown in Fig. 6. The model inputs a sequence of visemes (gray-scale) of dimension 15x30x48x1. The input is processed

Fig. 5. Cropped and Resized Visemes

```
---3D CNN for Visual Speech Recognition---
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1 (Conv3D)               (None, 13, 28, 46, 8)     224
_________________________________________________________________
batch_normalization (BatchNo (None, 13, 28, 46, 8)     32
_________________________________________________________________
activation (Activation)      (None, 13, 28, 46, 8)     0
_________________________________________________________________
conv2 (Conv3D)               (None, 11, 26, 44, 16)    3472
_________________________________________________________________
batch_normalization_1 (Batch (None, 11, 26, 44, 16)    64
_________________________________________________________________
activation_1 (Activation)    (None, 11, 26, 44, 16)    0
_________________________________________________________________
conv3 (Conv3D)               (None, 9, 24, 42, 32)     13856
_________________________________________________________________
batch_normalization_2 (Batch (None, 9, 24, 42, 32)     128
_________________________________________________________________
activation_2 (Activation)    (None, 9, 24, 42, 32)     0
_________________________________________________________________
pool1 (MaxPooling3D)         (None, 9, 11, 20, 32)     0
_________________________________________________________________
flatten (Flatten)            (None, 63360)             0
_________________________________________________________________
fc1 (Dense)                  (None, 32)                2027552
_________________________________________________________________
batch_normalization_3 (Batch (None, 32)                128
_________________________________________________________________
activation_3 (Activation)    (None, 32)                0
_________________________________________________________________
dropout (Dropout)            (None, 32)                0
_________________________________________________________________
fc2 (Dense)                  (None, 10)                330
_________________________________________________________________
softmax (Activation)         (None, 10)                0
=================================================================
Total params: 2,045,786
Trainable params: 2,045,610
Non-trainable params: 176
_________________________________________________________________
None
```

Fig. 6. Architecture of CNN used

by three Conv3D layers of 8, 16, and 32 neurons each. This stacking of three layers learns low-level features like edges and lines of the viseme and gradually high-level features such as lip movements and their sequence patterns. A batch normalization and activation layer follow each such layer. The features learned from these layers are sampled down by a max-pooling layer and vectorized using the Flatten layer. The features learned are then passed to a fully-connected layer of 32 neurons with L2 regularization, followed by batch normalization and activation layers, and given to the softmax classifier of 10 neurons matching the number of output classes. Table II shown below, presents the hyperparameters used in the CNN architecture.

V. Result Analysis

This paper uses accuracy to evaluate the experiment, along with precision, recall, and F-measure metrics. Eq. 1, 2, 3, and 4, respectively show the formulae for computing these metrics.

$$Accuracy\& = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision\& = \frac{TP}{TP + FP} \tag{2}$$

TABLE II. Hyperparameters for the CNN model

Parameter	Value
Kernel size	3x3x3
Stride	1x1x1
Pool size	1x3x3
Activation	ReLU
Optimizer	SGD
Learning rate	1e-2
Regularization factor	1e-2
Dropout factor	20%
Batch size	32

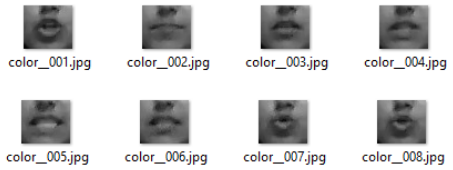


Fig. 7. Visemes of the Second Utterance of 'Choose'

TABLE III. Class-wise metrics

Metrics & Words	Precision (%)	Recall (%)	F-measure (%)
Begin	85.71	88.89	87.27
Choose	90.48	73.08	80.85
Connection	90.48	76.0	82.61
Navigation	78.26	75.0	76.6
Next	84.62	91.67	88.0
Previous	71.43	83.33	76.92
Start	73.68	70.0	71.79
Stop	87.5	60.87	71.79
Hello	45.95	89.47	60.71
Web	80.0	64.0	71.11
Weighted avg.	**80.24**	**76.89**	**77.40**

$$Recall\& = \frac{TP}{TP + FN} \tag{3}$$

$$F - measure\& = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

We concatenate the frames for each word to form a training example. Since the number of frames is different for each word due to utterance duration variation, we fixed the number of frames to 15 and padded the sequence with fewer than 15 frames with a viseme for a closed mouth. This padding method represents humans' closed-mouth position while we are not speaking, facilitating more human-like processing. The frames fewer than 15 for the word "Choose" are shown in Fig. 7.

Fig. 8 presents the confusion matrix obtained for the proposed model. This matrix shows that the model can robustly classify the viseme sequence to the target text. Table III lists the metrics class-wise. The F-measure for the model also shows that the classifier is more generalized and not biased towards any class.

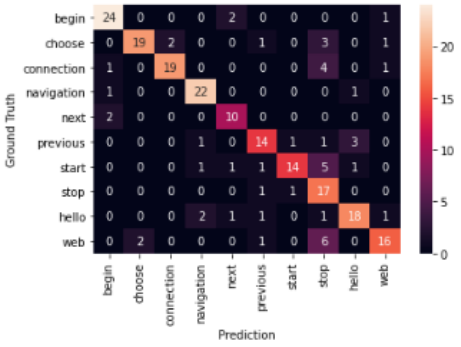


Fig. 8. Confusion Matrix

TABLE IV. COMPARISON OF THE PROPOSED METHODS WITH STATE-OF-THE-ART METHODS

Model	Accuracy
Borde et al., [8]	63.88%
Garg et al., [14]	56%
Proposed model	**76.89%**

Table IV compares the accuracy of various state-of-the-art models with the proposed model. Our model achieves about 76.89% accuracy, which is a significant improvement over the state-of-the-art models.

VI. CONCLUSION

This paper presented a combined approach of visemes concatenation and 3D Convolutional Neural Networks for Speaker-independent Visual Speech Recognition. We used dlib's face detection module to localize the face features in each frame, and with the help of 68-facial landmarks, we extracted the lip portion. The extracted visemes are cropped and resized to avoid them from being at different angles, improving the classifier's performance. We concatenated these frames of each word to generate an input feature. To fix the variation in the number of frames due to each word's utterance duration, we fixed the number of frames at 15. The 3D CNN learns from the sequence of visemes, the pattern for each word. The low-level and high-level features are appropriately learned from the hidden CNN layers. Our experiment shows that this approach outperforms the state-of-the-art models by improving the classification accuracy.

REFERENCES

[1] A. Thanda and S. M. Venkatesan, "Audio visual speech recognition using deep recurrent neural networks," in *IAPR workshop on multimodal pattern recognition of social signals in human-computer interaction*. Springer, 2016, pp. 98–109.

[2] J. Wang, J. Zhang, K. Honda, J. Wei, and J. Dang, "Audio-visual speech recognition integrating 3d lip information obtained from the kinect," *Multimedia Systems*, vol. 22, no. 3, pp. 315–323, 2016.

[3] E. Petajan, B. Bischoff, D. Bodoff, and N. M. Brooke, "An improved automatic lipreading system to enhance speech recognition," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1988, pp. 19–25.

[4] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, "3d convolutional neural networks for cross audio-visual matching recognition," *IEEE Access*, vol. 5, pp. 22 081–22 091, 2017.

[5] N. Alothmany, R. Boston, C. Li, S. Shaiman, and J. Durrant, "Classification of visemes using visual cues," in *Proceedings ELMAR-2010*. IEEE, 2010, pp. 345–349.

[6] T. N. Sengupta, "S.: Lip localization and viseme recognition from video sequences," in *Fourteenth National Conference on Communications*, 2008.

[7] U. Saeed and J.-L. Dugelay, "Combining edge detection and region segmentation for lip contour extraction," in *International Conference on Articulated Motion and Deformable Objects*. Springer, 2010, pp. 11–20.

[8] P. Borde, A. Varpe, R. Manza, and P. Yannawar, "Recognition of isolated words using zernike and mfcc features for audio visual speech recognition," *International journal of speech technology*, vol. 18, no. 2, pp. 167–175, 2015.

[9] I. Almajai, S. Cox, R. Harvey, and Y. Lan, "Improved speaker independent lip reading using speaker adaptive training and deep neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2722–2726.

[10] A. Rekik, A. Ben-Hamadou, and W. Mahdi, "A new visual speech recognition approach for rgb-d cameras," in *International conference image analysis and recognition*. Springer, 2014, pp. 21–28.

[11] P. Huber, *Real-time 3D morphable shape model fitting to monocular in-the-wild videos*. University of Surrey (United Kingdom), 2017.

[12] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874.

[13] S. L. Rose, L. A. Kumar, and D. K. Renuka, "Deep learning using python," 2019.

[14] A. Garg, J. Noyola, and S. Bagadia, "Lip reading using cnn and lstm," *Technical report, Stanford University, CS231 n project report*, 2016.

Security Issues in Near Field Communications (NFC)

Arwa Alrawais
College of Computer Engineering and Sciences
Prince Sattam Bin Abdulaziz University
Al-Kharj 11942, Saudi Arabia.

***Abstract*—Near Field Communications (NFC) is a rising technology that enables two devices that are within close proximity to quickly establish wireless contactless communications. It looks intuitively secure enough and various applications like ticketing, mobile payments, access grant etc. are taking advantage of NFC and flooding into the market in recent years. However, is it worth to trust such applications at the risk of leaking the user's private information? This paper surveys NFC vulnerabilities and exploits different kinds of security attacks. Upon surveying related materials, the paper covered possible solutions that could defend against those security threats. Furthermore, attacks and countermeasures evaluation in terms of practicality and cost have been further investigated.**

***Keywords*—Near Field Communications (NFC); NFC attacks; NFC countermeasures; NFC vulnerabilities**

I. Introduction

NFC (Near Field Communications) provides bidirectional, wireless, contactless communications for two NFC enabled devices or NFC tags within a short transmission range of less than 10 cm. It is derived from the Radio Frequency Identification technology, or RFID, whereas RFID is only capable of one-way transmission. NFC is based on inductive coupling to connect two NFC devices, or tags to establish communication at a central frequency of 13.56MHz, which is supported by ISO14443 standards.

NFC has three working modes, peer-to-peer mode, read/write mode, and NFC card emulation mode. The peer-to-peer working mode allows two NFC devices to transmit data between them. Read/write mode enables the NFC devices to access certain digital data. NFC card emulation mode, perhaps is the most interesting working mode, makes the NFC devices function as a NFC card. Based on activeness of the involved NFC devices or tags, the communication modes could be classified as active-active, active-inactive, inactive-active communication modes.

As an emerging technology, NFC has a promising and broad future to be applied in various kinds of applications. Currently, many key players of electronic communication market are involved in the NFC development [1], such as HP, Philips, Motorola, MasterCard, VISA, Panasonic, Microsoft, Gemalto, Vodafone, Siemens etc. NFC provides us with convenient tools like e-ticketing, electronic wallets, financial transactions, smart posters, etc. [2].

Among those wireless communications like WiFi, RFID, ZigBee and so on, why does electronic market favor NFC applications? NFC has the shortest transmission range and also the smallest data rate among the wireless communications. This means NFC has the advantage of quickly building a private communications within short distance. Table I shows that NFC is faster and easier to set up.

NFC appears to resist malicious attempts since they could only happen within a really short range. But is it safe enough for the users to rely on NFC applications instead of doing things in the conventional ways? The answer is negative. For one thing that NFC is a measure of wireless communication, which makes it vulnerable to eavesdropping, data corruption, and jamming attack. For another, NFC technology itself doesn't include strong security scheme to protect those applications that are built upon it. This leaves the job to the software designers and developers to seek ways to avoid any threats that could be caused by malfunction.

Though a NFC communication happens within close proximity, it doesn't mean that NFC is resistant to eavesdropping, jamming, data corruption, and other attacks towards wireless connection. Plus, NFC shares the basic standards and techniques with the proximity RFID technology. Some of the attacks that could be launched against RFID communications are major threats to NFC. How to defend against relay attack is an open problem in NFC communications, just like it's an unresolved problem in other wireless communications. Researches [4], [5], exploit the possibility of applying relay attack upon NFC communication. Imagine you have to use your NFC card to gain access to a building. What a malicious attacker could do is that he can attach a small receiver to the gate RF reader and record the signal sent by a legitimate NFC card. In this case, when a legitimate NFC card comes close to the reader, it thinks that it is sending signals to the reader. In fact, it is the "recorder" that is listening to the signal and tries to make a copy. Then the adversary could take advantage of this copy to do things like, clone a NFC card or use it to gain access.

As mentioned before, NFC doesn't provide any security mechanism to protect its communication, which leads to users' privacy being exposed to air. Even the Secure Element designed by Google to plot in a Google NFC device isn't secure as it sounds. The author in [6] addresses some issues when it comes to malware threats. The android operating system has proven to be vulnerable to malware attack, let alone the embedded SE, which is an obvious weak point when some malicious softwares try to gain information stored in SE through OS. For rooted devices, the SE access PIN is hashed using SHA256 and stored in the device rather than the SE. An attacker can brute-force the PIN and access the SE.

TABLE I. COMPARISON AMONG NFC, RFID, IRDA, BLUETOOTH [3]

	NFC	RFID	irDa	Bluetooth
Set -up time	¡ 0.1ms	¡ 0.1ms	˜ 0.5s	˜ 0.6s
Range	Up to 10cm	Up to 3m	Up to 5m	Up to 30m
Usability	Human centric Easy, intuitive, fast.	Item centric Easy	Data centric Easy	Data centric Easy
Selectivity	High given Security	Partly given	Line of sight	Who are you?
Use cases	Pay, get access, share, initiate service, easy set up	Item tracking	Control and exchange data	Network for data exchange, headset
Consumer experience	Touch, wave, simply connect	Get information	Easy	Configuration needed

The rest of this paper is organized as follows. In Section II, the paper summarize some related papers that talk about NFC technology and its vulnerabilities. In Section III, an illustration of the possible threats towards NFC technology and NFC applications along with the corresponding countermeasures to deal with these issues is introduced. At the end of Section III, an evaluation of the attacks and protection methods according is provided. At last, the author draw a conclusion of the study and point out some open problems about NFC in Section IV.

II. RELATED WORK

The growing number of released NFC applications raise concerns about its security issues. Lots of researchers have attempted to analyze the vulnerabilities of NFC technologies. The threats fall into two categories according to two characteristics the threats are aiming at, issues that intend to happen to wireless communications and dangers of malfunction of NFC applications or the operating system that carries the NFC softwares.

The authors in [6]–[10] make the point that eavesdropping attack is still possible in NFC connection. [7] claims that an antenna that is placed within distance of 10m can still "overhear" the data sent by an active NFC devices. This distance drops to 1m when the device is on passive working mode. Still, it makes eavesdropping possible. Thus, it opens the door for other threats like data corruption, data insertion, etc. Also, [7], [8] point out that using a RFID jammer or other devices that emit RF signals can easily jam or corrupt the data transmitted between two NFC devices. [9], [10] indicate that NFC is vulnerable to data modification attack, which is obvious since NFC doesn't encrypt the exchanged message. In addtion, [10] lists other potential attacks like data corruption, data insertion, and man in the middle attack. After trying to implement an secure offline payment application, Van Damme et al. [9] state that current technology is not sufficient to provide for a completely secured system not only because heavy use of cryptography will increase overhead, also because the hardware they used has limitations that slow down transaction speed and increase code complexity. In [11], the authors investigate man in the middle attack in NFC communications through performing a real time implementation in contactless payment system. They conduct man in the middle attack in NFC communications between passive tag and active terminal. Their results reported potential vulnerabilities in NFC communications due to the separation between payment card and point of sale.

In [4], [5], [12], the researchers claim that relay attack is also a big security concern in NFC. [12] even says that NFC is particularly vulnerable to relay attack and the authors provide countermeasures such as monitoring additional delay in propagation time, asking user to perform verification and integrating location information into transaction to protect NFC communication against relay attack. [4] points out that a malicious user could apply relay attack to gain access to Secure Element (SE) in a Google NFC device to pretend that he/she is in physical possession of the device. Michael Roland [4] mentions other malware threats about NFC devices. Both [5] and [4] talk about launching denial-of-service attack by simply touching a NFC devices using any arbitrary tags.

Upon cracking up the exchanged message between two NFC entities, [13] indicates that privacy infringement happens since NFC standards do not provide unlinkability between user message and public keys are constantly used in key agreements and the authors proposed a conditional anonymity using dynamic public key to solve this problem.

Collin Mulliner [14] analyzed vulnerabilities in NFC enabled smart phones. He developed a NDEF security toolkit to test the target NFC smart phone. His observation was NFC Data Exchange Format (NDEF) could be easily manipulated. He noticed that if the message sent by a smart poster is inserted with consecutive white spaces the user is unable to verify the security information displayed on the screen. And there is a weakness in the NDEF fuzzing process that if the value for the length of the payload field is set as two specific value, the phone will crash. After four crashes in a row, the phone automatically powered down. Another dangerous feature of the NFC enabled smart phone is that only the last 10 characters of the hostname is shown in the screen, which could fool the user into believing they are visiting the desired website while the phone is loading another malicious site.

NFC technology itself does not incorporate security mechanism to protect its framework and platform android operating system, for running NFC applications are weak to malware amongst OS [6]. Leakage of privacy happens in situation such as a Google NFC device user finishes grocery shopping using Google Wallet. [6] and [4] state that malicious software could plant itself without the user's knowledge and access SE to gain security information. Another work in [15] investigates the security of NFC in mobile payment system where the NFC

Fig. 1. Eavesdropping Attack

Fig. 2. Jamming Attack

tag could contains malicious threats to redirect users to install e.g., malicious code without users knowledge.

III. Analysis of NFC Security Issues

NFC isn't secure as it looks. In this section, a possible NFC related attacks and their countermeasures are surveyed. Each attack will be explained using illustrative example, and then suggested countermeasures will be introduced to mitigate the risks of these attacks. At the end of this section, an evaluation of these attacks and countermeasures based on several factors such as cost of the attack and countermeasure and practicality of the attack and countermeasure are provided.

1) Eavesdropping: One common attack on wireless communications is eavesdropping attack, and Unfortunately, NFC technology is not secure against this attack [6].

The limited communication range of NFC devices which is few centimeters (about 10 cm) doesn't prevent the risk of eavesdropping attack completely. Any attacker with sufficient equipments can listen to the communication between two NFC devices. The main issue is how close an attacker needs to be able to conduct eavesdropping attack against NFC devices. In fact, this depends on the equipments of attacker such as antennas used, receiver used, and the environment of the attack such as noise, emitted signal. Other important factors such as the location of the attacker and position, the location of NFC device affect the attack operation. In addition, the communication mode affects the attack since there is a difference between listening to a NFC device in passive or active mode [9]. It is more difficult to listen to a NFC device in a passive mode because the target device may draw its source power from the electromagnetic field that is generated by the active device. According to [10] eavesdropping attack can be conducted up to a distance of 10m, when a NFC device is sending data in active mode, whereas this distance is significantly decreased to about 1m when the sending device is in passive mode.

Fig. 1 illustrates how an attacker that is closed to NFC environment can listen to the communication between two NFC devices. An attacker with sufficient knowledge and equipment such as Proxmark is capable of capturing NFC communication. Proxmark is an open source and powerful device currently available for researching RFID and Near Field Communication systems. Proxmark has the feature to snoop NFC traffic between a reader and tag it costs less than $500, more information about Proxmark can be found here [16].

Establishing a secure connection and using standard encryption algorithms between two NFC devices can protect against eavesdropping attack. A standard key agreement protocol such as RSA or Elliptic Curves could be used to establish a shared secret key between two NFC devices. The secret key then can be used to encrypt the communication using symmetric key algorithm such as AES or 3DES [10]. This countermeasure will ensure the confidentiality in NFC communication and will protect against eavesdropping attack.

2) Denial of Service: Wireless communication can be very vulnerable to Denial of service attacks or as knows Dos attacks. The results of Denial of service attacks can be anything from degradation of the wireless communication to a complete loss of availability wireless service. By launching a Denial of service attack, a malicious attacker can attempt to make a NFC deceive or a reader unavailable to its intended users. In this section, several Denial of service attack scenarios has been discussed.

One scenario of denial of service attack is by using a jamming device that target NFC environment [10]. The goal of jamming is to disrupt communications between two NFC devices.

Fig. 2 shows a malicious attacker with jamming device such as RFID jammer transmit a signal that interfere with the transmission between a mobile NFC phone and a reader of a service provider. This interference can destroy the transmitted data and cause denial of service. Almost there is no way to prevent jamming; however, there is a solution to deals with this scenario by continuously trying to detect jamming attack.

The solution is to let NFC devices check the radio frequency field while transmitting. This means the sending device could continuously check for such an attack scenario and could stop the data transmission when someone tries to jam the transmission.

Another Denial of service attack has been explained by [14], where the goal of the attack is to destroy trust relationship between customers and the service provider.

The following steps explain the scenario of this attack.

- A malicious attacker or a malicious competitor creates a tag that causes an NFC mobile phone to crash after scanning.
- The malicious attacker will sneak to the victim or the service provider and place the malicious tag on top of service provider tag.
- Any customer visit the victim or the service provider to get a service using NFC mobile phone will crash after scanning.
- The malicious tag cannot be linked to phone crash accident since it looks just like a normal tag, and this incident can destroy trust relationship between customers and the service provider

There is no solution for this attack; however, it can be detected using some tools such as fuzzing. More detail about this tool can be found in [14].

Another scenario of denial of service attack can be launched using empty NFC tag. Riyazuddin [7] indicated that just touching an NFC device with an empty tag causes a reaction of the device. The device will generate and an error message which is an easy way to occupy the device and make it unavailable. Adding a mechanism of controlling the NFC device such as NFC switch can help to prevent this attack scenario. The drawback of this solution is that the user has to turn on and off the NFC functionality each time when he needs to scan.

3) Phishing: Phishing attack is the act of attempting to obtain sensitive information such as passwords, and credit card details by masquerading as a trustworthy entity in an electronic communication. Phishing attacks could easily be performed against NFC environment by modifying or replacing NFC tags.

The following steps and Fig. 3 explains how a malicious attacker can harvest sensitive information such as credit card information by launching a phishing attack against parking meter that uses NFC technology for completing the process of payment [8]:

- The attacker first create a malicious tag that contains false information such as the URL link that directs to a phishing site.
- The attacker will find a parking meter that uses NFC technology and replace the original tag at the parking meter with the malicious tag.
- In order to pay the meter fee, a victim with NFC mobile device such as Samsung phone scans the park meter tag in order to pay the required fee.
- The user will be asked to install a malicious app com.porkmobile which is basically a Web view to the phishing site.
- The user will enter sensitive information such as credit card information using the installed malicious app, and the attacker will collect these sensitive information.

There are several countermeasures can be used to prevent or mitigate phishing attack risk. One crucial factor of conducting phishing attack is to deceive the users by masquerading as a trustworthy entity, however; people who are aware about this phishing attack are difficult to deceive. User awareness and education about phishing attack is an important countermeasure since it helps to minimize the number of successful attacks. Cautious users will recognize the process of requiring installing new application with suspicious name and will investigate more about the name and originality of the application.

Gerald et. al. [5] suggests using signatures on tags and transporters and they indicate that would be suitable way to overcome this issue. Furthermore, applications market for NFC mobile such as Google's market can play a crucial role to prvent malicious applications that are suspicious to phishing.

A. Data Insertion

Data insertion attack goal is to insert a message into exchanged data between two NFC devices, when the answering device takes time to answer the original device. The attack can be launched only if the device has some delay that makes an attacker is able to transmit its message before the answering device. If both the attacker and the answering device transmit the data at the same time, the data will be overlapped and corrupted.

Data insertion attack can be launched between two NFC devices. The following scenario explains the attack steps:

- The attacker will place his malicious reader near the original reader device.
- The victim user will use the mobile NFC phone to transmit the data to the reader device.
- The malicious reader will reply directly to the victim user before the original reader.
- The original reader will reply to victim user after the attacker and the reply will be ignored by the victim's mobile NFC phone.

In order to prevent the data insertion attack between two NFC devices, there are three countermeasures can be employed. Firstly, the answering device should answer the original device with no delay. In this way the attacker would not be able to insert a message into the exchanged data between two NFC devices, because attacker can't be faster than answering device. Secondly, the answering device should listen to the channel while transmitting the data, so the device can detect any potential attack. Thirdly, establishing a secure channel between two NFC devices is the best approach to prevent any attack [10].

Fig. 3. Phishing Attack Targeting at Parking Meter

B. Data Modification

Data modification is different than data insertion where attacker inserts a message into the exchanged data between two NFC devices. In data modification, the attacker can modify the exchanged data between NFC devices, so the receiving device will receive some valid but manipulated data. The feasibility of data modification attack relies on the amplitude of modulation [10]. It is difficult to launch data modification attack against NFC environment when the coding modulation is 100% in modified Miller coding modulation, this is because in 100% modulation the attacker is not able to alter a bit of value 0 to a bit of value 1. Although if a bit of value 1 is coming first (i.e. with a probability of 0.5), the attacker is able to alter a bit of value 1 to a bit of value 0. In 100% modulation, two half bits for radio frequency signal on and radio frequency signal off are checked by the decoder. The attacker should perform two steps to make decoder recognize one as zero and zero as one. First step which is a feasible step where attacker makes a pause in the modulation that loaded with carrier frequency. Second step which is practically impossible, where the attacker makes a pause of radio frequency signal that is received by the valid receiver. In this step, the attacker tries to overlap the original signal and the sending signal to make the receiver's antenna get a zero signal. However, it is easy to conduct the data modification attack when the modulation is 10% modulation. In 10% modulation, the decoder compares and assesses signal levels 82% and full. The attacker attempts to insert a signal to the 82% signal, in order to make the 82% signal become visible as a full signal and the actual full signal appears as 82% signal. Therefore, the valid bit of the reverse value of the bit would be decoded by the decoder. In conclusion, the attacker is feasible in all bits for 10% modulation, whereas is not feasible for all bits in 100% modulation.

Another example of data modification is exchanging electronic business cards or pairing information. Because there is no encryption or authentication in the transaction protocol, the means of security to ensure authenticity, integrity, and confidentiality should be implemented in the application layer.

A current common protocol, NFCIP-1, does not include the means of security. In this situation, the attackers can disturb the communication and modify the data. As shown in Fig. 4, if B disturbs the communication between A and C, and the communication does not include encryption or authentication. Thus, B can modify the exchanged data between A and C.

There are several ways to protect against the data modification attack. First, the attacker would not be able to alter all the data transmitted by radio frequency link, if 106k baud in active mode has been used. It can be clearly seen that the active mode is important, however, this mode is vulnerable for eavesdropping attack. Furthermore, some bits in 106k baud can be modified. Second, the sending device checks constantly the radio frequency field while transmitting data to detect any potential attack. Third, establishing a secure connection between two NFC devices appears to be the best approach to protect against data modification attack [7].

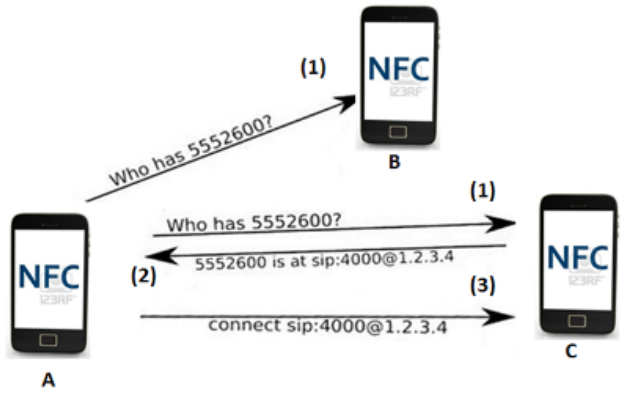


Fig. 4. Example of Data Modification over Peer Link.

C. Man In The Middle Attack

In Man in The Middle Attack (MITM), an attacker makes two parties believe that they are connecting to each other directly while in fact the whole conversation is directed by the attacker. In classical scenario, let assume Alice and Bob two parties want to talk to each other and Eve is the attacker who controls the entire conversation. Both parties Alice and Bob think they are receiving and sending data to each other whereas the whole data is coming from Eve.

Let's assume the same classic scenario but the link between the two parties Alice and Bob is an NFC link, where Alice would be in active mode and Bob uses passive mode. Alice wants to send data to Bob, so Alice generates the radio frequency field. The data can be eavesdropped by Eve, if Eve is closed enough, and actively disturbed the transmission to ensure the data hasn't been transmitted to Bob. In this situation, the attack can be detected by Alice through checking for any

active disturbance. Alice would disconnect the communications [10].

Let's assume the protocol continue and is not been checked by Alice. Eve would generate radio frequency field to be able to send the data to Bob. But, this would cause two actively radio frequency fields. The first one is generated by Alice and the second one is generated by Eve. Bob would receive a data that is not understandable. As a result, this situation is practically impossible for man in the middle attacks to be conducted.

Another scenario, let us assume the same classic scenario, but this time the two parties Alice and Bob would be in active mode. Alice sends data to Bob, and Eve is able to eavesdrop the data. Eve disturbs the transmission to ensure that bob has not received the data. Again, if Alice has not checked for any active disturbance, the protocol would continue. Let us assume that the protocol continues. In active – active communication radio frequency field has been turned off by Alice, so Eve can send data to Bob. Eve turned on the radio frequency field and sends the data. In this situation, Alice expects an answer from Bob. As a result she would listen, and receive a data from Eve. Alice would detect a problem in the protocol, and disconnect the protocol. Consequently, it is impossible for Eve to send and receive a data from the two parities. In conclusion, in real world man in the middle attack practically is unfeasible to be conducted between two NFC devices [10].

As mentioned that man in the middle attack is practically impossible in NFC link. However, it is highly recommended to use active – passive communication mode. In addition, in order to detect any disturbance that launched by any attack, the active party should listen and check the radio frequency field during the transmission.

D. Data Corruption

The attacker needs a high power to be able to corrupt the data while transmitting between two NFC devices. However, this attack can be detectable, because the NFC devices can check the radio frequency filed during the data transmission and detect the type of attack. In addition, to perform data corruption attack the attacker requires more power than that can be detected by NFC device. Consequently, this attack can be detected by NFC devices [7].

E. Relay Attack

Relay attack is a type of man in the middle attack where the attacker attempts to manipulate the communication through relays the verbatim messages between two devices. Relay attack can be performing only if at least one of the attack devices supports card emulation. There are many possible scenarios to perform this attack.

The first scenario is when NFC is always on in smart phone, even if the phone not in use. A smart phone with payment application can make a transaction easily. As a result, this makes the phone vulnerable for relay attack. In this scenario, there are two attackers that are connected to each other through the Internet. The first attack has a proxy device and the second attack has a relay device with two NFC enabled devices or smart phones. In a public place such as public transportation where many people gather waiting for bus or metro to arrive. The attacker with rely device can get close to the victim's smart phone. Then, the proxy device performs NFC payment at payment station. The connection between the payment station and the victim's smart phone relays on the two devices. Francis et al. described the relay attack in NFC environment as shown in Fig. 5 [17].

The second scenario can be performed in modern smart phone where there are some privileges (known as jail breaking or rooting) that give you a full control over the smart phone. But, it also loses some security features of the smart phone, such as the application sandbox. In addition, the security features protect the secure elements where the NFC payment application resides. Thus, on a rooted smart phone, the secure elements are vulnerable more. In this scenario, the attacker attempts to let the user install a malicious application. The victim believes that he got the application access rights for the feature. Then, the malicious application would get the access right to execute the features. In the meantime the application gets an access to the secure elements, and informs the attacker over the Internet. Now, the attacker is able to make a payment by the victim's payment details [17].

To protect against relay attack, the smart phone's user should ensure that NFC in the smart phone is always off. In addition, the smart phone's user should preserve the security features to detect any malicious activity in any installed application.

F. Skimming Attack

There are two modes of secure element — external mode and internal mode.

1) External Mode: To emulate a tag, it requires smart card chips in NFC devices. In external mode, an external reader accesses the secure element and cannot distinguish between a smart card and an NFC device with a secure element. For example, there is a credit card applet in the secure element that turns the NFC handset into a mobile payment device.

2) Internal Mode: In internal mode, the host controller accesses the secure element (reading and altering). The running applications on the host controller of the handset can alter the information in the secure element. Hence, the users can remotely manage the information in the secure element by online connection (GPRS, Wi-Fi and etc.), also known as Over The Air (OTA) management. For example, when users use NFC for ticketing, ordinary smart card is a good choice. The tickets or money can be stored in the secure element remotely online.

In secure element, an index of applications is provided by both memory cards (NXP's Mifare Application Directory e.g.) and processor cards (JCOP e.g.). Therefore, it is vulnerable to a third party players because other applications in secure elements are exposed. The problem exists not only in NFC technology but also in other smart card industry.

G. Spoofing Attack

There is an unique ID for each contactless smart card chip (ISO14443 A: UID, ISO14443 B: PUPI, Felica: IDm). The length of them are 4, 7 or 10 Bytes. When collision happens

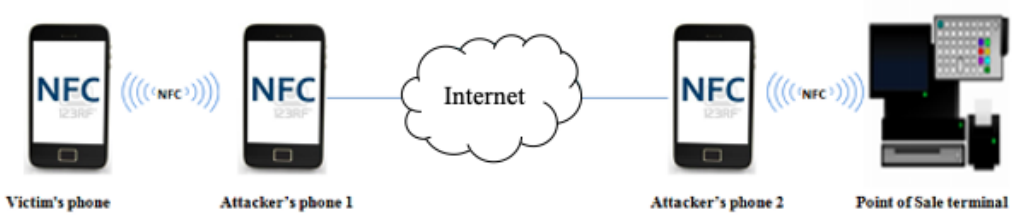


Fig. 5. NFC Relay Attack.

during the reading process, the unique ID is needed to prevent it by identification. The ID can be already acquired during the selection process of the transponder. The reading process does not include encryption or authentication of the reading device.

To prevent the collision, the unique ID is specified in the standard. A simple hardware like OpenPICC [18] can spoof someone's identity by simulating an ID. Therefore, if an application uses a fixed unique ID, it is easy to leak the holder's privacy. Because the reading process of the transponder does not include encryption, it is easy to eaves dropping the communication between the reader and the smart card chip to get the fixed unique ID. To avoid this situation, the unique ID can be created randomly when colliding, which is already used for NFC targets and e-passports [19]. It prevents the users from being tracked. However, it does not valid when victim is carrying an RFID transponder (a smart card or an NFC device).

H. Attacks and Countermeasures Evaluation

In this section, an evaluation and analysis between the surveyed attacks and their countermeasure is introduced. The evaluation is performed based on four factors which are attack cost, attack practicality, countermeasure cost and countermeasure practicality. The main goal of this evaluation and analysis is to evaluate and differentiate at the same time between mentioned attacks and countermeasures.

Attack cost describes the needed cost to perform an attack against near filed communication environment. The cost can be equipments such as jamming equipments or eavesdropping equipments, or the cost of the required time and effort to conduct the attack. Some attacks require purchasing extra equipments in order to launch the attack, for example, rely attack requires two NFC devices and proxy device and two involved attackers. On the other hand, there are several NFC related attacks that are not expensive and easy to launch such as denial of service attack.

Attack practicality is an important factor which describes the attacks quality of being practical, and the possibility of performing the attack. Not all NFC related attacks are practical; in fact some of the attacks are impossible to launch such as man in middle attack [10]. In addition, data modification attack is almost impossible for all bits in 100% modulation However; other attack such as eavesdropping, denial of service, phishing attack and rely attack are practical and can be launched with sufficient knowledge and equipment.

Countermeasures cost describes the needed cost to perform a countermeasure, such as extra resources or technical mechanisms. For some attacks such as denial of service attack the solution can be expensive since it requires hiring a security person or implementing a Closed-Circuit Television (CCTV) cameras system to monitor and prevent the access to the reader. However, other attacks such as relay attack require cheaper technical countermeasure such as turning off the NFC in the smart phone.

Another important factor is countermeasures practicality which describes the countermeasures quality of being practical, and the possibility of being performed. Some surveyed countermeasures are not practical to implement, for example, one of the data modification countermeasures is to use 106 K baud in active mode which will make NFC devices vulnerable to eavesdropping attack [7]. On the other hand, other suggested countermeasures are practical and very useful to implement such as establishing secure connection between NFC devices. Establishing a secure connection is very practical countermeasure and can be useful to prevent several NFC related attacks.

IV. Conclusions

Near filed communication is a promising technology and it is expected to be more integrated with future smart phones and to be an essential part of our daily lives. Master Card, Google and many payment services providers started to rely on near filed communication payment based technology which is anticipated to grow rapidly and broadly in the next few years. However; the security of near filed communication is still a concern and requires more deep analysis and further studies. This paper surveys many security threats, which are applicable to near filed communication, and covers countermeasures to protect against these threats. Near filed communication as technology cannot provide protection against many surveyed attacks such as eavesdropping or data modifications. Establishing a secure channel between NFC devices is a crucial mechanism to mitigate many security risks.

Future work of near filed communication could be how to design trustworthy operations of near field communication. Furthermore, other security related attacks should be more investigated such as NFC session hijacking, cloning attack, reply attack, and NFC skimming attack, which is reading an NFC device in a person's pocket. In addition, using near field communication in payment system impose many privacy issues which should be more studied and analyzed.

References

[1] A. N. Csapodi, Márton, "New applications for nfc devices," in *Mobile and Wireless Communications Summit*, 2007.

[2] F. Jeffrey, "The new paradigm for an interactive world [near-field communications]," *Communications Magazine, IEEE*, 2009.

[3] C. E. Ortiz, "An introduction to near-field communication and the contactless communication api," *http://goo.gl/icVkg*, 2008.

[4] M. Roland, "Practical attack scenarios on secure element-enabled mobile devices," in *International Workshop on Near Field Communications*, 2012.

[5] G. Madlmayr, J. Langer, C. Kantner, and J. Scharinger, "Nfc devices: Security and privacy," in *Availability, Reliability and Security*, 2008.

[6] S. H. Omkar Ghag, "A comprehensive study of google wallet as an nfc application," *International Journal of Computer Applications*, 2012.

[7] M. Riyazuddin, "Nfc: A review of the technology, applications and security." [Online]. Available: http://123seminarsonly.com/Seminar-Reports/023/46910687-Near-Field-Communications-Review.pdf

[8] I. Square, "Security risks of near field communication," *http://www.nearfieldcommunication.org/nfc-security-risks.html*, 2017.

[9] V. Damme, Gauthier, K. Wouters, and B. Preneel, "Practical experiences with nfc security on mobile phones," in *Workshop on RFID Security*, 2009.

[10] E. Haselsteiner and K. Breitfuß, "Security in near field communication (nfc)," 2006.

[11] S. Akter, S. Chellappan, T. Chakraborty, T. A. Khan, A. Rahman, and A. A. Al Islam, "Man-in-the-middle attack on contactless payment over nfc communications: Design, implementation, experiments and detection," *IEEE Transactions on Dependable and Secure Computing*, 2020.

[12] L. Francis, G. Hancke, K. Mayes, and K. Markantonakis, "Practical nfc peer-to-peer relay attack using mobile phones," in *RFIDSec*, 2010.

[13] E. H. et al., "Conditional privacy preserving security protocol for nfc applications," in *Consumer Electronics (ICCE), 2012 IEEE International Conference*, 2012.

[14] C. Mulliner, "Vulnerability analysis and attacks on nfc-enabled mobile phones," in *Availability, Reliability and Security*, 2009.

[15] M. Badra and R. B. Badra, "A lightweight security protocol for nfc-based mobile payments," *Procedia Computer Science*, vol. 83, pp. 705–711, 2016.

[16] "Proxmark3," *https://code.google.com/p/proxmark3/wiki/HomePage?tm=6*, 2007.

[17] R. Vermaas, "The security risks of mobile payment applications using near-field communication," 2013.

[18] "Openpcd http://www.openpcd.org/," 2007.

[19] I. Naumann, "Advanced security mechanisms for machine readable travel documents extended access," *Federal Office for Information Security, Tech. Rep.*, 2006.

Meezaj: An Interactive System for Real-Time Mood Measurement and Reflection based on Internet of Things

Ehsan Ahmad
College of Computing and Informatics
Saudi Electronic University
Riyadh, 11673

***Abstract*—Subjective well-being has a critical affect on progress and productivity vital for digital and strategical transformation. Increase in the suicide attempts of the college and university students is a clear indication of stress and anxiety among the students. Offering a fulfilling and healthy life to promote the life-long learning journey is also one of the important objectives of the Vision 2030 for modernization of the Kingdom of Saudi Arabia. Due to the multifaceted nature of subjective well-being, real-time mood measurement and reflection is a challenging task and demands using latest technologies. This paper aims to present Meezaj, an interactive system for real-time mood measurement and reflection leveraging the Internet of Things (IoT) technology. Architecture and workflow of the Meezaj system are discussed in detail. Meezaj not only promotes the sense of significance in the students, by indicating that their happiness matters in decision making, but also assists policy makers to identify factors affecting the happiness in an educational institute.**

***Keywords*—Subject well-being; happiness; IoT; Arduino; Vision 2030**

I. Introduction

Mental health contributes significantly in our perception about the world and how we react to it. Subjective well-being is considered to reflect the mental state of the people by exploring their experiences and state of mind about the life [1]. Although, subjective well-being investigation is affected by multiple parameters but is generally based on considering the emotional intelligence through measuring the positive and negative feelings [2]. Positive feelings are mainly reflected by happiness, pleasure, and joy while negative feelings are often reflected by stress, sadness, and angriness. Studies like [3], [4], [5], [6] and the research evidences provide by social psychologist D.G. Myers [7] suggest that happy people live longer, are more susceptible to engage in social interaction and undertake difficult tasks, are more willing to donate and help others, and are more open to new ideas.

The urge to find the most profound technologies for pervasive computing as visioned by *Mark Weiser* in 1988 has led to the emergence of intelligent integration of the physical and virtual objects and devices. The term *Internet of Things* (IoT) refers to the network of inter-connected heterogeneous *things* with the ability of data collection and exchange. Things may include portable personal objects like digital camera, smartphone, small sensors widespread in our surroundings like temperature sensor, or the large objects we interact with in our routine lives like smart TV, cars, buses, and trains [8], [9]. IoT technologies has revolutionized the measuring, monitoring, and controlling in almost all the aspects of human life. As reviewed by Suwimon and Sucha in [10], overall supply chain management process has been significantly improved by the use of RFID tags and communication infrastructure, crowd sourcing and participatory sensing is facilitating in finding parking and monitoring the overall health of vehicles in transportation, and ensuring assess to the healthcare in rural areas through integrating the devices and transmitting health data through smartphone has been possible due the exposure of the IoT technologies.

Due to the multi-faceted nature of subjective well-being, mood measurement and reflection in an institutional setting is a challenging task. Measures that are based on *subjective input* like Experience Sampling Method (ESM) and Day Reconstruction Method (DRM) are affected by personal bias and take a lot of work from the participant where they have to walk around with a beeper and papers to write down their input [11], [12]. Great deal of time and effort is then required to go through all of what the participants wrote. Measures that are based on *objective input* like physical activity, amount of sleep, and social interactions present the challenge of data collection (as the participants must use wearable devices) and the complexity of the data. Equally important as measurement, real-time reflection of the mood is essential for timely adjustment in the rule and regulations negatively affecting the mood of the employees or the students. Large amount and heterogeneity of the data hampers the real-time reflection of the mood. Another dimension of the mood reflection challenge is personalization of the working environment to maximize happiness among the individuals.

To cope with the above mentioned mood measurement and reflection challenges, this paper presents an interactive system named *Meezaj* (Arabic word of mood) to measure and reflect the mood of a community in an institutional setting. *Meezaj Mobile App* continuously collects responses to a short (and optionally long) mood surveys in specific durations of time. Questions for the long surveys are based on the Experience Sampling Method (ESM) approach. The data collected in terms of responses to the mood survey is then analyzed in real-time and mood of the participants is projected through the *Meezaj Mood Reflector*. Meezaj mood reflector is a IoT based custom

designed sculpture which not only aesthetically signifies the respective institute but can also interact with the other entities in the environment for personalization. The system is also equipped with a *Meezaj Web App* for detailed reports and administration tasks.

The next section describes need and background of this study. Section III presents a summary of the important existing applications for mood measurement along with their comparison. Section IV describes the proposed Meezaj system in detail for real-time mood measurement and refection. The architecture and workflow of the Meezaj system is specified with important algorithms. Section V concludes this paper with a glimpse of important future works.

II. Background

Kingdom of Saudi Arabia is going through a strategic and digital transformation for thriving knowledge-based economy. *Offering a fulfilling and healthy life* through healthcare transformation and promoting the *life-long learning journey* through educational transformation are two main objectives of this transformation under the umbrella of Vision 2030 [13]. Realization of such multifaceted transformation is quite challenging and demands for highly coordinated multidisciplinary efforts. The successful modernization of the existing infrastructure also requires to exploit modern information and communication technologies to measure, monitor, and reflect the effects and outcomes of policy updates. For example, in relation with health and education, it is essential to investigate that how the transformation of the learning paradigms (from traditional to blended to complete virtual) will affect the happiness and satisfaction of the students.

Increase in the suicide attempts of the college and university students is alarming. In United States alone almost 24,000 students attempt suicides annually [14]. Although not to that extant, but studies like [15], [16], [17], [18] have clearly indicated the signs of stress, and anxiety among the university students in Saudi Arabia as well. Traditionally, long paper-based surveys are conducted for subjective or objective input collection to investigate students' concerns about the infrastructure, curriculum, and educational paradigms. These surveys not only suffer with low response rate but also require considerable time and effort for data compilation and analysis. So for the middle and top management, timely response to a recent policy change may not be reflected in time. As a result, real-time mood measurement and reflection of the students and other stakeholders (faculty members, and the admin) is almost impossible with this paper-based approach.

According to Ministry of Communications and Information Technology (https://www.mcit.gov.sa/), with 188 registered mobile phones for every 100 Saudis, the Kingdom has the highest number of mobile users in the world. Equipped with various sensors, mobile devices also present the idea that survey participants don't have to carry any extra equipment for common mood measurement through ESM and DRM. These sensors can replace wearable devices for the objective input, which can drastically decrease the amount of resources needed [19]. Along with the mobiles phones, IoT being a convergence of multiple technologies, has also been successfully applied to different sectors of the society and has already shown promising results.

To realize real-time mood measurement and reflection, and to facilitate communication and data exchange among multidimensional stakeholders, we present an interactive system leveraging the mobile phone and IoT technologies. Our proposed Meezaj mobile app interacts with students to get their feeling in specific periods of time and analyzes the results to provide the true picture of mood to the middle-level (HoD, Dean) and top-level managements (Directors). Since students' mood is important to achieve the best abilities, skills and results, we seek to interact with their environment to improve the mood. Meezaj also supports interaction with several entities in the environment through IoT (hardware and software infrastructure) to maximise students' happiness. Meezaj provides automatic report generation for both individuals and the administration.

Although, Meezaj is targeting students, faculty members, and the admin staff in a typical University setting but due to it's scalable design, the system can easily be applied to other institutes like Banks, Airport, etc. by adjusting with their business rules.

III. Existing Applications

This section presents a summary of the most commonly used existing applications directly providing the mood measurement at institutional and personal level. Existing applications are also compared based on the features being offered to the perspective users.

a) Emooter: Equipped with web and mobile apps, *Emooter* facilitates team members to know work-related well-being [20]. Team members do the brainstorming to set well-being baseline and the feedback is provided on individual basis.

b) Culture Amp: To promote a healthy company culture, *Culture Amp* engages the employees by inquiring about their problems and helping HR in decision making [21]. Available is both web and mobile apps, the feedback tools cover entire employment life cycle; starting from the new hire surveys, onboarding, employee engagement surveys, single-question polls, and exit interviews.

c) Roundpegg: Recently acquired by the Achievers, *Roundpegg*, like Culture Amp, has a culture-first approach [22]. Employees are hired by assessing their fitness for an already defined "CultureDNA". Periodic customized surveys are conducted to know employees' feelings about their life at the company.

d) TINYPulse: Derived with several products for employees' feedback and performance management, *TinyPulse* conducts weekly surveys to know the pulse of the company [23]. To collect ideas, it also provides "Virtual Suggestions". Good work of the employees is acknowledged through "Cheers of Peers" approach.

e) MercuryApp: MrucuryApp, fetches employees' feelings from their responses to the daily reminder emails [24]. MercuryApp is considered a good fit for agile practitioners. To log their opinions a daily reminder feedback on the project is collected on daily basis.

TABLE I. COMPARISON OF POPULAR APPLICATIONS FOR MEASURING HAPPINESS

Application Name	Target Audience	Features			
		Analysis	*Web-based*	*Mobile app*	*IoT Communication*
Emooter	Companies	No	Yes	Yes	No
Culture Amp	Companies	Yes	Yes	Yes	No
RoundPegg	Companies	Yes	Yes	Yes	No
TINYPulse	Companies	Yes	Yes	Yes	No
MercuryApp	Companies	Yes	Yes	Yes	No
Mood Meter	Personal	Yes	No	Yes	No
Celpax	Companies	Yes	No	Yes	Yes
Peakon	Companies	Yes	Yes	Yes	No
Bluepulse	Companies	Yes	Yes	Yes	No
BambooHR	Companies	No	Yes	No	No
Heartcount	Companies	Yes	Yes	No	No

f) Mood Meter: Designed for personal emotional intelligence, the *Mood Meter* uses an evenly divided square to with different colors to get different categories of the mood [25]. Users tap in the respective color quadrant to reflect their mood. They can provide a reason, and then can select strategy to improve their mood.

g) Celpax: Celpax requires employees of a particular company to press green or red button on a specially designed device to express their feelings at the end of the day [26]. An extensive online dashboard is provided for periodic results.

h) Peakon: Peakon helps the management to drive measurable improvements with regular surveys, real-time insights, and collaborative action planning [27]. Supporting multiple languages, Peakon can engage large number of employees to collect their feedback through the web, mobile app, kiosk, and regular SMS.

i) Bluepulse: Managed under the umbrella of Explorance, *Bluepulse* is employee experience management software built from the ground up to investigate employees engament through surveys [28]. Bluepulse can easily be integrated with other related platforms for HR to enhance automation for rich insights.

j) BambooHR: BambooHR is administrated as a web application for providing complete HR solutions [29]. Employees engagement and moods are measured using surveys to set focus on what matters most for the human assets of the company.

k) Heartcount: Powered with AI algorithms, *Heartcounts* measures employees happiness by tracking their person progress, and relationship with the peers [30]. Immediate feedback is shared with the concerned authorities to improve decision making thus bringing the entire ecosystem together.

Table I presents comparison of the existing applications. It is clear that most of these applications target companies. They can collect user input, present the result to the management, and analyze the results to give some guidelines. Most of the existing solutions are provided in English language and are focused on encouraging the improvements in the business policies to increase the happiness of their employees. Some of the solutions are web-based only while others can also be worked on using mobiles phones.

In the context of measuring mood of the students in education institutes in Saudi Arabia, an application with Arabic interface is required. Furthermore, existing applications still require to wonder around several screens with heterogeneous data to understand mood of the participants, which is again time consuming. Due to the stand alone nature of the existing application, none of them provides interaction with other smart devices in environment for change to enhance positive feelings.

Our proposed Meezaj system supports both English and Arabic rather than only English giving us a wider client base in Saudi Arabia. Using IoT technology, Meezaj can interact with other smart devices in the environment for customization to maximize the happiness of the people engaged that particular area. To appeal the sense of urgency, the results are displayed at prominent places within the company through custom-designed sculptures.

IV. MEEZAJ SYSTEM

This section presents architecture and workflow of the Meezaj system developed for mood measurement and reflection. Meezaj measures the mood in a specific institutional setting by collecting data in terms of responses to mood surveys. The participants are notified on Android based smart phones to answer short or long survey periodically. Survey responses are then analyzed in real-time and the emotions of participants are reflected on a custom designed RGB LED-based sculpture. If most of the people are feeling happy, the sculpture glows yellow, but when they are sad, the sculpture will glow blue. The sculpture will glow red if most of the people are feeling angry.

A. Architecture

As depicted in Fig. 1, Meezaj system architecture is structured in three layers: *User Interaction*, *Application Logic*, and *IoT Platform*.

1) User interaction layer: User interaction layer contains the interface and presentation logic. User's visual experience is enhanced using the latest technologies for both web and the mobile apps. Developed using Bootstrap technology, the web interface provides visually appealing screens for all the required features to perform Create, Read, Update, Delete (CRUD) operations. Required reports are enriched with inforgraphics for detailed description and for improving understanding. Responsive design methodology has been followed to improve the visibility on small screens.

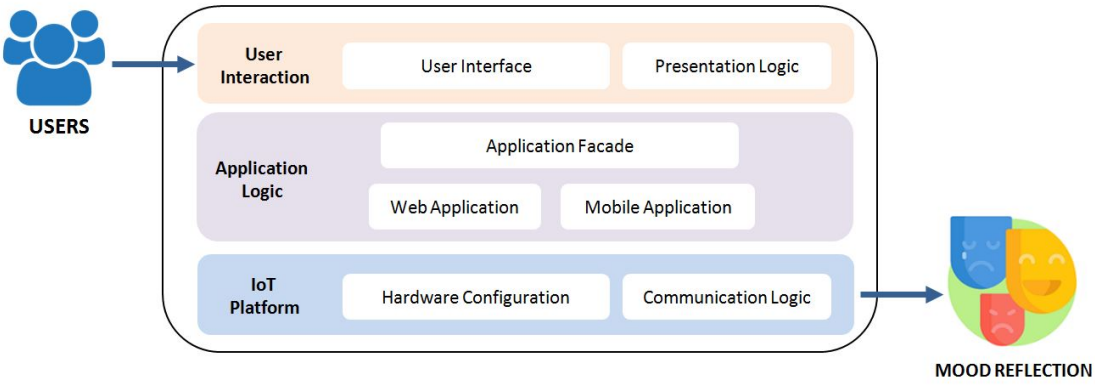


Fig. 1. Meezaj System Context Diagram

2) Application logic layer: Application logic layer consists of the following two main components:

- *Meezaj Web App*: Each registered institute is provided a unique code to be used by the students and employees of the for registration. The registered institutes can then perform all the required administration activities; like creating or updating detailed surveys by adding or removing questions, and activation or deactivation of the user accounts after login to Meezaj web app at https://meezaj.net/. Users activated with *Institute Manager* role can analyze the survey results for further insights.
- *Meezaj Mobile App*: To engage users for mood measurement an Android mobile app is developed using Android Studio and is available at the Google Play store. Upon installation, Meezaj mobile app requires user registration and respective institute code provided by the manager. Users can set the language to either English or Arabic. Meezaj app notifies users about the mood survey consisted of one question, "*How to do feel*" with three specially designed icons for happy, sad, and angry (see Fig. 2a). The user can press the icon to record in mood. After tenth response, users start getting individual mood report in the form of a pie graph (see Fig. 2b). User can also participate in long mood survey by keep pressing any particular icon for couple of seconds. Although, Meezaj as already provided sample questions for long surveys based on the Experience Sampling Method (ESM) approach but the institute managers can add/remove questions for customized surveys.

3) IoT platform layer: Sitting at the bottom, this layer defines basis of the hardware configuration used to control display of the sculpture and communication logic with other smart device in the environment. Hardware configuration consists of Arduino MKR1000 controller board and the required circuitry for RGB LED-based sculpture. MKR1000 board has been selected due to low cost and WiFi connectivity [31]. Although, Arduino boards are easy to configure for programming but MKR1000 requires some extra libraries to be installed for correct configuration. As a unique IP address is assigned to particular MKR1000 board, it is now a uniquely identified "*thing*" over the internet and can communicate with other things and computing platforms.

For communication, POST method of the HTTP protocol is used to communication with Meezaj web server by assessing the respective web application (https://meezaj.net/). As every MKR1000 WiFi board has a unique MAC address, so after a certain period of time (set through Admin panel) it connects with available WiFi and sends MAC address and recives the mood values form the *Meezaj web server*. Upon receiving the calculated mood values, the controller then sets the color values for the RGB LEDs to glow is appropriate colors to reflect the mood of the institute.

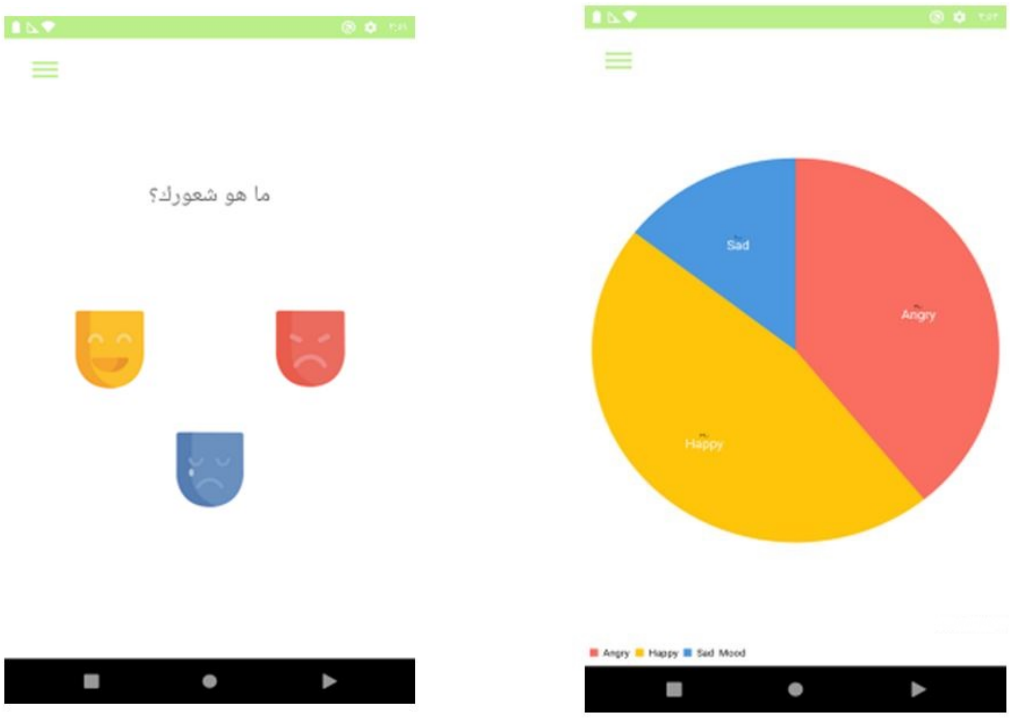
B. Workflow

This section presents the workflow of the Meezaj system. In a typical university setting Meezaj supports three types of roles:

- *Users* can be students, faculty members, or any member of the institutes work force. They are the ones going to respond to the mood surveys and their mood will be reflected. They can also check their mood history.
- *Institute Manager* is a member of middle management within the institute responsible for the providing the institute codes to the students, customizing the long mood surveys, setting the periodic mood notification timings, and analyzing the mood data.
- *System Administrator:* are members of the Meejaz team responsible for the system and its services. With privileges viewing the analyzed mood data, they can arrange discussions with the institute managers to discuss patterns and abnormal spikes in the data.

As depicted in Fig. 3, mood measurement and reflection through Meezaj system is composed of six steps described below.

1) Mood survey notification: After registration, the users get a mood survey notification after a certain period of time, set by the institute manager through admin panel on Meezaj web app. Answering a survey is one of the most important features of the Meezaj system giving that it is currently the only way



(a) Short mood survey (in Arabic)

(b) Individual mood report (in Engish)

Fig. 2. Meezaj Mobile App User Engagement

of getting user mood information. As Meezaj can support multiple educational institutes at the same time, so the users are identified by the institute code provided by the institute managers. Survey notification remains in the notification until responded by the user. It is important to note that the Meezaj mobile app need to run in the background specially when the users in the premises of their institutes. The permission is exclusively asked for while installation.

2) Survey response: Users can tap the mood survey notification to open Meezaj mobile app. The respective screen has three icons as shown in Fig. 2a, wherein, yellow icon with happy face is used for happiness, blue icon is for sadness, and the red icon is used to represent the angriness. Users can press appropriate icon to express their feelings. This short mood survey is specially designed to improve user response time and interest. A really simple and short survey (with only one question and three possible choices) is also one of the unique features of our Meezaj system for happiness measurement and is based on our own experience and reviews of the existing applications. Users get frustrated very often while answering multiple questions. This not only affects the response ratio (as explained in [32] for surveys in general) but may also produce invalid results in case of happiness measurement. For example, the user was feeling happy when the survey notification was received but gets frustrated while answering the long mood survey. So the short mood survey facilitates getting the immediate feelings accurately.

To investigate the reasons for the immediate feelings, users can also provide the details (through long mood survey) by long-pressing the appropriate icon. Meezaj mobile app then leads to the screen with multiple questions to know the reasons of the current feelings. Although, we have already provided sample questions based on theoretical foundations of the ESM approach but institute managers can customizing this question set by rephrasing, adding and removing possible answers, and by adding or deleting a complete question. Response to the long mood surveys directly indicate the reasons of the immediate feelings. Later patterns can be identified using Natural Language Processing (NLP) techniques and the related policies can be updated to maximize the happiness of the students.

Users of the Meezaj mobile app are also asked for permission to share their location information. Upon submitting response for a short or long survey, GPS location of the respondent is also exchange with the web sever. This information can further assist in identifying the happy, sad, or angry spots within the campus. As shown in step 5 of Fig. 3, other IoT devices in that particular spot can then be engaged to customize the environment for positive change. For example, the light can be adjusted in a close environment and a fountain (if present) can be turn on for the change in the environment.

3) Mood data submission: Appropriate relational schema is defined for each institute's data storage. Meezaj facilitates role-based user management module is implemented to realize data protection. Successful data modeling not only establishes administration of acquisition and storage but also ensures the reliability, and timeliness of the data for relevant users.

Algorithm 1 specifies important steps followed for survey data submission to the web server. It requires response to a survey with unique identification *Surv_Id* issued to a respondent *Resp_Id* in an institute with institute code *Inst_Code*.

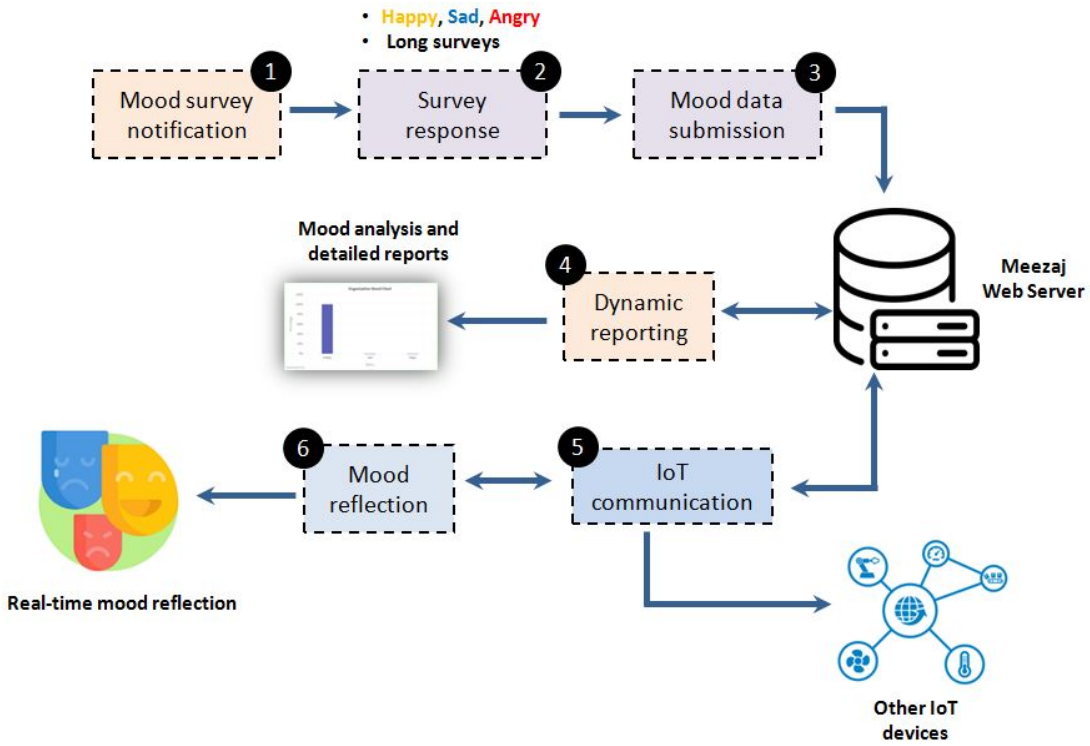


Fig. 3. Stepwise Workflow of the Meezaj System

The algorithm then ensures that all the respective relations are updated accordingly. In case of a short survey (with one questions and three possible answers), a new record is added to the *Survey* relation with mood values (happy, sad, or angry) using *Surv_Id.mood_value*, *Resp_Id*, and *Inst_Code*.

For long surveys after updating the *Survey* relation, every *question* in the question set for a particular survey is explored for the answer and *questions.answer_value* is added to the *Detailed_Survey* relation.

4) Dynamic reporting: Along with their growth, institute are also looking for improving their operational performance. Efficient reporting, with right data at the right time, for well-informed decision making is decisive for any technological application. Meezaj also provides a feature-rich module for managing dynamic reporting. Complex ideas of subjective assessment are realized with infographics to improve understandability of the important data. Individual mood reports are automatically generated and shared with the respondents after each survey response (as shown in Fig. 2b). With details provided as responses to the long the mood surveys, Meezaj also provides an opportunity to identify mood patterns in relation with the people, activities, and places. Users can then easily identify the personalities, places, and activities that make them happy and can try to engage with them more often.

At the institute level dynamic reports can be generated with a user-friendly interface. Institute's mood spikes can further assist to recognize the reasons and update the institutional policies accordingly.

5) IoT communication: Due low cost, built-in WiFi shield, and ease to use, Arduino MKR1000 board is selected for IoT hardware configurations in the Meezaj system [31]. It is based on Amtel ATSAMW25 SoC (System on Chip) which is specifically designed for IoT projects and devices. MKR1000 board also contains a Li-Po charging circuit enabling it to run on battery. Switching from external 5V to Li-Po battery is maintained without any degradation.

For programming, Arduino MKR1000 board requires WiFi101 library allowing WiFi shield to connect the available WiFi connection and further communication. This library supports crypto-authentication and allows the board to be configured as a server–accepting multiple client connections or as a client–sending requests to the server [31]. The library also

Algorithm 1: Mood data submission to Meezaj web server

Require: Mood survey response with institute code *Inst_Code*, respondent Id *Resp_Id* and survey id *Surv_Id*
Ensure: Update respective database relations for *Surv_Id*
1: **if** *Surv_Id* is a short survey **then**
2: update *Survey* relation with *Surv_Id.mood_value*, *Inst_Code*, and *Resp_Id*
3: update *Surv_Id.time_stamp*
4: **end if**
5: **if** *Surv_Id* is a long survey **then**
6: repeat step 2-3
7: **for all** *question* $\in$ *Surv_Id.Questions* **do**
8: update *Detailed_Survey* relation with *question.answer_value*
9: **end for**
10: **end if**

supports both WEP and WPA2 Personal encryption methods.

Another reason for selecting Arduino MKR1000 board is its ability to support different M2M communication protocols. In Meezaj, we have used it for HTTP communication but it can easily be configured as *publisher* and *subscriber* with an MQTT broker. This allows it to communicate with other IoT devices in the environment for customization and then maximizing the happiness of the students.

6) Mood reflection: Mood reflection is an important final step of the Meezaj system workflow. The data collected in terms of responses to the (short and long) mood surveys is analyzed in real-time and emotions of the participants are reflected on a custom designed MKR1000 and RGB LED-based sculpture. If most of the people are feeling happy, it will glows yellow, but when they are fearful, the sculpture will glow blue. The sculpture will glow red if most of the people are feeling angry.

Algorithm 2 presents the mood reflection logic based on the *MAC_address* of the MKR1000 connected to the custom RGB LED-base sculpture and the *update_period* parameter set by the institute manger through Admin panel on Meezaj web app. Fist of all, in the *setup()* method of the Arduino sketch, programming logic is specified to connect respective MKR1000 with *MAC_address* to the available WiFi network. An instance *client* of the *WiFiClient* class, present in the included WiFi101 library, is created. Because of the IP address assigned to MKR1000 WiFi shield with respective *MAC_address*, this instance is now uniquely identified as a "*thing*" on the internet and represents the sculpture to reflect the mood of the particular institute or a department of an institute.

After establishing connection with an available WiFi router, the MKR1000 board then connects with the Meezaj web app (*http://www.meezaj.net*) through HTTP connection and sends its MAC address using POST method request to activate the *sendmood* PHP script. This PHP script fetches mood values from the survey responses stored in the respective relations and sends them back on the serial port of the MKR1000 board to be stored as a String. This String is then further processed to get the required *mood values*. Appropriate color values are then set for RGB LEDs connected with the MKR1000 board to glow appropriately. The *client* then waits for the next *update_period* and queries Meezaj web server again for mood values.

Algorithm 2: Real-time mood reflection with MKR1000

Require: Active WiFi connection with MAC address *MAC_address* of the MKR1000 WiFi shield and the update time period *update_period*

Ensure: Appropriately glow LED-based sculpture using MKR1000 with *MAC_address* to reflect mood of the institute

```
connect MKR1000 with an active WiFi connection
create an instance client of the WiFiClient class with IP address assigned to MKR1000 with MAC_address
while client.available do
    generate HTTP connection string for www.meezaj.net
    generate POST request method by sending MAC_address to PHP script file sendmood.php
    get mood values from the HTTP response
    set values for RGB LEDs connected to the particular MKR1000 according to the mood values
    wait for update_period
end while
```

V. Conclusion and Future Work

It is a proven fact that students' level of happiness has a significant effect on their performance. In this paper, we have presented Meezaj–an interactive system for mood measurement and reflection of the relevant stakholders based on the Internet of Things (IoT). Meezaj supports both short and long mood survey, to collect data about the feelings of the participants, to be notified on Meezaj mobile app. Responses to these surveys are submitted to the Meezaj web server. Institute level mood is then calculated and reflected through Meezaj web app and a custom designed sculpture. Arduino MKR1000 and RGB LED-based sculpture is controlled to glow in appropriate colors to reflect the mood of the institute. One of the important future work is to further enhance mood data acquisition with on sight cameras and use IBM Watson NLP techniques for sentiment analysis on mood survey response data.

Acknowledgment

The author would like to thank Saoud Abdulmajeed and Abdulrahman Ismail from Saudi Electronic University for their work on developing an initial prototype of the Meezaj system as part of their capstone project.

References

[1] N. R. Council, *Subjective Well-Being: Measuring Happiness, Suffering, and Other Dimensions of Experience*, A. A. Stone and C. Mackie, Eds. Washington, DC: The National Academies Press, 2013. [Online]. Available: https://www.nap.edu/catalog/18548/subjective-well-being-measuring-happiness-suffering-and-other-dimensions-of

[2] D. A. Morand, "Family size and intelligence revisited: The role of emotional intelligence," *Psychological Reports*, vol. 84, no. 2, pp. 643–649, 1999, pMID: 10335078. [Online]. Available: https://doi.org/10.2466/pr0.1999.84.2.643

[3] V. Applasamy, R. A. Gamboa, M. Al-Atabi, and S. Namasivayam, "Measuring happiness in academic environment: A case study of the school of engineering at taylor's university (malaysia)," *Procedia - Social and Behavioral Sciences*, vol. 123, pp. 106 – 112, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877042814014414

[4] S. Lyubomirsky, L. King, and E. D. Diener, "The benefits of frequent positive affect: Does happiness lead to success?" in *Psychological bulletin*, 2005.

[5] A. J. Oswald, E. Proto, and D. Sgroi, "Happiness and productivity," *Journal of Labor Economics*, vol. 33, no. 4, pp. 789–822, 2015. [Online]. Available: https://doi.org/10.1086/681096

[6] C. O. Walker, T. D. Winn, and R. M. Lutjens, "Examining relationships between academic and social achievement goals and routes to happiness," *Education Research International*, vol. 2012, p. 643438, Sep 2012. [Online]. Available: https://doi.org/10.1155/2012/643438

[7] D. G. Myers, "Happiness articles," http://www.davidmyers.org/Brix?pageID=47, (accessed July 21, 2020).

[8] L. Coetzee and J. Eksteen, "The internet of things-promise for the future? an introduction," in *2011 IST-Africa Conference Proceedings*. IEEE, 2011, pp. 1–9.

[9] R. H. Weber and R. Weber, *Internet of things*. Springer, 2010, vol. 12.

[10] S. Vongsingthong and S. Smanchat, "Internet of things: a review of applications and technologies," *Suranaree Journal of Science and Technology*, vol. 21, no. 4, pp. 359–374, 2014.

[11] R. Larson and M. Csikszentmihalyi, *The Experience Sampling Method*. Dordrecht: Springer Netherlands, 2014, pp. 21–34.

[12] K. Ludwigs, L. Henning, and L. R. Arends, "Using the day reconstruction method - same results when used at the end of the day or on the next day?" *International Journal of Community Well-Being*, vol. 2, no. 1, pp. 61–73, May 2019. [Online]. Available: https://doi.org/10.1007/s42413-019-00017-x

[13] Kingdom of Saudi Arabia, "Vision 2030," https://vision2030.gov.sa/sites/default/files/report/Vision/20Realization /20Programs/20Overview.pdf, (accessed July 21, 2020).

[14] M. D. Fernandez Rodriguez and I. B. Huertas, "Suicide Prevention in College Students: A Collaborative Approach," *Interam J Psychol*, vol. 47, no. 1, pp. 53–60, Jan 2013.

[15] A. G. Abdel Rahman, B. N. Al Hashim, N. K. Al Hiji, and Z. Al-Abbad, "Stress among medical saudi students at college of medicine, king faisal university," *Journal of preventive medicine and hygiene*, vol. 54, no. 4, pp. 195–199, Dec 2013, 24779279[pmid]. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/24779279

[16] T. A. Bahhawi, O. B. Albasheer, A. M. Makeen, A. M. Arishi, O. M. Hakami, S. M. Maashi, H. K. Al-Khairat, O. M. Alganmy, Y. A. Sahal, A. A. Sharif, and M. S. Mahfouz, "Depression, anxiety, and stress and their association with khat use: a cross-sectional study among jazan university students, saudi arabia," *Neuropsychiatric disease and treatment*, vol. 14, pp. 2755–2761, Oct 2018, 30425493[pmid]. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30425493

[17] Z. J. Gazzaz, M. Baig, B. S. M. Al Alhendi, M. M. O. Al Suliman, A. S. Al Alhendi, M. S. H. Al-Grad, and M. A. A. Qurayshah, "Perceived stress, reasons for and sources of stress among medical students at rabigh medical college, king abdulaziz university, jeddah, saudi arabia," *BMC Medical Education*, vol. 18, no. 1, p. 29, Feb 2018. [Online]. Available: https://doi.org/10.1186/s12909-018-1133-2

[18] M. Soliman, "Perception of stress and coping strategies by medical students at king saud university, riyadh, saudi arabia," *Journal of Taibah University Medical Sciences*, vol. 9, no. 1, pp. 30 – 35, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1658361213000814

[19] K. Ludwigs and S. Erdtmann, "The happiness analyzer – developing a new technique for measuring subjective well-being," *International Journal of Community Well-Being*, vol. 1, no. 2, pp. 101–114, Feb 2019. [Online]. Available: https://doi.org/10.1007/s42413-018-0011-3

[20] Emooter, "*Emooter*," https://www.emooter.com/, (accessed July 21, 2020).

[21] C. Amp, "*Culture Amp*," https://www.cultureamp.com/, (accessed July 21, 2020).

[22] RoundPegg, "*RoundPegg*," https://www.achievers.com/roundpegg/, (accessed July 21, 2020).

[23] Tinypulse, "*Tinypulse*," https://www.Tinypulse.com/, (accessed July 21, 2020).

[24] MercuryApp, "*MercuryApp*," https://www.mercuryapp.com/teams, (accessed July 21, 2020).

[25] M. Meter, "*Mood Meter*," https://moodmeterapp.com/, (accessed July 21, 2020).

[26] Celpax, "*Celpax*," http://www.celpax.com/, (accessed July 21, 2020).

[27] Peakon, "*Peakon*," https://peakon.com/, (accessed July 21, 2020).

[28] Bluepulse, "*Bluepulse*," https://explorance.com/products/bluepulse/, (accessed July 21, 2020).

[29] BambooHR, "*BambooHR*," https://www.bamboohr.com/, (accessed July 21, 2020).

[30] Heartcount, "*Heartcount*," https://www.heartcount.com/, (accessed July 21, 2020).

[31] ARDUINO, "*ARDUINO MKR1000 WIFI* ," https://www.arduino.cc/, (accessed July 21, 2020).

[32] B. Burchell and C. Marsh, "The effect of questionnaire length on survey response," *Quality and Quantity*, vol. 26, no. 3, pp. 233–244, Aug 1992. [Online]. Available: https://doi.org/10.1007/BF00172427

A Repeated Median Filtering Method for Denoising Mammogram Images

Hussain AlSalman
Department of Computer Science
College of Computer and Information Sciences
King Saud University, Riyadh, Saudi Arabia

***Abstract*—In the medical field, mammogram analysis is one of the most important breast cancer detection procedures and early diagnosis. During the image acquisition process of mammograms, the acquired images may be contained some noises due to the change of illumination and sensor error. Hence, it is necessary to remove these noises without affecting the edges and fine details, achieving an effective diagnosis of beast images. In this work, a repeated median filtering method is proposed for denoising digital mammogram images. A number of experiments are conducted on a dataset of different mammogram images to evaluate the proposed method using a set of image quality metrics. Experimental results are reported by computing the image quality metrics between the original clean images and denoised images that are corrupted by different levels of simulated speckle noise as well as salt and paper noise. Evaluation quality metrics showed that the repeated median filter method achieves a higher result than the related traditional median filter method.**

***Keywords*—Mammogram images; image denoising; median filter; repeated median filtering; speckle noise; salt and paper noise**

I. Introduction

Nowadays, image processing methods have been applied for diagnosis in several medical applications, such as liver image analysis [1, 2], brain tumor classification [3, 4], breast image enhancement, and cancer diagnosis [5-7], and so on. Image denoising process is used to eliminate noises from noisy images and improve their quality. However, it faces difficulty to distinguish between noises and other important images' components such as edges and textures due to they have approximately the same high frequencies, which might lead to lose some details of the images [8]. Therefore, image denoising without losing significant information from a noisy image is still a vital problem in the image processing field [8]. In recent years, great achievement has been accomplished in the field of image denoising [9-12].

In medical imaging systems and applications, image denoising plays an important role as a pre-processing step to enhance the quality of digital images and improving the process of medical diagnosis [13]. Even though the medical image denoising process has been studied in many research types for a long time, it is still a challenging issue and an open task. One of the key reasons for this is that the medical image denoising is an inverse problem from a mathematical perspective, and its solution is not unique and not flexible.

The rest of the paper is organized as follows: Section II presents a literature review about the methods used for medical and breast image denoising. Section III describes the applied research methods. Section IV presents the experimental results and discussion. Finally, Section V concludes and discusses the research work.

II. Literature Review

Currently, there are many approaches to image denoising for medical imaging systems. Some of the common approaches are median filtering, Wiener filtering, morphological filtering, wavelet-based filtering, and curvelet transform, among other significant approaches. The median filter [14, 15] is a statistical approach for noise reduction in images with blurred edges. Wiener filter is another statistical approach that calculates unknown signals using a related known signal as input [16]. Morphological filtering is a local non-linear transformation of geometric features; its fundamental operations are closing, opening, erosion, and dilation. It has been applied in different areas, especially image denoising [17]. Wavelet-based filtering has also been used for the d-noising of images of all kinds, specifically for medical image systems. It is a mathematical calculation that is able to perceive local features of the image.

Additionally, it is used to decompose 2D signals into diverse resolution levels [18]. For a functional performance of image denoising, an adaptive procedure for image discontinuities are applied. Accordingly, a multi-resolution approach is adapted. Here, the curvelet transform can be used to improve image resolution [19]. Breast cancer is currently the most common type of cancer with the highest mortality cause among women in the world[1]. The number of deaths from breast cancer has doubled in 22 years, affecting both industrialized and less developed countries. Its main known risk factors are associated with prolonged exposure to estrogens, are indicators of lifestyle and reproductive patterns, and therefore are difficult to modify and reducing mortality, then, it requires improving early detection and treatment strategies. Among screening procedures, which also include self-examination and clinical examination, mammography is the only technique that can offer sufficiently timely detection. In which a low energy X-rays is used to screen breast in order to assist breast detection. To ensure an accurate diagnosis, breast x-rays images should be of high quality. In this

[1] https://www.who.int/news-room/fact-sheets/detail/cancer

direction, multiple approaches are used for de-noising the mammogram is one approach to improve quality. In [20, 21], convolutional neural networks (CNNs) are applied to minimize the noise in mammograms. Recently, Total Variation (TV) and Non-Local Mean (NLM) algorithms are developed to mitigate some shortages of repeatable noise elimination in medical images [22].

In summary, the previous solutions in-depth related studies are still not flexible and need to be improved in terms of developing a method to remove the noise from a noisy image at different levels for getting a high-quality image depending on the selected level. Thus, this paper proposes a repeated median filtering (RMF) method that applies a median filter at a different number of iterations with different filter sizes making it more flexible for user choice.

III. Research Methods

A. Median Filter (MF)

In image processing, before further processing, such as edge detection, it is usually necessary to first perform a certain degree of noise reduction. The filtering process using MF is a common step in image processing. It is especially useful for speckle noise and salt-and-pepper noise. Preserving the edges makes it useful in situations where edge blur is not desired.

Median filter (MF) is a non-linear digital filter technology that is often used to remove noise from images or other signals. The design idea is to check the samples in the input signal and determine whether it represents the signal. Use an observation window composed of an odd number of samples to achieve this function. The values in the observation window are sorted, and the median value in the middle of the observation window is used as the output. Then, the oldest value is discarded, new samples are obtained, and the above calculation process is repeated. The main idea of the MF is to traverse the signal entry through the entry and replace each entry with the median of the neighbor entry. The neighbor's pattern is called a "window," and it slides through the entrance to cover the entire signal. For one-dimensional signals, the most obvious windows are only the front and back items, while 2D (or higher-dimensional) signals (such as images) may have more complex window modes (such as "box" or "cross" modes). Note that if there is an odd number of entries in the window, the median is easy to define: after all entries in the window are sorted numerically, this is only the middle value. For even entries, there is more than one possible median.

Median filtering is a smoothing technique similar to linear Gaussian filtering. All smoothing techniques can effectively remove noise in smooth or smooth areas of image signal but have an adverse effect on the edges. In general, it is essential to maintain edges while reducing noise in the signal. For example, edges are critical to the visual appearance of an image. For small to medium levels of Gaussian noise, the MF is significantly better at removing noise than Gaussian blur while preserving edges for given fixed window size. However, for high noise, its performance is not better than the Gaussian blur, and it is particularly effective for speckle noise and salt and pepper noise (impulse noise). Therefore, MF is widely used in digital image processing [3].

The naive implementation described above sorts each entry in the window to find the intermediate value; however, since only the intermediate value in the list is needed, the selection algorithm can be more efficient. In addition, some types of signals (usually the case of images) are represented using integers: in these cases, the histogram is simple because updating the histogram from window to window and finding the median of the histogram is not incredibly tedious that makes it to be much more efficient. Fig. 1 illustrates an example of a median filter calculation.

B. Proposed Repeated Median Filtering (RMF) Method

The repeated median filtering (RMF) method is a non-linear median-based processing approach that applies an MF on an image *N* times to remove noises at different levels for getting a high-quality image depending on the suitable selected level. The main idea of the RMF method is very simple but more effective. In this method, the number of iterations and filter size should be defined that makes it flexible for user choice. To illustrate how the method works, Algorithm 1 describes the steps of the RMF process. In the application of the RMF method, the user needs to initialize the method's parameters, such as the number of iterations and the filter size. The large size of MF used in the approach is not suitable due to a large set of pixels that makes the MF values deviate from the values of pixels. The number of iterations makes the method is more flexible to perform the filtering process at different times. The filter size forms a 2D window that is a central symmetric shape that replaces the pixel at the center by the median value of pixels values inside that window.

Algorithm 1. The main steps of the RMF method.

```
1.  Input: inputImage, number of iterations (N), filter size (s)
2.  Output: outputImage
3.  Begin
4.  [imageWidth, imageHeight]=imageSize(inputImage);
5.  For r=1; r<=N; r++
6.     For i=0; i<imageWidth; i++
7.        For j=0; j<imageHeight; j++
8.           Initialize array int[s*s] temp
9.           int index=0
10.             for k=i-1; k<i+1; k++
11.                for m=j-1; m<m+1; m++
12.                   temp[index]=inputImage [k, m]
13.                   index++;
14.                sort(temp)
15.             End for
16.          End for
17.       outputImage [i, j]=temp[s+1];
18.    End for
19.  End for
20. End for
21. End
```

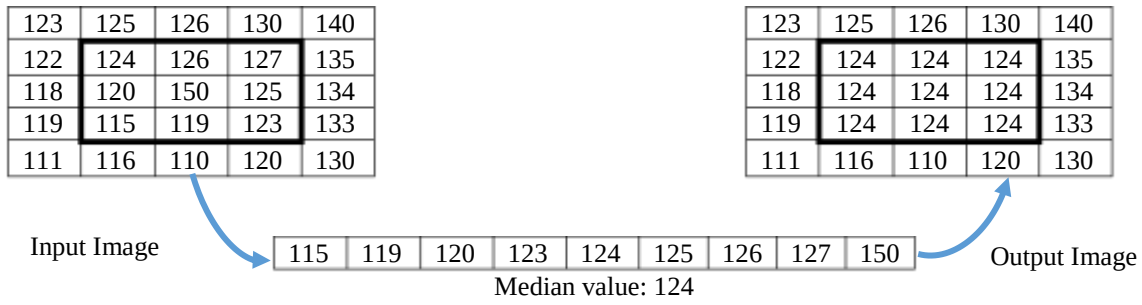


Fig. 1. An Example of a Filtering Process using MF.

IV. Experiments and Discussion

In this section, a set of experiments are conducted on a number of breast mammogram images taken from the database of mini- Mammogram Image Analysis Society (MIAS) to evaluate the proposed method. In addition, the results of the proposed method will be compared with the results of MF on the same images. The proposed method is implemented by using the MATLAB R2016b programming tool.

The implementation was performed on a laptop that has an Intel CPU I7 2.2 GHz with 16 GB of RAM and a Windows 10 operating system. With these experiments' configurations, the evaluation results on the test images are assessed using the peak signal-to-noise ratio (PSNR) and mean squared error (MSE) performance metrics. The experimental results and comparisons will be introduced in the following subsections.

A. Dataset Mammogram Images

The dataset mammogram images are five samples, selected from the Mini-mammogram Image Analysis Society (MIAS) database and shown in Fig. 2. All test images are in PGM format. In the experiments, these images are converted into PNG format and resized to 256×256 pixels.

B. Image Quality Evaluation Measures

The evaluation measures used to assess the proposed image denoising method are quantitative image quality measures include Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR). These measures are computed based on original and denoised images. MSE is a cumulative value of squared errors between an original image (O) and a denoised image (D) with 2D matrices with m rows and n columns. MSE has a small value if the method performs well and can be computed as [23]:

$$MSE = \frac{1}{M*N}\sum_{M,N}[O(m,n) - D(m,n)]^2 \quad (1)$$

The second measure is the PSNR that can give a good indication of the capability of the method to remove the noises. The small value of PSNR for the denoised image means it has a poor quality [23]. PSNR can be calculated as in the following equation.

$$PSNR = 10log_{10}\left(\frac{R^2}{MSE}\right) \quad (2)$$

The variable R in the previous equation is the maximum fluctuation of image's pixels if the image has a data type of double floating-point, then R is one, and if the image has a data type of 8-bit unsigned integer, then R is 255.

C. Results and Discussion

To validate the proposed denoising method, all evaluation images are degraded artificially using speckle noise with a different ratio of 0.5% and 1%. Besides, they are also degraded using salt and pepper noise with a different ratio of 10% and 20%, respectively. The experimental results of the proposed filtering method on noised images are assessed based on the adopted image quality evaluation measures. During the experiments, the number of iterations and filter size are initialized. Initializing the number of iterations is critical to determine the quality of denoised image. Therefore, a set of experiments are conducted to select the best value of the number of iterations. Tables I and II list the quantitative results of MSE and PSNR measures under a different number of iterations (N).

From Tables I and II, the best quantitative results of MSE and PSNR are when the value of N is 2. Thus, this value is selected for the method to remove the images' noise. Tables III and IV exhibits the experimental results of MSE and PSNR for the proposed method at different noise levels of test images degraded with speckle noise.

TABLE I. Quantitative Results of MSE and PSNR Measures of RMF Method under a different Number of Iterations (N) for Speckle Noise with a 1% Noise Ratio on mdb015.pgm Image

N	MSE	PSNR
1	0.00036	82.5685
2	0.00030	83.3691
3	0.00032	83.2037
4	0.00034	82.9138
5	0.00036	82.5974

TABLE II. Quantitative Results of MSE and PSNR Measures of RMF Method under a different Number of Iterations (N) for Salt and Pepper Noise with a 20% Noise Ratio on mdb015.pgm Image

N	MSE	PSNR
1	0.00110	77.6215
2	0.00035	82.7175
3	0.00037	82.5298
4	0.00040	82.1220
5	0.00044	81.7712

As shown in Tables III and IV, the proposed method achieves high values of PSNR and low values of MSE. These results validate the effectiveness of the method to remove the speckle noises from the test images. Fig. 3 visualizes an example of noised and denoised images.

Fig. 2 shows how the noisy images are improved by using the RMF method that removes the speckle noise that has a ratio of 1%. Tables V and VI demonstrate the quantitative results of MSE and PSNR for the proposed method on the test images that are degraded with salt and pepper noise at different noise levels of 10% and 20%. Fig. 4 visualizes an example of noised and denoised images.

To compare the proposed RMF method with the MF method, Tables V and VI show the results of PSNR of denoised images using RMF and MF methods. Table V displays the results of PSNR on the test images that are degraded with speckle-noise at 1% of the noise level. Similarly, Table VI exhibits the results of PSNR on the test images that are degraded with salt and pepper noise at 20% of the noise level.

TABLE III. QUANTITATIVE RESULTS OF MSE AND PSNR MEASURES OF RMF METHOD FOR SPECKLE NOISE WITH A 0.5% NOISE RATIO ON TEST IMAGES

Image No.	MSE	PSNR
mdb005.pgm	0.00030	83.4473
mdb010.pgm	0.00014	86.6788
mdb012.pgm	0.00038	82.3583
mdb013.pgm	0.00025	84.2229
mdb015.pgm	0.00021	84.8947

TABLE IV. QUANTITATIVE RESULTS OF MSE AND PSNR MEASURES OF RMF METHOD FOR SPECKLE NOISE WITH A 1% NOISE RATIO ON TEST IMAGES

Image No.	MSE	PSNR
mdb005.pgm	0.00042	81.9013
mdb010.pgm	0.00023	84.4980
mdb012.pgm	0.00055	80.7610
mdb013.pgm	0.00036	82.5686
mdb015.pgm	0.00030	83.3691

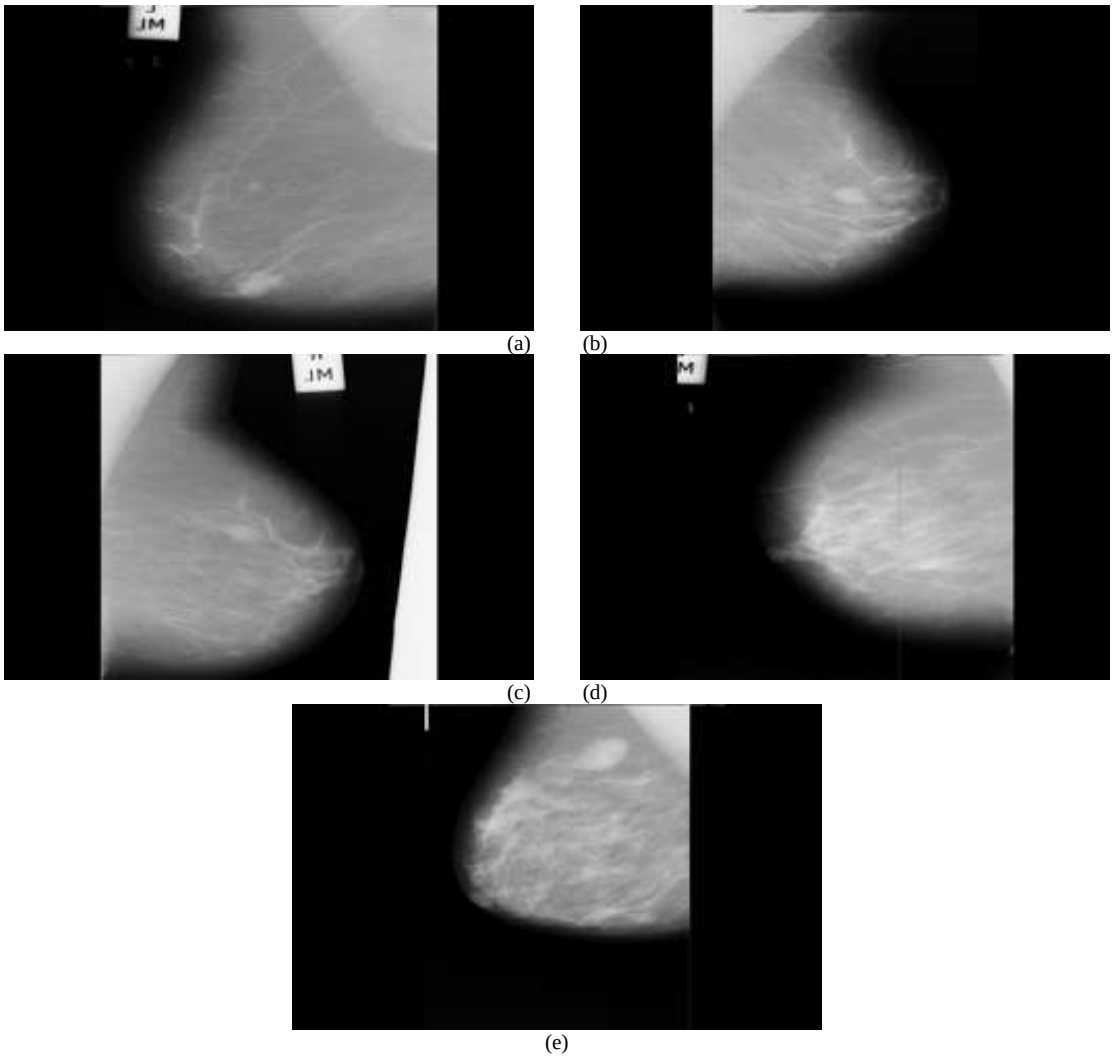
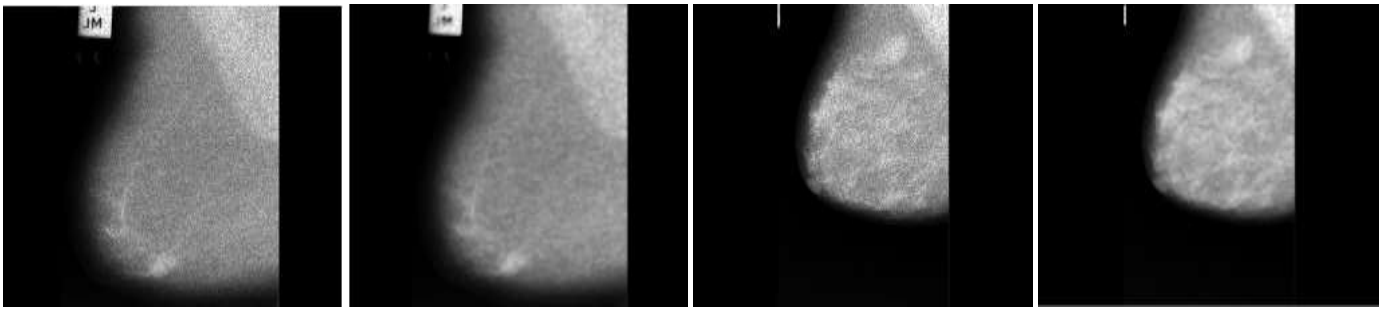


Fig. 2. Test Evaluation Images Taken from Mini- Mammogram Image Analysis Society (MIAS) Database under Images Names: (a) mdb005.pgm, (b) mdb010.pgm, (c) mdb012.pgm, (d) mdb013.pgm, and (e) mdb015.pgm.



(a) (b) (c) (d)

Fig. 3. A Visualization Example of Noised and Denoised using the RMF Method for Speckle Noise Removal with 1% Noise ratio: (a) and (c) Noisy Images, (b) and (d) Denoised Images.

TABLE V. COMPARISON RESULTS OF PSNR MEASURE FOR RMF AND MF METHODS FOR SPECKLE NOISE WITH A 1% NOISE RATIO ON TEST IMAGES

Image No.	RMF PSNR	MF PSNR
mdb005.pgm	81.9013	80.9391
mdb010.pgm	84.4980	83.0336
mdb012.pgm	80.7610	80.0645
mdb013.pgm	82.5686	81.4102
mdb015.pgm	83.3691	82.5685

TABLE VI. COMPARISON RESULTS OF PSNR MEASURE FOR RMF AND MF METHODS FOR SALT AND PEPPER NOISE WITH 20% RATIO ON TEST IMAGES

Image No.	RMF PSNR	MF PSNR
mdb005.pgm	81.8447	77.6831
mdb010.pgm	84.6203	77.8431
mdb012.pgm	77.0829	74.7745
mdb013.pgm	83.0218	77.8119
mdb015.pgm	82.7175	77.6215

Besides, Fig. 5 and 6 visualize the results of MSE for the RMF and MF methods applied on the test images that are corrupted by speckle noise with 1% noise ratio and salt and pepper noise with 20% noise ratio.

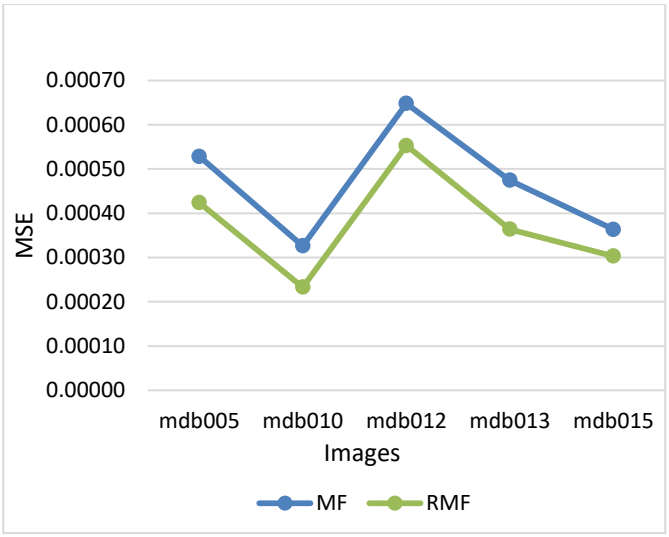


Fig. 4. A Visualization of MSE Results for the RMF and MF Methods on the Test Images Corrupted by Speckle Noise with a 1% Noise Ratio.

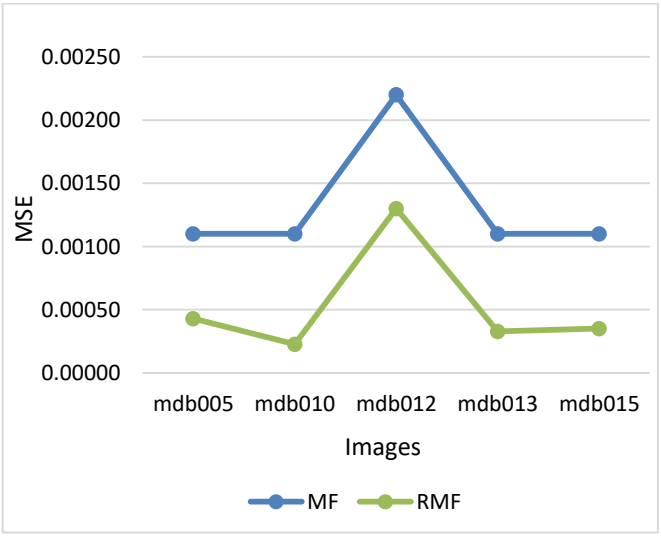


Fig. 5. A Visualization of MSE Results for the RMF and MF Methods on the Test Images Corrupted by Salt and Pepper Noise with a 20% Noise Ratio.

From comparison results in Tables V and VI, as well as Fig. 4 and 5, it is clear that the proposed RMF method outperforms the MF method in terms of PSNR for all test images. For the speckle noise, an improvement of the denoised images is greater than 1%. Furthermore, for the salt and pepper noise, the proposed method achieves a significant high PSNR result.

V. CONCLUSIONS AND FUTURE WORK

In the medical field, mammogram analysis is one of the most important procedures for breast cancer detection and early diagnosis. During the image acquisition process of mammograms, these images may be contained some noises due to the change of illumination and sensor error. Hence, it is necessary to remove these noises without affecting the edges and fine details, achieving an effective diagnosis of beast images. Therefore, in this paper, a repeated median filtering (RMF) method is proposed for denoising mammogram images. This method is able to enhance the digital mammogram images in a special domain to preserve the useful information of images. To evaluate the proposed method, a number of experiments are conducted on a dataset of different mammogram images to evaluate the proposed method using a set of image quality metrics. Experimental results are reported by computing the image quality metrics

between the original clean images and denoised images that are corrupted by different levels of simulated speckle noise as well as salt and paper noise. Evaluation quality metrics showed that the repeated median filter method achieves a higher result than the related traditional median filter method.

ACKNOWLEDGMENTS

The author is thankful to the Deanship of Scientific Research, College of Computer and Information Sciences (CCIS) at King Saud University for funding this research.

REFERENCES

[1] E.-L. Chen, P.-C. Chung, C.-L. Chen, H.-M. Tsai, and C.-I. Chang, "An automatic diagnostic system for CT liver image classification," IEEE transactions on biomedical engineering, vol. 45, no. 6, pp. 783-794, 1998.

[2] W. Cao, X. An, L. Cong, C. Lyu, Q. Zhou, and R. Guo, "Application of Deep Learning in Quantitative Analysis of 2 - Dimensional Ultrasound Imaging of Nonalcoholic Fatty Liver Disease," Journal of Ultrasound in Medicine, vol. 39, no. 1, pp. 51-59, 2020.

[3] A. Gumaei, M. M. Hassan, M. R. Hassan, A. Alelaiwi, and G. Fortino, "A hybrid feature extraction method with regularized extreme learning machine for brain tumor classification," IEEE Access, vol. 7, pp. 36266-36273, 2019.

[4] J. Amin, M. Sharif, N. Gul, M. Yasmin, and S. A. Shad, "Brain tumor classification based on DWT fusion of MRI sequences using convolutional neural network," Pattern Recognition Letters, vol. 129, pp. 115-122, 2020.

[5] H. AlSalman and N. Almutairi, "IDSS: An Intelligent Decision Support System for Breast Cancer Diagnosis," in 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), 2019, pp. 1-6: IEEE.

[6] A. Gumaei, A. El-Zaart, M. Hussien, and M. Berbar, "Breast segmentation using k-means algorithm with a mixture of gamma distributions," in 2012 Symposium on Broadband Networks and Fast Internet (RELABIRA), 2012, pp. 97-102: IEEE.

[7] H. AlSalman, "Enhancing Digital Mammogram Images using Bandpass Filters in Frequency Domain," International Journal of Computer Science and Network Security, vol. 19, no. 11, pp. 107-113.

[8] L. Fan, F. Zhang, H. Fan, and C. Zhang, "Brief review of image denoising techniques," Visual Computing for Industry, Biomedicine, Art, vol. 2, no. 1, p. 7, 2019.

[9] M. C. Motwani, M. C. Gadiya, R. C. Motwani, and F. C. Harris, "Survey of image denoising techniques," in Proceedings of GSPX, 2004, pp. 27-30.

[10] P. Jain and V. Tyagi, "A survey of edge-preserving image denoising methods," Information Systems Frontiers, vol. 18, no. 1, pp. 159-170, 2016.

[11] M. Diwakar and M. Kumar, "A review on CT image noise and its denoising," Biomedical Signal Processing Control, vol. 42, pp. 73-88, 2018.

[12] P. Milanfar, "A tour of modern image filtering: New insights and methods, both practical and theoretical," IEEE signal processing magazine, vol. 30, no. 1, pp. 106-128, 2012.

[13] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, Digital image processing using MATLAB (Pearson Education India). 2004.

[14] H. Ibrahim, N. S. P. Kong, and T. F. Ng, "Simple adaptive median filter for the removal of impulse noise from highly corrupted images," IEEE Transactions on Consumer Electronics, vol. 54, no. 4, pp. 1920-1927, 2008.

[15] S. Agarwal and S. Chand, "Median Filtering Detection Using Markov Process in Digital Images," in Advances in Biomedical Engineering and Technology: Springer, pp. 35-42.

[16] S. Ghael, A. M. Sayeed, and R. G. Baraniuk, "Improved wavelet denoising via empirical Wiener filtering," in SPIE Technical Conference on Wavelet Applications in Signal Processing, 1997.

[17] J. Mehena, "Medical Image Edge Detection Using Modified Morphological Edge Detection Approach," 2019.

[18] Y. Yang, D. S. Park, S. Huang, and N. Rao, "Medical image fusion via an effective wavelet-based approach," EURASIP Journal on Advances in Signal Processing, vol. 2010, pp. 1-13, 2010.

[19] J.-L. Starck, E. J. Candès, and D. L. Donoho, "The curvelet transform for image denoising," IEEE Transactions on image processing, vol. 11, no. 6, pp. 670-684, 2002.

[20] D. Abdelhafiz, C. Yang, R. Ammar, and S. Nabavi, "Deep convolutional neural networks for mammography: advances, challenges and applications," BMC bioinformatics, vol. 20, no. 11, p. 281, 2019.

[21] L. Li, Y. Chen, Z. Shen, X. Zhang, J. Sang, Y. Ding, X. Yang, J. Li, M. Chen, and C. Jin, "Convolutional neural network for the diagnosis of early gastric cancer based on magnifying narrow band imaging," Gastric Cancer, pp. 1-7, 2019.

[22] S. Lee, S. J. Park, J. M. Jeon, M.-H. Lee, D. Y. Ryu, E. Lee, S.-H. Kang, and Y. Lee, "Noise removal in medical mammography images using fast non-local means denoising algorithm for early breast cancer detection: a phantom study," Optik, vol. 180, pp. 569-575, 2019.

[23] R. Sammouda, A. M. S. Al-Salman, A. Gumaei, and N. Tagoug, "An efficient image denoising method for wireless multimedia sensor networks based on DT-CWT," International Journal of Distributed Sensor Networks, vol. 11, no. 11, p. 632568, 2015.

Developing an Information Management Strategy for e-Government in Saudi Arabia

Fatmah Almehmadi
College of Computer and Information Systems
Umm Al-Qura University (UQU), Makkah, Saudi Arabia

***Abstract*—Given the current Corona virus pandemic, the role of e-government in both developed and developing countries is becoming more important than ever. This study aims to assess the development of e-government in Saudi Arabia and to compare it with that of two world e-government leaders: USA and the Republic of Korea, during the period of 2003-2020. Data analysis consists of: 1) a comparative, cross-country, longitudinal analysis of the e-government development index (EGDI) relating to Saudi Arabia, the USA and the Republic of Korea; 2) a trend analysis of the online services, telecommunication infrastructure, and human capital indicators; and 3) a gap analysis to pinpoint the gap between Saudi Arabia and the USA, and the gaps between Saudi Arabia and the Republic of Korea. The results reveal a continuous rise in the rankings of Saudi Arabia's EGDI over the years. However, findings also indicate some areas that require more improvement. An information management strategy for the support of e-government in Saudi Arabia has been developed, describing the current e-government situation and setting high, medium, and low-level priorities that the country needs to consider in order to better its compliance with international e-government practices.**

Keywords*—*e-Government; Information technology; Information management; Trend analysis; Saudi Arabia; the USA; the Republic of Korea

I. Introduction

Digital transformation and, in particular, the role of e-government is becoming more prominent worldwide, given the current Corona virus (COVID-19) pandemic, which requires people to adapt their previous daily working styles so as to work primarily online. Previous research suggests that e-government incorporates many elements such as e-administration, e-services, and e-society. However, at its core, e-government focuses on delivering governmental services to people via the adoption and use of information and communication technologies (ICT). While e-government; as a concept, emerged in the United States in 1995, as a new administrative project for the state of Florida's central mail, its acknowledged, widespread use began in 2001 [1]. Since then, there has been a growing interest in exploring the topic. As a term, e-government has been defined a number of ways in previous literature; however, it is often referred to as the use of information and communication technologies (ICTs) to enable the provision and delivery of government services for citizens, businesses, and organizations. For example, in 2001, the World Bank defined e-government as "the government owned or operated systems of information and communication technologies that transform relations with citizens, the private sector and/or other government agencies" [2] (p. 4). e-Government is also conceptualized in [3] as governments' use of ICTs together with organizational change to enhance their structures and operations.

The significance of e-government is well recognized by both developed and developing countries. However, previous reports that have explored e-government on a global scale, such as the series of reports conducted by the United Nations (UN), point to a gap between developed and developing countries in terms of the adoption and implementation of e-government, and suggest that this gap needs to be addressed [4,5]. Saudi Arabia is one of the developing countries that needs to learn from more developed countries so as, to ensure compliance with the best global e-government practices.

In Saudi Arabia, e-government is regarded as beneficial for a variety of different purposes, most notably those relating to the management of organizational processes and religious events, such as pilgrimages or the Hajj. In fact, there have been many e-government initiatives in the country in recent years, as it is in a process of transition to e-government. Indeed, reports of the National Transformation Program 2020 and of the Saudi Vision 2030 highlight the importance of further development of e-government initiatives. These reports also indicate the Saudi government's commitment to the adoption and use of ICT to fully transition to e-government. However, research investigating the evolution of e-government in Saudi Arabia over time is still limited, especially when compared to that conducted in other developed countries. For example, a study of e-government [1] (p. 4) calls for further research "to evaluate the accomplishment of e-government projects in the Arab World." This paper, therefore, aims to investigate and assess the development of e-government in Saudi Arabia.

The research questions that the current study aims to address are: 1) How has e-government in Saudi Arabia evolved during the period of 2003 to 2020?, 2) How does the development of e-government in Saudi Arabia differ from the development of e-government in the USA and in the Republic of Korea (RoK)? These two countries have been chosen because, according to UN e-government reports, they have been ranked first globally more than one occasion; the USA was ranked first in 2003, 2004, and 2005, while the RoK was ranked first in 2010, 2012, 2014.

This research aims to evaluate the development of e-government in Saudi Arabia during the period 2003-2020, and to compare it with the development of e-government in the

fmmehmadi@uqu.edu.sa.

USA and the RoK, given that these two countries are considered to be among e-government world leaders. The objectives of the study are as follows:

- To track the evolution of e-government in Saudi Arabia during the period 2003 to 2020.
- To compare the level of progress of e-government in Saudi Arabia with that of world leading countries, namely the USA and the RoK.
- To identify the similarities and differences that relate to e-government development between Saudi Arabia, the USA, and the Rok.
- To identify the most important aspects that need to be considered to enhance e-government in Saudi Arabia.
- To develop an information management strategy prioritizing the key aspects of e-government that need to be addressed to further improve e-government in Saudi Arabia.

The present study adds to existing knowledge in the field. Firstly, it differs from most prior studies, which, while revealing valuable insights about e-government, often focused on case studies of a single country. In this study, however, the approach is different, in that it offers a cross-country comparison by evaluating the level of e-government progress in Saudi Arabia alongside that in the USA and the RoK. Secondly, it is hoped that the findings of this study and the lessons that can be learned from e-government development in the USA and the RoK can be used to reposition Saudi Arabia on its roadmap to improve its EGDI ranking in the upcoming UN review reports. As one author [6] (p. 7) notes that for successful implementation of e-governments, there is a need to learn from the experience of other countries especially those which were considered successful or highly ranked. Thirdly, the findings of this study can be used by policy and decision makers in Saudi Arabia to consider specific factors or areas that may help to further develop e-government in the country.

The subsequent sections provide a review of the relevant literature, an explanation of the research methodology, a detailed presentation of the results, a discussion of the key findings, and the conclusions of the study.

II. Review of Literature

A. Benefits of e-Government

The adoption and implementation of e-government can bring various benefits for citizens, businesses, and public sector organizations [1,2,6]. In terms of benefits for citizens, it can lead to the delivery of services in a way that is not only convenient, but also cost-effective. Another benefit is reducing the time citizens need to spend applying for and following-up on government services. In addition, e-government can benefit by providing online access to government services via different devices, such as computers and smart phones. As discussed in [7,8], a further advantage is transparency, since citizens can easily monitor their applications for government services and the different procedures required to process these applications.

E-government can also be beneficial for businesses because it is likely to enhance the quality of services and facilities that are provided by the government to businesses and private sector companies. This, in turn, can lead to the provision and delivery of convenient, cost-effective, and time-saving services to these businesses [9]. In addition, according to [10], e-government systems that are used by private sector entities can be beneficial in terms of the simplification and improvement of supply chain and marketing process management.

There are also many benefits of e-government when it comes to public organizations, including the efficient and effective management of the data and information created, processed, and delivered by these organizations. Indeed, the adoption and implementation of e-government can offer an opportunity to develop a knowledge base system that can be collaboratively used for better services provision. In this regard, e-government can also further improve the performance of the public sector, as it supports information sharing among the sector's different organizations [6].

B. Barriers to e-government

There are various challenges that can hinder the successful adoption and implementation of e-government. These, according to [6], can also delay the progress made towards realizing its full potential. Although such obstacles vary from one country to another, there are common barriers that most countries encounter [11]. One of the key challenges to e-government, especially in developing countries, is inadequate ICT infrastructure [12]. E-government initiatives in developing countries often face technical obstacles, such as a lack of compatible technical infrastructure and of shared standards among different government departments [6]. This, in turn, can lead to partial or complete failure of e-government initiatives. Indeed, one study [13] finds, in developing countries, this technical factor contributes to the complete failure of 35% of e-government projects, and the partial failure of 50% of e-government projects. Additionally, three recent studies [14] investigating barriers to e-government in the Arab world, come to the conclusion that inadequate ICT infrastructure is the technical factor most likely to impede the successful implementation of e-government in Arab countries.

Another obstacle to e-government implementation is related to policy and regulation issues. The transition to e-government systems, especially in developing countries, often only focuses on technical issues, like sustaining suitable hardware and software. However, this transitional process should also consider organizational issues such as, developing a range of adequate rules, policies, and guidelines to deal with such matters as archiving, e-signatures, and data protection through e-government systems [6]. One of the challenges that faces developing countries when they transit to e-government systems is the lack of adequate rules to regulate e-government activities or processes. This can delay or lead to the potential failure of the implementation of e-government. E-government websites often ask for personal data, and people maybe be reluctant to make use of these systems if they feel that their privacy is at risk or that information security is not sufficiently guaranteed [11,15].

A further challenge that can negatively affect the adoption and implementation of e-government is the lack of qualified staff and training. Two e-government studies have stated that "government should train its employees and citizens in basic skills of dealing with the computer and Internet in order to let them participate in e-government development applications" [6] (p. 40). Previous studies indicate that a lack of ICT skills among e-government personnel can be a common problem, and is often reported as a significant challenge, particularly fort developing countries [16]. This is frequently coupled with inadequate staff training [14], which may further hinder the proper implementation of e-government, given that it requires qualified employees with the technical skills vital for e-government logistics, such as the installation of ICT systems and the management of online processes and applications [6].

C. e-Government in Saudi Arabia

Saudi Arabia is located in South-West Asia and is one of the largest Arab countries, covering an area of around 2,000,000 square kilometers. The country has a population of around 33,414,660 people [17], and an approximate population growth of 2.52% every year [18].

The kingdom of Saudi Arabia (KSA) recognizes the significance of the adoption and use of e-government. A Royal Decree directive to the Ministry of Communications and Information Technology (M.C&IT) was issued in 2003 to develop an e-government transitional plan [19]. In 2005, the M.C&IT, in collaboration with the Ministry of Finance and the Commission of Communications and Information Technology (C.C&IT) developed the e-government national program, called Yesser [20]. The two main objectives of this program are: 1) to raise productivity and efficiency of public sector organizations, and 2) to provide easy-to-use services for both citizens and businesses. According to [19–21], the Yesser program has put forward several initiatives to address these objectives. These initiatives are related to human capital, e-services provision, national databases, e-procurement systems, infrastructure, and institutional frameworks including rules and regulations.

Nonetheless, the transition to an e-government system in a developing country like Saudi Arabia is often a complex process with various challenges along the way. Indeed, e-government in developing countries can be fraught with difficulties arising from barriers such as a lack of adequate ICT infrastructure, and bureaucratic inefficiency [6]. A study by [16] indicates that Arabian Gulf countries, including Saudi Arabia often have incomplete ICT infrastructure, owing to the absence of local IT industries, which, when compared to foreign IT industries, may be better able to accommodate ICT infrastructure requirements. Another study by [22] highlights the inadequacy of ICT infrastructure and websites for people with special needs such as, blind citizens, and indicates that this problem can hinder the adoption and implementation of e-government. It concludes that there is a need to review accessibility policies in all Arabian Gulf countries including Saudi Arabia so as to "accelerate the transition to accessible e-government websites" which should also "serve all kinds of citizens, including [the] disabled" [22] (p. 6). This is in line with the findings of another study [23], which argues that a significant effort should be made to ensure that websites in Saudi Arabia become more accessible to all citizens.

An additional challenge is the lack of staff qualified to operate and manage e-government. One study [6] argues that this challenge, if not properly addressed through adequate training, can hinder the implementation of e-government initiatives. Another study [16] indicates that a lack of qualified staff can be regarded as a problem in Saudi Arabia and other developing countries. Human resources and qualified personnel are considered essential to realize success in terms of e-government, since it requires different human capacities, whether technical or managerial [6,14]. To address the challenge relating to human resources in developing countries, including Saudi Arabia, two scholars [6] suggest developing a knowledge management strategy, focusing on the provision of adequate training, accessible to all e-government staff.

Policy issues represent another obstacle. One study [22] contends that it is essential for different government agencies to operate with and adhere to the same set of rules. However, this is often not the case in developing countries, including Saudi Arabia, where "many government agencies have their own regulatory environment and strategic priorities" [16,24] (p. 3). Two studies [6] [14] also reveal similar results, and suggest developing a range of policies that address different e-government activities such as archiving, signing, and data transmission and protection.

This study argues that learning lessons from world-leading countries in terms of e-government implementation is an essential step that developing countries like Saudi Arabia should consider. A good starting point from which to evaluate the evolution of e-government in Saudi Arabia and to compare it to that of other developed countries is the series of e-government reports conducted by the UN since 2003. The next section provides an overview of these reports.

D. United Nations' e-Government Survey Reports

The UN, which is a leading global organization, recognizes the significance of e-government and the range of different benefits it can bring to citizens, businesses, and organizations. Indeed, through its series of e-government reports starting in 2003, the UN encourages countries all around the world to adopt and implement e-government. These reports are based on the collaborative work of the United Nations Department of Economic and Social Affairs (DESA), and the Division for Public Institutions and Digital Government (DPIDG), previously known as the Division for Public Administration and Development Management (DPADM) [5]. The reports assess the status of e-government in different countries around the world, according to a specific set of measurements. Specifically, the assessment of e-government in the UN reports is based on the following three major measurements [4,5,25]:

- Online services index (OSI), also called web measurement index: This includes the measurement of four stages of online services, which are emerging presence (stage1), enhanced presence (stage2), transactional presence (stage3), and connected presence (stage4).

- Telecommunications infrastructure index (TII): This includes the measurement of four indicators, which are the number of internet users per 100 people, the number of mobile cellular subscribers per 100 people, the number of active mobile broadband subscribers per 100 people, and the number of fixed broadband subscribers per 100 people.
- Human capital index (HCI): This includes the measurement of two indicators, which are adult literacy and the combined primary, secondary, and tertiary gross enrolment ratio.

Based on the above three measurements, the UN developed the e-government development index (EGDI), according to which each country is given a value indicating its world rank. The EGDI is a weighted average of the OSI, the TII, and the HCI. The EGDI can be mathematically presented as follows: EGDI = 1/3 ((OSI) + (TII) + (HCI)) [4,5,25]. Further information about the indicators of e-government can be found in Appendix A.

III. Research Methodology

The study made use of data collected from a number of sources, including e-government official reports, portals, and ICT statistics, all relating to the three countries considered in the study. However, the research focused primarily on the UN e-government survey reports, not only because these reports were considered thorough, rigorous, and comprehensive, but also because they were regarded as precise, and relevant to the study's research questions. The UN is considered a pioneering organization that assesses e-government on a global level.

For the purpose of the study, the comparison between the development of e-government in Saudi Arabia and in the USA and the RoK was made according to a specific set of numerical data taken from UN e-government reports ranging from 2003 to 2020.

It is important to consider data analysis as a critical part of any research methodology. It involves the process of using various methods in order to reveal specific results that address certain questions or objectives. In the case of this study, a number of data analysis techniques were used. These techniques are detailed as follows.

A. Comparative, Cross-Country Analysis

The study used comparative analysis to identify similarities and differences between Saudi Arabia, the USA, and the RoK, in relation to e-government development indicators, namely, EGDI, OSI, TII, and HCI. The timeframe of this analysis was the period 2003 to 2020.

B. Trend Analysis

Trend analysis was used to measure the development of e-government in Saudi Arabia, the USA, and the RoK over the period 2003 to 2020. Specifically, trend analysis in the present study was employed to determine the uptrend (positive slope) and the downtrend (negative slope) in relation to the e-government indicators under consideration: EGDI, OSI and its sub-indicators (OSI1, OSI2, OSI3, OS4), TII and its sub-indicators (TII1, TII2, TII3, TII4), and HCI and its sub-indicators (HCI1, HCI2). In research methods literature, trend analysis is often referred to as a special data analysis technique of regression analysis. In this form of data analysis, the independent variable is time, whereas the dependent variable is the variable to be forecast [26]. In this study, the independent variable is represented by the period 2003 to 2020, while the dependent variables are represented by the above indicators.

C. Gap analysis

According to research methods literature, gap analysis can be used for various purposes. In the present study, it was used is to measure the gap between the development of e-government in Saudi Arabia and the development of e-government in the USA and the RoK, with reference to the indicators: EGDI, OSI and its sub-indicators (OSI1, OSI2, OSI3, OS4), TII and its sub-indicators (TII1, TII2, TII3, TII4), and HCI and its sub-indicators (HCI1, HCI2).

In terms of data analysis software, SPSS was used to conduct the trend analysis and to deal with missing data, by using the multiple imputation statistical method. MS Excel was also used to make the tables required for the comparative, cross-country analysis, the trend analysis, and the gap analysis, and to draw figures from these tables.

IV. Results

A. An Overview of the Development of e-Government in Saudi Arabia, the USA, and the RoK

1) EGDI for Saudi Arabia, the USA, and the RoK: The EGDI values for each country considered in the study can be used to give an overview of the level of overall progress made in relation to the adoption and implementation of e-government in these countries. Fig. 1 shows the EGDI values for Saudi Arabia, the USA, and the RoK during the period 2003 to 2020.

Fig. 1 indicates that, although the EGDI values for Saudi Arabia are still low, the country has made good progress in that these values have steadily increased over the years. Nonetheless, these values also indicate that Saudi Arabia needs to further improve e-government in the future. By contrast, in the case of e-government in the USA, the EGDI values were high, especially during the period 2003-2005. However, these values then dropped, particularly, in 2016. The EGDI values for the RoK began low during the period 2003-2010, then subsequently increased during the period 2012-2020.

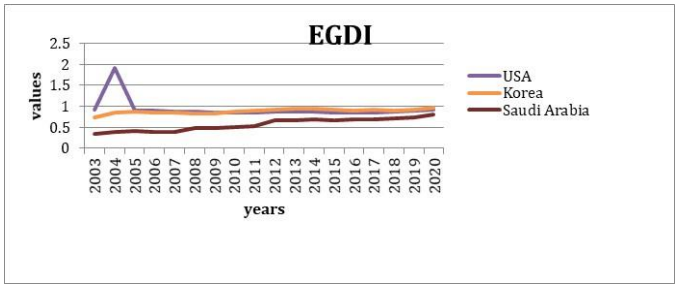


Fig. 1. EGDI Values for the USA, the RoK, and the KSA during the Period 2003 to 2020 [3, 4, 25].

As shown in Fig. 1, the EGDI values for Saudi Arabia have generally improved over the past 15 years. This improvement is particularly clear in 2010, 2012, 2014, 2016, 2018, and 2020. In a similar vein, the EGDI values for the USA have significantly improved over time. However, there was a notable drop in the EGDI for the USA in 2008, although it did begin to increase again in 2010. Fig. 1 also indicates a rapid improvement of the EGDI values for the RoK, particularly, over the period from 2012 to 2020.

2) Indicators of e-government for Saudi Arabia, USA, and RoK: The e-government indicators (OSI, TII, and HCI) for Saudi Arabia, the USA, and the Rok are based on the UN e-government reports, shown in Appendix A. These indicators can be used to give a descriptive overview of the level of overall progress that has been made. Fig. 2 shows the values of OSI, TII, and HCI for the USA, the RoK, and Saudi Arabia during the period 2003 to 2020.

As can be seen from the values shown in Fig. 2, Saudi Arabia has made good progress in terms of all three indicators of the EGDI during the period 2003 to 2020. However, these values, if compared with those relating to the USA and the RoK, indicate that further development of e-government in Saudi Arabia is still required.

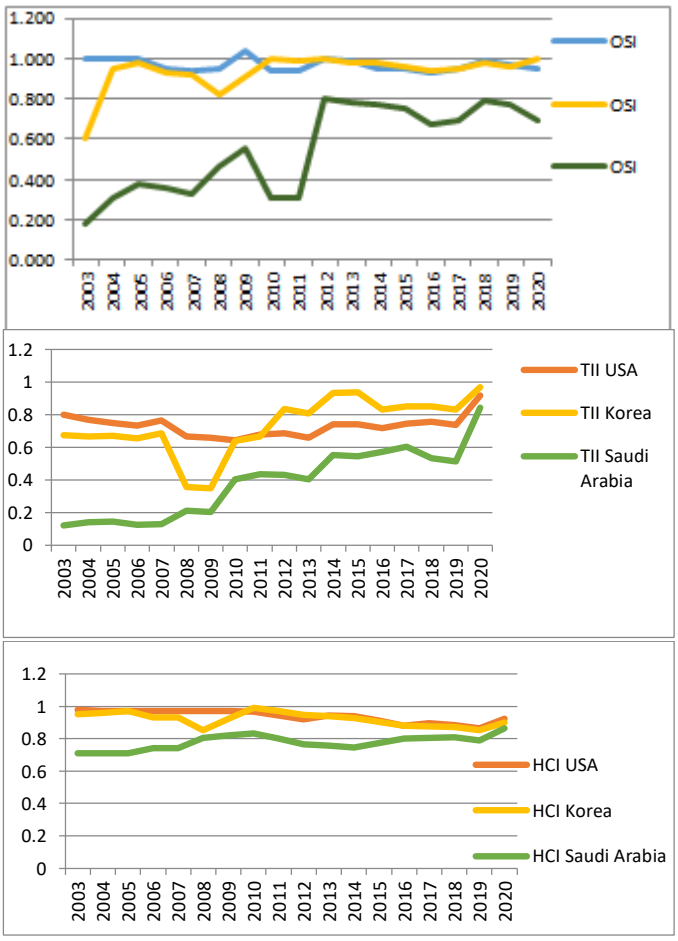


Fig. 2. OSI, TII & HCI Values for the USA, the Rok, and the KSA during the Period 2003 to 2020 [3, 4, 25].

In terms of OSI, the Rok ranks highest with the value of 1, but the USA also has a high value, at 0.94. OSI for Saudi Arabia, on the other hand, has improved from only 0.18 in 2003, to 0.68 in 2020. Similarly, the TII values for both the USA and Saudi Arabia indicate an improvement, although the TII values for the RoK for the same period are better still. Nonetheless, it is worth noting that while the TII values for Saudi Arabia show constant improvement, the TII values for both the USA and the RoK show unsteady progress, most notably between 2008 and 2013 in the case of the USA, and between 2008 and 2011 in the case of the RoK.

The HCI values for Saudi Arabia indicate constant progress from 2003 to 2020 except during the period from 2012 to 2016, where these values dropped slightly. Similarly, HCI for both the USA and the RoK has higher values initially, especially from 2003 to 2014, but then decreases, particularly from 2016 to 2019 in the case of the USA, and from 2016 to 2020 in the case of the RoK.

Fig. 2 shows that Saudi Arabia has made steady progress in all indicators of e-government development during the period from 2014 to 2020, although a drop was seen in the period from 2008 to 2012. Fig. 2 also indicates that, over time, HCI for Saudi Arabia has improved more than TII or OSI. This is also the case for both the USA and the RoK, where HCI values were higher than TII and OSI values.

B. Trend Analysis Results

Trend analysis was used to discover the uptrend (positive slope) as well as the downtrend (negative trend) in the UN e-government indicators for the USA, the RoK, and Saudi Arabia. This form of data analysis is a distinct case of regression analysis in which time is considered the independent variable, whereas the variable to be forecasted is considered the dependent variable. In this study, dependent variables represent the EGDI and its indicators.

1) Trend analysis of EGDI: Table I shows the EGDI trend analysis for the USA, the RoK, and Saudi Arabia during the period 2003 to 2020.

As shown in Table I, the trend analysis of the EGDI values for Saudi Arabia indicates a steady improvement, with a positive trend value of 0.268 compared with a value of -0.0166 for the USA, and a value of 0.0077 for the RoK. This result highlights the fact that the EGDI for Saudi Arabia has generally improved over time, with the peak of its EGDI value in 2020 (0.7991). On the other hand, the trend analysis of the EGDI values for the USA points to a negative trend value, and the trend analysis of the EGDI for the RoK points to a positive trend value. This indicates that while the RoK has maintained high EGDI values over time, the EGDI values for the USA have dropped over time even though the USA had high EGDI values during earlier periods, especially during the period 2003 to 2005.

2) Trend analysis of OSI and its sub-indicators: The trend analysis of the OSI values for the countries considered in this study can be used to infer the level of progress made in relation to online services and its sub-indicators, and whether the results of the trend analysis indicate a positive or a

negative values. Table II shows the OSI trend analysis for Saudi Arabia, the USA, and the RoK during the period 2003 to 2020.

The trend analysis of the OSI for Saudi Arabia indicates that online services have gradually improved, and this improvement has resulted in a positive trend value of 0.0388635. In a similar vein, the trend analysis of the OSI for the RoK points to a positive value of 0.01042673. On the other hand, the trend analysis of the OSI for the USA indicates a negative value of -0.0024602. This is because the OSI values were lower during the period 2012 to 2020 than during the period 2003 to 2005, and in 2012.

In terms of the OSI sub-indicators, Table II shows that Saudi Arabia has made good improvement with regards to the four online services stages that represent the OSI sub-indicators, with positive values of 1.35367647, 1.93143382, 6.49834559, and 2.93839614. In comparison with the OSI sub-indicators for the USA, Table II shows that stages 1, 2, and 4 have negative trend values of -0.275, -0.7416131, and -0.1834559, respectively whereas only stage 3 has a positive trend value of 0.30857077. This is because values of the OSI sub-indicators for the USA were higher during the period 2003 to 2008 than during the period 2016 to 2020, where a decrease in these values can be particularly noted. On the other hand, the OSI sub-indicators for the RoK have positive trend values for three of the stages of online services, which are stage 1, 3, and 4, with values of 0.005, 1.28524816, and 1.78713235, respectively, and one negative trend value of -.04082491 for stage 2.

3) Trend analysis of TII and its sub-indicators: Table III shows the results of the trend analysis of TII and its sub-indicators for the USA, the RoK, and Saudi Arabia during the period 2003 to 2020.

TABLE I. TREND ANALYSIS RESULTS OF THE EGDI FOR THE USA, THE ROK, AND THE KSA DURING THE PERIOD 2003 TO 2020

Country	Trend (Slope) of EGDI
The USA	-0.0166
The RoK	0.0077
The KSA	0.0268

TABLE II. TREND ANALYSIS RESULTS OF OSI AND ITS SUB-INDICATORS FOR THE USA, THE ROK AND THE KSA DURING THE PERIOD OF 2003 TO 2020

Indicator	Country		
	The USA	*The RoK*	*The KSA*
OSI	-0.0024602	0.01042673	0.0388635
OSI1	-0.275	0.005	1.35367647
OSI2	-0.7416131	-0.4082491	1.93143382
OSI3	0.30857077	1.28524816	6.49834559
OSI4	-0.1834559	1.78713235	2.93839614

TABLE III. TREND ANALYSIS RESULTS FOR TII AND ITS SUB-INDICATORS FOR THE USA, THE ROK AND THE KSA DURING THE PERIOD 2003-2020

Indicator	Country		
	The USA	*The RoK*	*The KSA*
TII	0.00372	0.01933	0.03858
TII1	2.12107	2.41639	5.00924
TII2	4.5183	3.75414	10.0155
TII3	3.11295	4.31919	4.90803
TII4	1.15821	1.012439011	1.19233

TII for Saudi Arabia improved gradually, especially in 2020 when its value jumped to 0.8442 from only 0.211 in 2008. Thus, the results of the trend analysis of the TII values for Saudi Arabia during the period 2003 to 2020 show an annual improvement with a positive trend value of 0.03858. In a similar vein, the results of the trend analysis of the TII values for both the USA and the RoK indicate positive values of 0.00372 for USA and 0.01933 for RoK. The following explains the results of the trend analysis of the TII sub-indicators.

The first sub-indicator of TII to consider here is the number of estimated internet users per 100 persons. The values of this indicator have been slowing improving over time. However, this indicator notably increased for Saudi Arabia in 2020. Similarly, it improved remarkably for the USA and the RoK in the same year (2020). The estimated number of internet users in Saudi Arabia increased to 87.27 in 2020, although it was just 18.66 in 2008. This has resulted in a positive trend value of 5.00924 for Saudi Arabia. Likewise, the estimated number of internet users in the USA as well as in the RoK increased in 2020, and this has resulted in positive trend values of 2.12107, and 2.41639 for the USA and the RoK, respectively.

A second indicator to consider is the number of mobile subscribers. The number of mobile subscribers in Saudi Arabia was just about 78 in 2008, but this number gradually increased over time especially during the period 2012 to 2020. This has resulted in a positive trend of 10.0155. Similarly, in the USA the estimated number of mobile subscribers was just about 74 in 2008, but this value increased remarkably, especially during the period 2014 to 2020. This has resulted in a positive trend value of 4.5183. On the other hand, the estimated number of mobile subscribers in the RoK was just about 83 in 2008, but this value increased greatly, especially during the period 2012 to 2020. This has resulted in a positive trend value of 3.75413. E-government can be accessed and utilized via different devices, and the results of the trend analysis of this indicator points to great potential for e-government, particularly with smart phones users.

In terms of active mobile broadband subscriptions, which is the third sub-indicator of TII, Table III shows that the USA, the RoK, and Saudi Arabia, all have positive trend values of 3.11395, 4.31919, and 4.90803, respectively. The increasing number of mobile subscribers, which, in turn enables subscription to active mobile broadband, may be the reason why the trend values of this sub-indicator are also increasing.

On the other hand, these countries have varied but positive trend values of 1.15821, 1.012439011, and 1.19233, respectively, in relation to fixed broadband subscription, which is the fourth sub-indicator of the TII. Nonetheless, the values of this indicator for Saudi Arabia during the period 2003 to 2020 are still lower than those of both the USA and the RoK, which indicates that this area needs further development.

4) Trend analysis of HCI and its sub-indicators: Table IV shows the results of the trend analysis of the HCI and its sub-indicators for the USA, the RoK, and Saudi Arabia during the period 2003 to 2020.

The HCI for Saudi Arabia, the USA, and the RoK generally improved during the period 2003 to 2020. However, the HCI for Saudi Arabia decreased at times, especially during the period 2012 to 2016. As a result, as shown in Table IV, the HCI for Saudi Arabia has a negative trend value of -0.0020209. Similarly, in 2008 the HCI for the RoK was 0.8056, but this value then decreased, especially during the period 2012 to 2018, and this resulted in a negative trend value of -0.0114045. Likewise, in 2008 the HCI for the USA was 0.9711, but during the period 2010 to 2018 this value began to decrease, which resulted in a negative trend value of -0.00097055.

In terms of the HCI sub-indicators, Table IV shows a positive trend value of 1.296 for Saudi Arabia in relation to HCI1, which represents adult literacy. On the other hand, Table IV shows a neutral value of 0 for both the USA, and the RoK in relation to HCI1, and this indicates maintenance of the same position over time. In contrast, Table IV shows a negative trend value of -0.2147455 for the RoK in relation to HCI2, which represents gross enrolment ratio, while positive trend values can be noted from Table IV in relation to HCI2 for both Saudi Arabia (2.20163636), and the USA (0.37845455).

C. Gap Analysis Results

Gap analysis is an approach for the steps to be taken to enable change from a present situation regarding an issue or a factor to a future, more desirable situation [26]. This can be achieved by studying, comparing, and understanding gaps between the examined factors or indicators. In the context of the current study, this form of analysis refers to an evaluation of the gaps between the e-government indicators for Saudi Arabia and those of the USA, and the gaps between the e-government indicators for Saudi Arabia and those of the RoK. Because data that represent the three main indicators of e-government exist in different forms, a per unit analysis was used to convert them to percentages to enable the proper comparison and to identify accordingly the priorities for indicators that require more attention.

1) Gap analysis of OSI indicators: Table V shows the results of the gap analysis between Saudi Arabia and the USA, and that between Saudi Arabia and the RoK in relation to indicators of their online services index.

In Table V, the indicators of the OSI are OSI1, OSI2, OSI3, and OSI4, which respectively represent: stage 1 (emerging presence), stage 2 (enhanced presence), stage 3 (transactional presence), and stage 4 (connected presence). Please see Appendix A for a description of these stages. The per unit analysis values for these four indicators were calculated according to the following formulas:

- P_OSI1= ((OSI1-Min value)/ (Max value -Min value)) *100
- P_OSI2= ((OSI2-Min value)/ (Max value -Min value)) *100
- P_OSI3= ((OSI3-Min value)/ (Max value -Min value)) *100
- P_OSI4= ((OSI4-Min value)/ (Max value -Min value)) *100

Gap analysis values, on the other hand, were calculated according to the following formulas:

- The USA P_OSI1 - the KSA P_OSI1
- The USA P_OSI2 - the KSA P_OSI2
- The USA P_OSI3 - the KSA P_OSI3
- The USA P_OSI4 - the KSA P_OSI4
- The Rok P_OSI1 - the KSA P_OSI1
- The Rok P_OSI2 - the KSA P_OSI2
- The Rok P_OSI3 - the KSA P_OSI3
- The Rok P_OSI4 - the KSA P_OSI4

According to Table V, there is a significant gap in relation to all four sub-indicators of the OSI between the values of Saudi Arabia and the values of both the USA and the RoK, which were equivalent. The results of the gap analysis shown in Table V indicate that the gap between the values of Saudi Arabia and those of the USA and the RoK are 26.93%, 19.05%, 28.58%, and 54.55% in relation to OSI1, OSI2, OSI3, and OSI4, respectively. Notably, this gap is greater for OSI4 than for the other three sub-indicators of the OSI. This indicates that Saudi Arabia needs to further invest in developing its online services at different levels to address the current gap.

2) Gap analysis of TII indicators: Table VI shows the results of the gap analysis between Saudi Arabia and the USA, and that between Saudi Arabia and the RoK in relation to the indicators of their TII.

TABLE IV. TREND ANALYSIS RESULTS OF HCI AND ITS SUB-INDICATORS FOR THE USA, THE ROK AND THE KSA DURING THE PERIOD 2003-2020

Indicator	Country		
	The USA	***The RoK***	***The KSA***
HCI	-0.0097055	-0.0114045	-0.0020209
HCI1	0	0	1.296
HCI2	0.37845455	-0.2147455	2.20163636

TABLE V. OSI INDICATORS PER UNIT AND GAP ANALYSIS

Country	**OSI1**	**OSI2**	**OSI3**	**OSI4**
The USA	100	100	100	100
The RoK	100	100	100	100
The KSA	74.07	80.95	72.09	45.45
Max value	100	100	100	100
Min value	3.70	0	2.33	0
Per Unit Analysis	**P_OSI1**	**P_OSI2**	**P_OSI3**	**P_OSI4**
The USA	100	100	100	100
The RoK	100	100	100	100
The KSA	73.07	80.95	71.42	45.45
Gap Analysis	**Gap_P_OSI1**	**Gap_P_OSI2**	**Gap_P_OSI3**	**Gap_P_OSI4**
Between the KSA and the USA	26.93	19.05	28.58	54.55
Between the KSA and the RoK	26.93	19.05	28.58	54.55

TABLE VI. TII INDICATORS PER UNIT AND GAP ANALYSIS

Country	**TII1**	**TII2**	**TII3**	**TII4**
The USA	93.3	120	120	33.8
The RoK	96.02	120	113.63	41.6
The KSA	87.27	120	111.09	20.24
Max value	99.01	120	120	51.24
Min value	0	20.74	0	0
Per Unit Analysis	**P_TII1**	**P_TII2**	**P_TII3**	**P_TII4**
The USA	94.23	99.105	100	65.96
The RoK	96.98	99.105	94.69	81.18
The KSA	87.86	99.105	92.57	39.500
Gap Analysis	**Gap_P_TII1**	**Gap_P_TII2**	**Gap_P_TII3**	**Gap_P_TII4**
Between the KSA and the USA	6.37	0	7.43	26.46
Between the KSA and the RoK	9.12	0	2.12	41.68

In Table VI, the column headers TII1, TII2, TII3, and TII4, respectively represent the sub-indicators of the TII: the number of internet users per 100 persons; the number of mobile cellular subscribers per 100 persons; the number of active mobile broadband subscribers per 100 persons; and the number of fixed broadband subscribers per 100 persons. The per unit analysis values of these indicators for each country were calculated according to the following formulas:

- P_TII1= ((TII1-Min value)/ (Max value -Min value)) *100
- P_TII2= ((TII2-Min value)/ (Max value -Min value)) *100
- P_TII3= ((TII3-Min value)/ (Max value -Min value)) *100
- P_TII4= ((TII4-Min value)/ (Max value -Min value)) *100

The gap analysis values of these four indicators, on the other hand, were calculated according to the following formulas:

- The USA P_TII1 - the KSA P_TII1
- The USA P_TII2 - the KSA P_TII2
- The USA P_TII3 - the KSA P_TII3
- The USA P_TII4 - the KSA P_TII4
- The Rok P_TII1 - the KSA P_TII1
- The Rok P_TII2 - the KSA P_TII2
- The Rok P_TII3 - the KSA P_TII3
- The Rok P_TII4 - the KSA P_TII4

According to Table VI, there is a significant gap between Saudi Arabia and both the USA and the RoK in relation to two sub-indicators of the TII, which are TII3, and TII4. The results of gap analysis that are shown in Table VI indicate that, in

terms of TII3, there is a gap between Saudi Arabia and the USA by 7.43%, and by 19.05% between Saudi Arabia and the RoK, whereas, in terms of TII4, the gap is even greater between Saudi Arabia and the USA (26.46%) and between Saudi Arabia and the RoK (41.68%). This indicates that Saudi Arabia needs to further develop its national telecommunication infrastructure, especially in relation to TII3 and TII4, to address this current gap.

On the other hand, there is a moderate gap between Saudi Arabia and the USA of 6.37%, and between Saudi Arabia and the RoK of 9.12% with regards to TII1. Notably, there is no gap between Saudi Arabia and either the USA or the RoK in relation to TII2, which indicates improvement and progress for Saudi Arabia.

3) Gap analysis of HCI indicators: Table VII shows the results of the gap analysis between Saudi Arabia and the USA, and that between Saudi Arabia and the RoK in relation to indicators of their HCI.

In Table VII, HCI1 and HCI2 respectively represent indicators of the HCI which are Adult Lteracy (AL) and Gross Enrolment Ratio (GER). The per unit analysis values for these two indicators for each country were calculated according to the following formulas:

- P_HCI1= ((HCI1-Min value)/ (Max value -Min value)) *100
- P_HCI2= ((HCI2-Min value)/ (Max value -Min value)) *100
- P_HCI3= ((HCI3-Min value)/ (Max value -Min value)) *100
- P_HCI4= ((HCI4-Min value)/ (Max value -Min value)) *100

The gap analysis values of these two indicators, moreover, were calculated according to the following formulas:

- The USA P_HCI1 - the KSA P_HCI1
- The USA P_HCI2 - the KSA P_HCI2
- The USA P_HCI3 - the KSA P_HCI3
- The USA P_HCI4 - the KSA P_HCI4
- The Rok P_HCI1 - the KSA P_HCI1
- The Rok P_HCI2 - the KSA P_HCI2
- The Rok P_HCI3 - the KSA P_HCI3
- The Rok P_HCI4 - the KSA P_HCI4

According to Table VII, there is a moderate gap of 4.54% between Saudi Arabia and both the USA and the RoK in relation to HCI1. This indicates that further effort is required to address the gap. On the other hand, the HCI2 values for Saudi Arabia are even higher than the HCI2 values for the USA and the RoK, which indicates significant improvement and progress for Saudi Arabia.

TABLE VII. HCI INDICATORS PER UNIT AND GAP ANALYSIS

Country	**HCI1**	**HCI2**
The USA	95.33	100
The RoK	99	98.38
The KSA	99	97.48
Max value	100	115.41
Min value	19.1	17
Per Unit analysis	**P_OSI1**	**P_OSI2**
The USA	94.22	84.34
The RoK	98.76	82.69
The KSA	98.76	81.78
Gap analysis	**Gap_P_OSI1**	**Gap_P_OSI2**
Between the KSA and the USA	4.54	-1.65
Between the KSA and the RoK	4.54	-2.56

V. DISCUSSION

While e-government is well-instituted in developed countries, this is not usually the case in developing countries due to the various challenges these countries face in terms of e-government adoption and implementation [12]. These challenges, whether technical, organizational, or legal, contribute to the overall e-readiness index values, which are published in UN e-government reports, of these countries. Thus, for each developing country, it is vital to identify, and accordingly to address each of these challenges. Saudi Arabia is one of these developing countries, and it also encounters technical, organizational, and legal challenges [1,14,23] that, if adequately addressed, could boost the country's rank and its EGDI index, which are shown in the 2020 UN e-government report. The results of the current study reveal that the values of the EGDI and its components for Saudi Arabia have steadily increased in recent years. However, the results also indicate that there are different areas that need further attention and, consequently, additional improvement. An information management strategy (IMS) in support of e-government in Saudi Arabia is therefore proposed and is shown in Fig. 3 below.

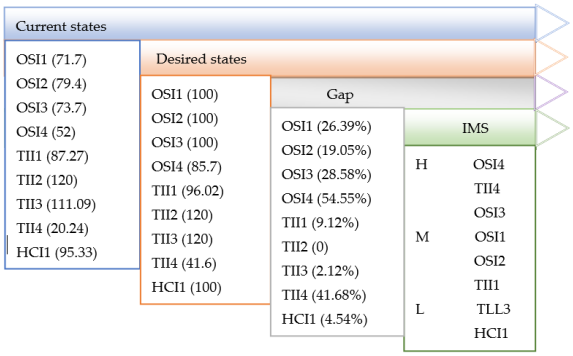


Fig. 3. Information Management Strategy for e-government in Saudi Arabia.

The strategy shown in Fig. 3 describes the current and desired states of e-government in Saudi Arabia in relation to e-government indicators that have been analyzed in the current study. The values for the current state represent the results of the trend analysis of the OSI, TII, and HCI indicators for Saudi Arabia, whereas the values of the desired state represents the results of the trend analysis of the OSI, TII, and HCI indicators for the RoK, as these were the highest values of the three countries considered in this study. Furthermore, the values in the third column represent the gaps percentages based on the gap analysis that was conducted in the current study between Saudi Arabia and the RoK in relation to the OSI, TII, and HCI indicators.

The information management strategy shown in Fig. 3 also sets three levels of priorities for areas that Saudi Arabia needs to consider to better comply with international e-government practices. The highest priorities include OSI4, TII4, and OSI3. Saudi Arabia needs to further develop its online services, particularly its connected and interactive services, its Internet fixed-broadband infrastructure, and its transactional online services. Furthermore, the medium-level priorities include the need to re-consider increasing the presence of additional and more developed online services and increasing the number of internet users all over the country. E-government online services for all citizens, including those with special needs, requires more attention. Finally, Fig. 3 sets low priorities for e-government in Saudi Arabia, which include further development of active mobile broadband subscriptions and further investment in human capital. It is important to re-consider the development of the HCI in Saudi Arabia, as this factor may have an impact on other e-government indicators, namely, the OSI and TII. Indeed, a recent study by [14] (p. 10) suggested that "whenever the good HCI exist, the growth of OSI and TII will increase" and "when there is no good HCI, the growth of OSI and TII will decrease".

The significance of the proposed strategy in this study is that it depicts the main factors that the Saudi government needs to account for during the continued process of developing and implementing e-government. Specifically, this study argues that technical factors, especially those relating to online services and infrastructure issues, need to be first re-considered to achieve successful implementation of e-government and to consequently improve the current state of country's EGDI. This is in line with a recent study [14] that found that infrastructure was the key challenge impacting the successful implementation of e-government in the Arab world.

VI. Conclusion

The aim of the current study was to benchmark the level of progress of the e-government in Saudi Arabia with two global e-government leaders: the USA and the RoK. This researcher conducted: a) a comparative, cross-country, longitudinal analysis of the e-government development index (EGDI) values that relate to Saudi Arabia, the USA and the RoK, b) a trend analysis of the online services, telecommunication infrastructure, and human capital indicators, and c) a gap analysis to pinpoint the gap between Saudi Arabia and the USA and the gap between Saudi Arabia and the RoK. The results reveal a continuous and steady rise in the rankings of the Saudi EGDI over the years. However, the results also indicate that some areas require more improvement. An information management strategy for the support of e-government in Saudi Arabia is therefore proposed.

Given the current situation of the Coronavirus pandemic, the role of digital transformation and e-government in both developed and developing countries is considered highly significant. Academic research that comparatively evaluates the development of e-government longitudinally is still limited. This research paper is significant in that it extends the existing literature on the development of e-government by adopting a cross-country, longitudinal analysis, based on EGDI indicators of UN e-government reports during the period 2003-2020, to assess the overall progress of e-government in three countries: Saudi Arabia, the USA, and the RoK. It should be noted that the results of this study are based on specific data analysis techniques, and thus interpretations of these results should be considered in the context of such techniques. Nonetheless, the current study is significant in that it is one of the few studies that not only evaluates and compares the evolution of e-government in Saudi Arabia with that of the USA and the RoK, but also presents an information management strategy that prioritizes a number of areas that need to be considered to further develop e-government in Saudi Arabia.

The findings of the current study can be used by policymakers and IT managers in Saudi Arabia not only to focus on areas that need further consideration and development, but also to monitor and evaluate each step made that would enable the full realization of the benefits of e-government. The model proposed in the study can be thus used by e-government policymakers to consider the main factors that need to be taken into account during the continual process of developing and implementing e-government.

References

[1] Gharibi, W.; Khan, M.A. E-government in the Arab world: Analysis and perspective. In Proceedings of the 2014 World Congress on Computer Applications and Information Systems (WCCAIS), Hammamet, Tunisia: Curran Associates, Inc.: New York, USA. 2014.

[2] Ndou, V. E – government for developing countries: opportunities and challenges. The Electronic Journal on Information Systems in Developing Countries, Vol 18, 1-24, 2004.

[3] Fidel et. al The Case for e-government: Excerpts from the OECD Report: the e-government Imperative. OECD Journal on Budgeting, Vol 3, pp. 61–96, 2003.

[4] United nations department for economic and social affairs. United nations E-government survey 2018.; United nations: New York, 2019.

[5] United nations department for economic and social affairs. United nations E-government survey 2020.; United nations: New York, 2020.

[6] Alshehri, M.; Drew, S. Challenges of e-government Service Adoption in Saudi Arabia from e-Ready Citizen Perspective. IADIS International Conference ICT, Society and Human Beings, Freiburg, Germany: IADIS Press, Lisbon, Portugal, 2010.

[7] Benedetti, M.; Ghezzi, C.; Russo, C.; Lamberti, L. Towards a user-centric e-government service design: evidences from Italy. In A.R. Shark e Toporkoff, S. (eds.) Beyond e –Government Measuring Performance: A global perspective: BookSurge Publishing, 2009.

[8] Lamberti, L.; Benedetti, M.; Chen, S. Benefits sought by citizens and channel attitudes for multichannel payment services: Evidence from Italy. Government Information Quarterly, Vol 31, pp. 596–609, 2014.

[9] Detlor, B.; Hupfer, M.E.; Ruhi, U.; Zhao, L. Information quality and community municipal portal use. Government Information Quarterly, Vol 30, pp. 23–32, 2013.

[10] Arendsen, R.; Peters, O.; ter Hedde, M.; van Dijk, J. Does e-government reduce the administrative burden of businesses? An assessment of business-to-government systems usage in the Netherlands. Government Information Quarterly, Vol 31, pp. 160–169, 2014.

[11] Mundy, D.; Musa, B. Towards a Framework for eGovernment Development in Nigeria. Electronic Journal of e-government, Vol 8, pp. 148-166, 2010.

[12] Twizeyimana, J.D.; Andersson, A. The public value of E-Government – A literature review. Government Information Quarterly, Vol 36, pp. 167–178, 2019.

[13] Heeks, R. Most eGovernment-for-Development Projects Fail: How Can Risks be Reduced? iGovernment working paper number 14, Manchester: The University of Manchester, 2003.

[14] Qasem, M.H.; Elkadi, H.K.; Ghoneim, S.G. E-Government in Arab Countries: challenges and Evaluation. Journal of Computer Engineering, Vol 20, pp. 1-11, 2018.

[15] Lau, E.M. 5 Th Global Forum on Reinventing Government Mexico City, 5 November 2003 Challenges for E-government Development Oecd E-government Project Organisation for Economic Co-operation and Development, 2003.

[16] Al-Busadiy, M.; Weerakkody, V. E-government services in Oman: An employee's perspective. Electronic Government (EG), Vol 8, pp. 185-207, 2011.

[17] General Authority for Statistics in Saudi Arabia offical website. Available online: https://www.stats.gov.sa/en (accessed on 1st Oct 2020).

[18] Almufarrij, T. The Kingdom's annual population growth rate. Saudi Gazette Newspaper report, April 2, 2018.

[19] The Ministry of Communication and Information Technology in Saudi Arabia. Digital government strategy 2012. Available online: https://www.yesser.gov.sa/en/for-government/digital-government-strategy-2012 (accessed on 2nd Oct 2020).

[20] The Ministry of Communication and Information Technology in Saudi Arabia. E-government Program (Yesser). Available online: https://www.my.gov.sa/wps/portal/snp/pages/agencies/agencyDetails/AC319 (accessed on 1st Oct 2020).

[21] The Ministry of Communication and Information Technology in Saudi Arabia offical website. Available online: https://www.mcit.gov.sa/en/page/175492 (accessed on 2nd Oct 2020).

[22] Omari, A. Technology Adoption in the Arabian Gulf Countries: The Case of E-Government. IJCSEIT, Vol 3, pp. 1-8, 2013.

[23] Alghaith, W.; Sanzogni, L.; Sandhu, K. Factors Influencing the Adoption and Usage of Online Services in Saudi Arabia. EJISDC, Vol 40, pp. 1-32, 2010.

[24] Borras, J. International Technical Standards for E-Government. Electronic Journal of E-Government, Vol 2, pp. 75-80, 2004.

[25] United nations department for economic and social affairs. United nations E-government survey 2016; United nations: New York, 2016.

[26] Alshomrani, S. A Comparative Study on United Nations E-Government Indicators between Saudi Arabia and USA. Journal of Emerging Trends in Computing and Information Sciences, Vol 3, pp. 411-420, 2012.

APPENDIX A: E-GOVERNMENT INDICATORS AS DEFINED IN THE UN E-GOVERNMENT REPORTS OF 2016 AND 2020

- Online services index (OSI): A main indicator of the EGDI that evaluates the scope, nature, and quality of online e-government services.
- OSI1: stage1: Provision of basic e-government services.
- OSI2: stage2: Provision of greater information resources and services.
- OSI3: stage3: Provision of a two-way services approach between citizens and government agencies.
- OSI4: stage4: Advanced e-government services that connect government to government, government to citizens, and citizens to government.
- Telecommunication infrastructure index (TII): A main indicator of the EGDI that evaluates the status and quality of the telecommunication infrastructures for enabling the provision and delivery of online e-government services.
- Human capital index (HCI): A main indicator of the EGDI that evaluates the status and levels of human capital that relate to a particular country.

The Automatic Agricultural Crop Maintenance System using Runway Scheduling Algorithm: Fuzzyc-LR for IoT Networks

G. Balakrishna[1], Nageswara Rao Moparthi [2]
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation, India

***Abstract*—In this framework, the crop diseases have been identified using three types of methods, fuzzy-c as a clustering algorithm, runway scheduling trains like classification algorithm, and logistic regression as prediction algorithm. These techniques are meaningful solutions for losses in yields and the quantity of agriculture production. In this work, crop disease and corresponding fertilizers are predicted based on pattern scalability by the above algorithms. It proposes a Sensor Calibration and Feed Back Method (SCFM) with RWSA for better agriculture crop maintenance with automation and Fuzzy-c, Logistic regressions are helpful in studying the datasets of the crops for classifying the disease. This research tries to identify the leaf color, leaf size, disease of plant, and fertilizer for the illness of crops. In this context, RWSA-Agriculture gives the solution for current problems and improves the F1-Score. The data collected from local sensors and remote station is estimated with the dataset, these sensor based L.R., and Fuzzy-c controls disease prediction system in SCFM and RWSA. This technique accurately regulates the dispensing of water as well as chemicals; fertilizers for crop monitor and prevent the diseases of crops. This investigation gives performance metrics values i.e PSNR=44.18dB, SSIM = 0.9943, BPP =1.46, Tp=0.945 and CR = 5.25.**

Keywords*—Runway Scheduling Algorithm (RWSA); Sensor Calibration and Feedback Method (SCFM); IoT; fuzzy-c; logistic regression (LR)

I. INTRODUCTION

This framework introduced an IoT based sensor calibration technique with the Runway Scheduling algorithm. Agriculture is the leading and emerging subject everywhere, and automation is compulsory. The population is increasing at a fast rate, so food gradients are necessary for people. The conventional agriculture method serves intensified results because Health monitoring of crops and disease detection is a critical task. Because of this reason, increasing the in-depth research regarding plants for the rising of the production rate is the need of the hour, which can be changed with the help of traditional methods like IoT, Machine Learning, and Artificial Intelligence. Until we cover edge computing, manual machine correlation, and static threshold methods [1][2], etc. these methods give satisfactory results, but improvement is a need for disease identification. Crop maintenance and disease detection for a plant is a crucial task in agriculture; it is complex to recognize the plant disorders manually. Therefore it requires tremendous work and maintenance, along with more processing time, export suggestions. Hence IoT based agriculture disease monitoring is used for the recognition of plant diseases for image processing and future extraction, classification. This work discusses various methods of identifying plant diseases selecting their leaves as images and applies 3-algorithms. This work also discussed some clustering and classification algorithms used in the plant disease monitoring system. AgriTech, for instance, relates to the overall use of innovation in agriculture, another side, smart agriculture is frequently used to indicate the use of IoT alternatives in agriculture. The identical refers to the definition of intelligent farming. IoT technologies also can transform several agricultural aspects. In other phrases, there are more than five methods in which IoT can improve agriculture: data, tons of details collected by smart farming sensors, e.g., weather conditions, soil quality, the advancement of crop growth, or cattle health. This information could be used to monitor your crop's specific status as well as personnel appearance, machinery effectiveness, etc. India is a cultivated country of the United States, and about 70% of the population depends on agriculture [25]. Farmers have a significant kind of variety for choosing various suitable plants and finding first-class pesticides for the plant. Disease on plant results in a sizeable reduction in the tremendous amount of agricultural merchandise. The research of plant disorder complements with the studies of visually observable styles on the flowers. Monitoring of fitness and sickness on plant plays an essential characteristic in cultivation of plant life inside the farm. In the early days, the monitoring and evaluation of plant illnesses have been performed manually utilizing the know-how person in that subject. This calls for first-rate quantity of work and calls for excessive processing time. The image processing techniques may be used inside the plant sickness detection. In most of the instances, ailment signs are visible at the leaves, stem, and fruit. The plant leaf for the detection of ailment is considered, which indicates the sickness signs. This paper offers the advent of the image processing method used for plant disorder detection.

Fig. 1 explains about generalized IoT architecture framework. In this investigation, this work is to be taken as an example. Various sensor modules collect the information and send it back to respective modules. So sensor measurements and estimations are compulsory. This estimation of the process is to perform with the SCFM mechanism.

Fig. 1. General IoT Network.

A. Automation of Crops Condition

This development and changing desires need to be utilize IoT of an agriculture industry that is going through labor shortages and increasing prices for farm work. Many farmers are finding it difficult looking to produce other, higher-outstanding crops and finding the workers to plant, hold, and harvest one's plants. This anxiety isn't new in agriculture. For all of the human records, growing agricultural cultivation has been a function of either consisting of greater laborers or finding more green tools to do the system [27]. Modern agriculture is no tremendous. In the face of labor shortages, farmers are turning to technology to make farms extra inexperienced and automate the crop manufacturing cycle. The growing interest in technology and automation is evident in venture capital investments for ag-tech startups. In 2007, in step with TechCrunch, investors contributed US$200 million to entire of 31 ag-tech startups. Ten years later, in 2017, total investment rose to over US$1. Five billion dollors are split amongst 160 startups. These startups are addressing each element of the agriculture rate chain.

Others create a software program to govern the seed, soil, fertilizer, and irrigation, and make predictions approximately on timing and yield. Some startups use drones to show conditions remotely and even comply with fertilizers, insecticides, and other treatments from above. A developing cohort of businesses is operating on agricultural robotics to construct self-sufficient tractors, combines, and even fruit and vegetable deciding on robots. Before you may automate farm operations, however, you need accurate information approximately on the state of the farm. We also need a way for self-maintaining gadgets to connect; this is the realm of the Internet of Things (IoT). IoT devices are the sensors, gauges, and machines that are connected for the duration of a farm, the use of Bluetooth, a cellular network, or a few one-of-a-kind forms of connection. More IoT devices permit growers to build up more records approximately the dominion of their farms, and IoT is showing high-quality promise for optimizing useful resource delivery and driving precision agriculture to collect maximum performance

B. Crop Monitoring System

In this section, an IoT based agriculture, cultivation, automatic cropping system has been discussed. Because of this, the accuracy of the cropping has been increased for various climatic conditions such as temperature, rain, and cold, respectively. According to this, water supply and fertilizers have to be supplied to crops. This automation continuously monitors the LCD screens. When various diseases occur, then immediately finding the diseases and suggesting particular fertilizers has been performing quickly. For crop monitoring RGB values of a leaf are required.

RGB plant colour estimation by using below eq. (1).

$$f(x) = 0.2989xR + 0.5870xG + 0.114.xB \quad (1)$$

II. Parallel Research

This innovation relates as a rule to a framework for computerized control and all the more explicitly to a structure for checking and overseeing crop development. Farming has been a significant part of human presence for a long time. Upgrades in thinking about yields, quickening crop development, guaranteeing the nature of harvests, and accommodating a copious and productive collect have kept on adding to the pleasure and improvement of our populace's satisfaction. Significant zones for the robotization of farming incorporate water system, security against climate, creepy crawlies, and infection, and accommodating plant nourishment. Likewise, it is critical to have the option to conjecture crop development and collect with the goal that the financial aspects of reaping and appropriation can be increasingly productive. One case of a sort of yield that has profited incredibly from slow drifts in computerized agribusiness is the grape, which proves to be fruitfully used to make wine. The present vineyards incorporate distinctive administering frameworks for giving water to yields to a water system. Instances of such structures are "trickle" or "sprinkler" frameworks where water is steered among lines of vines by a cylinder having emanating gaps dispersed at ordinary interims. The water stream can be turned on or off physically or can be mechanized with clock control, P.C., and so on. The cylinders can be raised over the ground, or at or subterranean level.

The situation of diminishing water tables, evaporating of streams, and tanks, the capricious condition introduces an earnest requirement for the legitimate use of water. To adapt up to this utilization of temperature and dampness sensors at appropriate areas for checking of yields is actualized in [8]. A scheming formed through edge estimations of temperature also soil dampness can be modified hooked on a microcontroller-based door to regulator water quantity. Photovoltaic sheets that can organize the outline also can have a duplex correspondence connection dependent on cell – Internet interface that permits data review and water system planning to be personalized over a web page [9]. The mechanical improvement in open source programming and equipment make it simple to build up the gadget which can improve observing and remote sensor system made it conceivable to use in checking and control of nursery parameter inaccuracy agriculture [7].

In papers [2][3][4] projected a rural utilization of remote sensor organize for yield field checking. These frameworks wholly furnished with two sort sensor hubs to quantify dampness, temperature, also a picture for detecting the center to think about data by taking pictures of yields. Parameters assume a significant job in settling on a decent necessary

leadership for stable return inside a period. The limitations are temperature, mugginess, and pictures. By subsequent, these techniques can accomplish great soundness of sensors through low utilization of intensity. With it, is an extensive stretch of checking the agribusiness field region. Author in [5] anticipated a nursery Monitoring System dependent on agribusiness IoT among a cloud. In a nursery, the board can screen diverse ecological parameters viably utilizing sensor gadgets, for example, light sensor, temperature sensor, relative mugginess sensor, and soil dampness sensor. Occasionally (30 seconds), the sensors gather data of agribusiness field zone and are actuality logged then put away web-based utilizing distributed also calculating the Internet of Things. [6][13][16] Documents clarify an IOT Based Crop-Field Monitoring along with Irrigation Automation framework.

In their effort to screen crop-field, a framework is created through utilizing sensors as well as indicated by choice starting a server dependent on detected information, the water scheme framework computerized by using remote broadcast, the detected information sent in the direction of a web server database. On the off chance that the water system is mechanized, at that point, that implies if the dampness of temperature fields drop beneath of the probable territory. The client container screen regulate the framework remotely through the assistance of a submission, which gives a web interface to the client. In [7] a keen dribble water organization framework is planned. In this, a versatile Android application is utilized to decrease the inclusion of humans. Also, it is used to the regulator to screen the yield region remotely. Water depletion could be reduced through the Drip Irrigation framework as it works depending on data commencing water level sensors. Selected progressively, various situated sensors are utilized to screen the earth. The field climate information gathered and sensed together with weather information from internet repositories can be used to make several efficient choices to increase crop output. If the environmental condition is warm, dry, sunny, windy, then plants require a large quantity of water, and if these variables are like a cold, wet, cloudy, low wind, then the crop needs less water. The previous research model abstracted a scheme consisting of six components that are monitoring, managing, planning, distributing information, supporting the decision, and tracking action [22].

The practical and perfect quantity of agriculture manufacturing, storing, and exporting has been significantly enhanced by the IoT cloud service system by farmers. In this research, we are presenting various architectures of IoT multi-layer platforms for the agriculture sector by using IoT technologies [25-26]. This research contributes to significant suggestions for agriculture with developing countries [23].

A sensible model for automatic farming is suggested by investigating the layer IoT model. Before that, let us identify the general construction of IoT. Establishing numerous physical gadgets by and by IoT permanently consumes a 3-layer structure. The primary layer is the incorporated request layers, which in horticulture associated requests work because it is deliberating as a U.I. Layer. It is an agriculture client, and it includes rancher's mobile applications, and individual gadgets happen to screen the farming region. As per this layer, the farmers can take a choice to secure their harvest as more and improve sustenance creation yield. The subsequent layer is the data board layer, which contains a few obligations like arrangement and grouping of information, making, checking, essential leadership, and so forth. These jobs keep up also accomplished in this layer. The 3rd layer is a system executive's layer which speaks to the correspondence innovations like Gateway, RFID, GSM, Wifi, 3G, UMTS, as well as Bluetooth Low Energy.

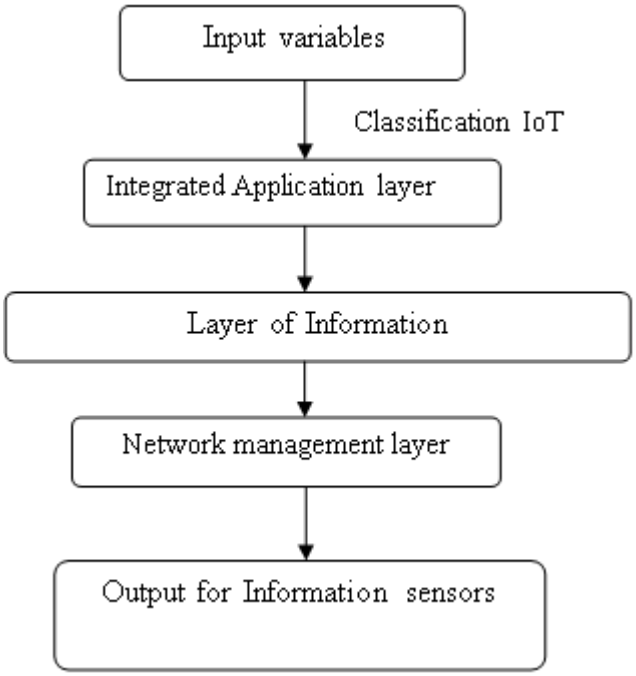


Fig. 2. Conventional Layer Model.

Zigbee and so forth. The fourth layer is a data accumulation layer that comprises a wide range of sensors, cameras, and so on. They are utilized to gather the data of harvest for enhanced in addition to simple field checking of horticulture zone. Fig. 2 demonstrates the four-layer IoT structure. But maintenance is more complex, and the energy consumption of IoT networks also increases, so move to propose a FUZZY-C RWSA-SCFM with L.R. machine-learning model is suggested.

III. METHODOLOGY

The significant objective of SCFM-RWSA is to develop and monitor various tasks of agricultural IoT systems and corresponding visualization sensors estimation system. In this system, IoT information is decelerated to SCFM. This technology has implemented and collected the multiple sensors information by using runway scheduling algorithm. SCFM gives the crop images-data to sensors visually. This representation can help for fast and accurate comparison with dataset classification and belongs to proposed agriculture systems. In this research, the Fuzzy-c model is to be used for clustering the dataset and classification perform with the Logistic regression model (Fig. 3).

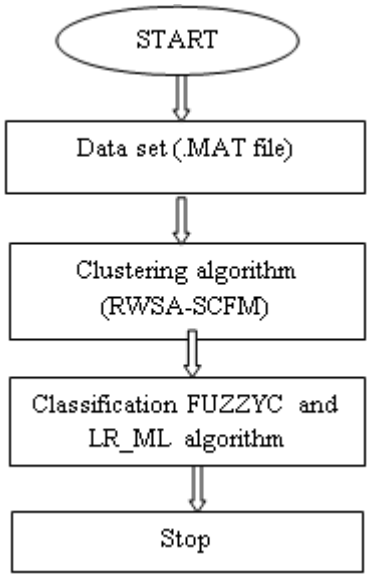


Fig. 3. Demonstrate the Complete block Diagram of the Proposed Method.

Here dataset and input images information is collected from various technical fields, and these are submitted with RWSA, FUZZY, and L.R. algorithms to achieve more throughput at the disease of crops monitoring using IoT.

A. RWSA (Runway Scheduling using SCFM)

1) Examples of iThings

Create two functions get smart (index)

Function 1 retrieving electricity plugs and sensors data frames

Index zero_fan, humidifier, and bulb

……
……....
…….
{

Set plugs (Boolean state, index)

}

Function 2 controlling smart plugs 2

Index zero_Fan, humidifier, and blub

State explanation (1: on, 0:off)

……
……....
…….
{

Stop
}

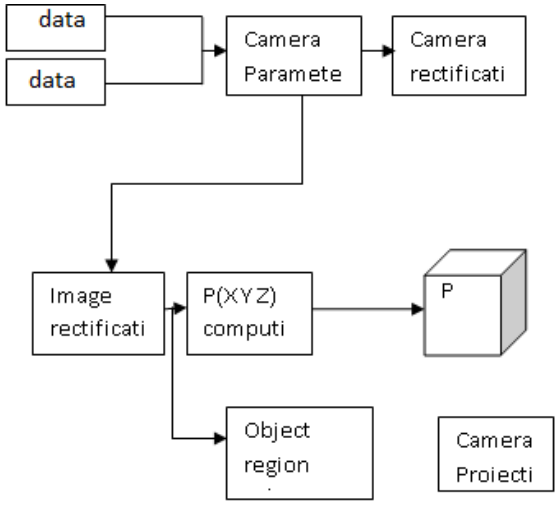


Fig. 4. RWSA-SCFM.

Fig. 4 demonstrate that the RWSA_SCFM-3D model, using this model image has been observed in three-dimensions such that 3600 monitoring is performed. In the above block diagram, the left side part demonstrates IoT things, which are connected to various electronic devices and sensors. The middle of the block diagram explains storage supporting things, which are real-time data storage devices. In front of the storage system, sensors have been divided into three stages named as offline processing stages, online processing stages, and graphical user interface unit (GUI). This GUI has been implemented by MATLAB 2015b software. Coming to the first stage, i.e. offline processing unit system, it is used for camera processing parameter estimation. In this, a relative camera position, angle, and tilting has adjusted for IoT instructions [10]. For example, if the location of the camera orientation varies concerning geometry, the following guidelines are generated in the below manner.

If camera 1 tilting > camera 2 tilt

Vary the camera 2 position

Else

Vary camera 1 position

Coming to the second stage online processing unit continually providing the data from sensors (X, Y, Z, coordinates). Using this data steps, one and three automatically updated the positions of the sensors. In the third step, the graphical Processing unit monitors the IoT information and using this information, 3D virtual instructions are automatically configured. According to this, all three phases are performed with the above function, respectively. The final right side of the block demonstrates that SCFM developed IoT information, as discussed all relates to SCFM implementation. In this case, all farmers continually and efficiently monitor the crops using sensors. This discussion is illustrated in the below section.

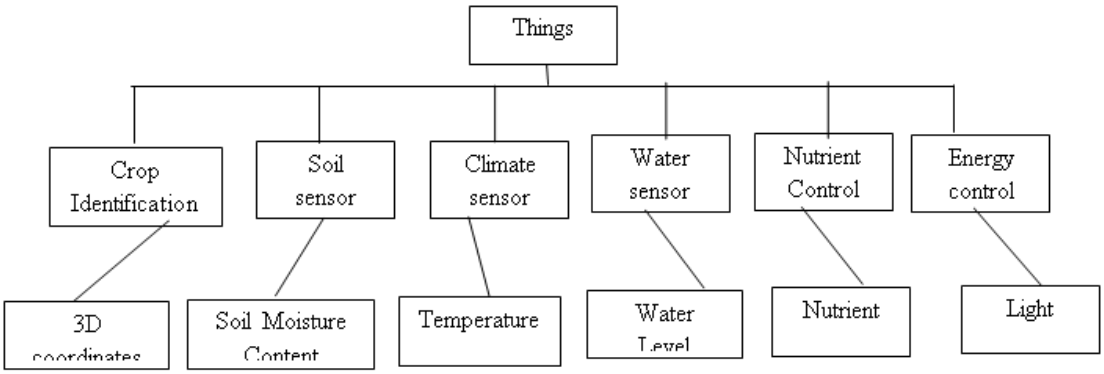


Fig. 5. 3D Parameters.

B. Parameters Related to 3D Communication

In this section, the usage of various IoT devices, and functionality has been explained. Various electronic devices like sensors, motors, and humidity sensors are arranged hierarchically. Further, agriculture systems like soil, crop, water, nutrients, and climatic conditions are associated with the following IoT crop monitoring system [6]. Various sensors provide some data such as crop identification, soil sensor, climate sensor, water sensor, nutrient control, and energy control system.

Fig. 5 explains that SCFM – IoT system of 3D parameters mechanism in this step by step crop monitoring System composed of a farm – crop mechanism. This model is illustrated below.

- region farm :farmer management
- region sensor : sensor monitoring action
- region plant : crops disease monitoring
- Region IoT: electronic sensor devises composition.

Above all steps are related to SCFM region-based modal. Hear all multiple cameras are adjusted with a particular position by using visualize the 3D coordinate system. Therefore all coordinates of crops are recognized by utilizing the horizontal visualization. In this investigation, IoT based crop monitoring with a multihull camera has been deployed to measure accuracy and efficiency. According to the wireless sensor system, this implementation has been developed and verified in [11] shown in Fig. 4.

C. Algorithm-Runway Scheduling Algorithm (RWSA)

Fig. 6 demonstrates the runway scheduling algorithm for crop disease prediction and classification. In this figure total, nine steps are utilized to classify the disease and suggest the proper fertilizer.

IoT's effective operation, and especially execution, is critical to the technology system as a whole throughput. Scheduling run-off arrivals and departures is a complicated issue that needs to tackle various and often competing factors of effectiveness, safety, and equity between any scenarios. One strategy to runway scheduling that emerges from operational and fairness considerations is the restricted position shift (CPS), which demands that an operating position of the device in the optimized sequence does not deviate considerably from its place in the first-come-first-served series.

RWSA rules, as shown in Fig. 6 are the rules for former, plant and sensing development to receive the storage data from the transmitter. The main advantage of RWSA is the data available in the storage devices if it does not appear than it retransmits the information from one node to another node. A simple example of RWSA is illustrated here, i.e., plant one is located at x-node, and plant two is located at y-node if the exchange of information has performed by using [12] shown in runway scheduling algorithm.

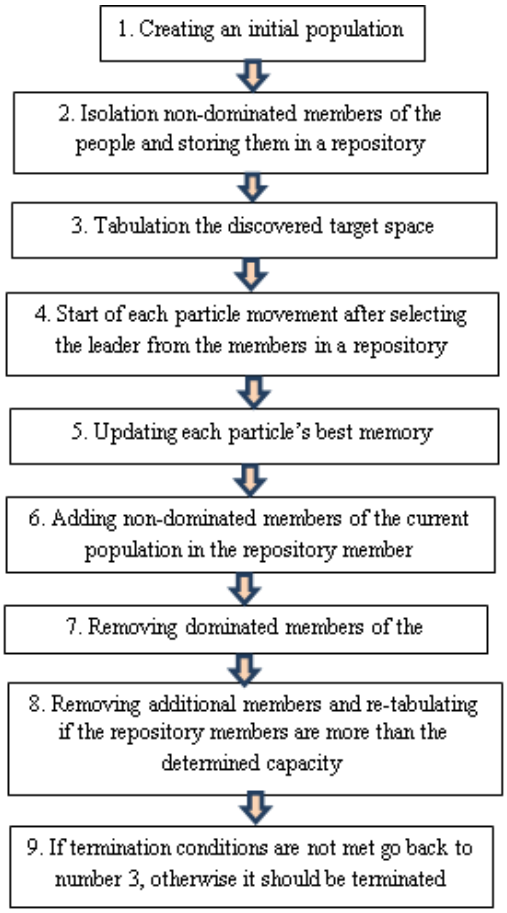


Fig. 6. Rwsa Algorithm.

D. IoT Agriculture Monitoring with Multi-Cameras

The IoT model, which is explained above, works based on below Fig. 10 principle. In this, x,y,z assessors are randomly adjusted by longitudinal directions of the screen. In this, parallel screens are utilized to continuously monitor the crops and leaves. Here, principle access changed the focal length, and optical access adjusted the lens diameter and image plane is customized by the quantum segment. Sensor camera calibration is correctly regulated by the RWSA algorithm. Hence, accurate coordinates are confined for capturing. So, evaluation and improvement of cameras automatically adjusted.

$$C = \begin{bmatrix} R \\ t \end{bmatrix} k \tag{2}$$

$$C = \begin{bmatrix} f_x & 0 & 0 \\ s & f_y & 0 \\ c_x & c_y & 1 \end{bmatrix} \tag{3}$$

Fx= x-axis focus, Fy= y –axis focus, Cx= center of x-axis point, Cy= center of y-axis point, C= rank of Eigen matrix.

Where stiff changes since 3D world organize to 3D camera coordinates are a rotation R and a conversion t, termed extrinsic parameters, and c represents the rank of the system. K is the intrinsic parameter representing a two-dimensional (2D) image coordinate projective transformation from 3D camera coordinates. The inherent matrix of the camera is described as equation 1 and 2 also in Fig. 7.

Fig. 8 explains sensor calibration flow chart, in 1st step preparation of data from camera is collected in pattern manner in addition to next step, calibration is processed with weight balancing method in the third step to evaluate the data set and improve the functionality of cameras & processing of IoT network[15].

E. Screening Analysis

Above all, the review contains four parts here. Identification of screening eligibility include the blocks depending upon the n no of samples find the records. Frame size is decided by sample availability and records exclusion.

F. Offline Preparation Stage

This offline processing stage consists of the superimposition of IoT data associated with physical objects. The examples of these objects are plants, leafs, and space; the view of cameras grabbing this digital data in an effective manner. In this constrain, the offline processing stage estimates the parameters of a camera with respect to the intrinsic and extrinsic composite behavior. The overall calculations of orientation image classifications rectify the disease by the RWSA algorithm among this orientation. Distorted images are adjusted by automatic lengths rectification system. However, to estimate the condition by the calibration method, which is shown below. The essential information which is acquired from Fig. 8 gives the apparent functionality of screening analysis of data. In this total, 4 stages are screening the duplicate and original records based on included, eligibility, screening, and identification parameters.

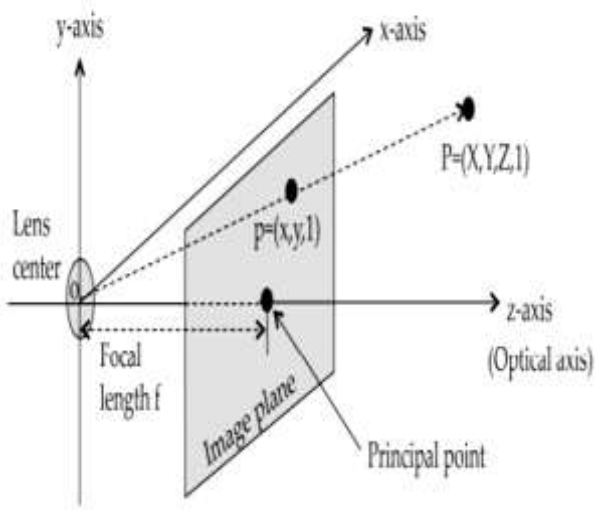


Fig. 7. Sensors of Camera Calibration.

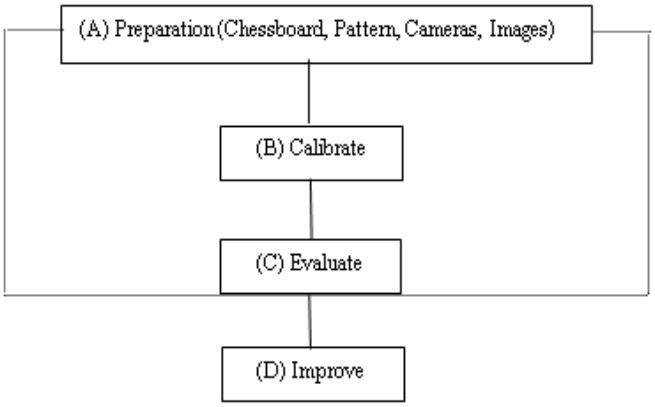


Fig. 8. Evaluation of Data.

The overall functionality is designed by a software called MATLAB 2015b, and cameras are calibrated with respect to their images clearly shown in Fig. 9.

Here sensing unit is an essential module; it can be searching signals continually from conditional blocks using eq.3 gives the ADC to application algorithm nothing but SCFM.

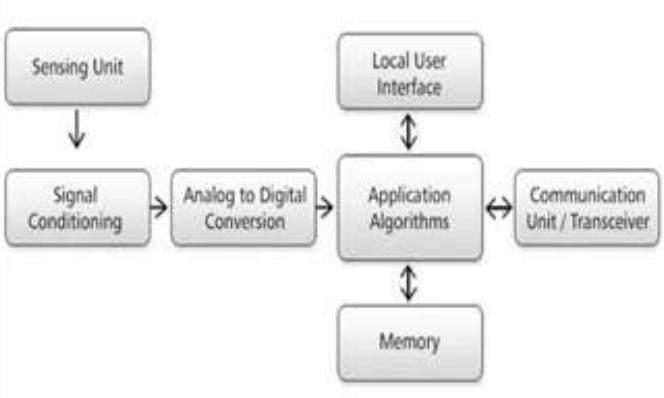


Fig. 9. IoT Data Evolution Process.

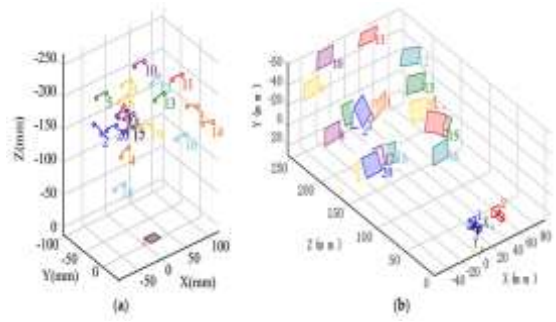


Fig. 10. Graphical Representation of Data.

Memory is a local storage. It can be a cloud or a storage unit and the final transceiver cloud be sending and receiving data continuously and calibrating the feedback if correct controlling is achieved. If SCFM is not reached to calibration, then it processes for feedback modeling, which is mentioned in equation 3 and shown in Fig. 10.

G. *Fuzzy-C Algorithm (Clustering of IoT –Agricultural Datasets)*

In this section, we are using the grouping of datasets that have been analyzed using a fuzzy-c algorithm. This algorithm gives a better accuracy point with the help of weight balanced load equalization system.

Fig. 11 demonstrates the fuzzy-c clustering model using the fuzzification technique. In this fuzzy logic involves phase adjustment, rule evaluation, and aggregation. Camera direction ROI (region of interest) is analyzed with calibration by proposed SCFM and clustering with the FUZZY-C method. At every step, this adjustment is observed clearly. In this phase, the input variables are represented as fuzzy sets with three values high, Medium, and low. Table I shows the Fuzzy load function for the input and output variables. The triangular fuzzy (trimf) set is used in our model.

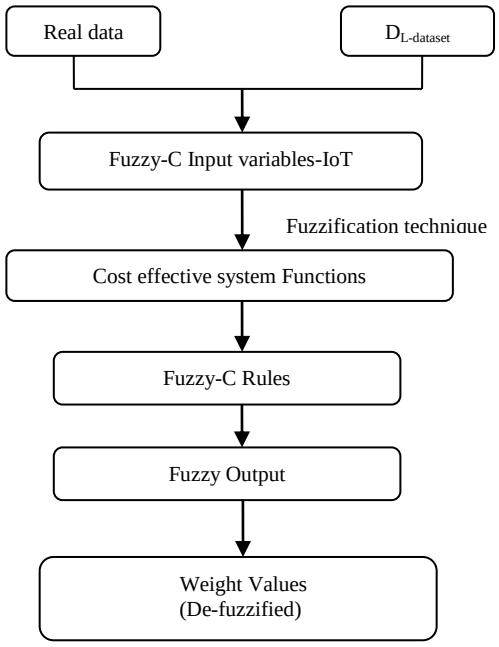


Fig. 11. Flow of FUZZY_C Model.

For Rp, Low = 0 to 3 weights, Medium = 2 to 5 weights, High=4 to 9 weights

For DZL, Low = 0 to 3 cluster, Medium = 2 to 5 cluster, High =4 to 7 cluster For w, Low = 0 to 3, Medium = 1 to 3, High = 2 to 5.

Ep= interphase fuzzy, DL= de-fuzzification

$$\text{fuzzy\_C}=[\sum Z_{\text{allrules}}\ f_i * \alpha(f_i)/\sum_{\text{all rules}} \alpha(f_i)\ ] \quad (4)$$

Zall= all rules of fuzzy logic, fi= current stage of function, $\alpha(f_i)$= coefficient of fuzzy-C

Table I demonstrates that When E .P. of a node is low, then it could not move ahead of the packets to Z.L. Therefore, it needs to be in a sleeping kingdom for a longer time. However, if the node is closer to Z.L., it has to be lively a piece bit earlier. Hence in rule no 2, 3, w is assigned as High, and in rule 1, it's far assigned Medium. If the E .P. is medium, then the w is assigned a Medium cost, regardless of the gap to Z.L. Hence rule no, 4, 5, and six, Medium price is assigned to w. Finally, if the E_P is excessive, then the node may be in the energetic node for a greater time. Hence in rule eight and 9, w is about as Low fee. However, if the space to Z.L. is less, then w rate is about to Medium, in rule 7.

TABLE I. DECISION USING FUZZY LOADS

Rule no.	E_P	D_L	weights value
01	High-1	Low-0	Medium-M
02	High-1	Medium-M	Low-0
03	High-1	High-1	Low-0
04	Low-0	Low-0	Medium-M
05	Low-0	Medium--M	High-1
06	Low-0	High-1	High-1
07	Medium-M	Low-0	Medium-M
08	Medium-M	Medium-M	Medium-M
09	Medium-M	High-1	Medium-M

Algorithm

1. source $S_i = 1,2, \ldots N$
2. intermediate node $N_j, j = 1,2 \ldots . K$
3. N_j estimates E_P and D_L
4. RE_P and D_L are passed as input variables to a Fuzzy –C model
5. Fuzzification is performed over the input and output variables
6. Fuzzy Rules are applied as per Table I, and fuzzy output is returned.
7. Estimation is performed, and the value w is returned.
8. Estimate the duty cycle of N_j based on the output w
9. N_j is put in sleep mode for the period of $Time_{sleep}$
10. End For
11. End For

This above FUZZY-C algorithm is applied on datasets. From this, we get information like similar elements and variables. Using of this method gives accurate clustering among real data and reference data.

H. Logistic Regression

LR (logistic regression) is a basic supervised machine learning model. This model is mainly using a classification of designs in every field. This L.R. is used for the classification of disease finding and fertilizer estimation purposes. In category trouble, the target variable (or output), y, can take the handiest discrete values for a given set of functions (or inputs), X. Contrary to well-known perception, logistic regression is a regression model.

Mathematical computations of LR

The hypothesis for linear regression is h (X) = θ 0 + θ1*X

$$h(X) = \frac{1}{1+e^{-(\theta o+\theta 1^{*}\times)}} \quad (5)$$

The hypothesis of linear function which is used for regression analysis

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 \quad (6)$$

$J(\theta_0, \theta_1)$ is the regression coefficent for LR

$$h(X) = \frac{1}{1+e^{-(\theta o+\theta 1^{*}\times)}} \quad (7)$$

$$h(X) = \theta 0 + \theta 1^{*} \times \quad (8)$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(h_\theta(x(i)), (y(i)) \quad (9)$$

$$= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}\log(1 - h_\theta(x^{(i)}))] \quad (10)$$

Eq 6 to 7 explains about linear functionality of tree classification for weight balancing.

$$P(y=1/x;\theta)=h_\theta(x) = \frac{1}{1+e^{-\theta T_x}} \quad (11)$$

Eq 10 explains about the inverse of hypothesis function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(h_\theta(x(i)), (y(i)) \quad (12)$$

$$-\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}\log(1 - h_\theta(x^{(i)}))] \quad (13)$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(h_\theta(x(i)), (y(i)) \quad (14)$$

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Note: y=0 or 1 always. Equation 12 to 15 demonstrate that cost-effective function classification and estimating the accurate pathology for disease prediction. Entire LR mathematical computations give classification of plant diseases. Based on this calculation the tp and efficiency. h(x)= LR classification factor.

Fig. 12 shows that 3D camera image training of land selected for testing, here wheat plants are taken into account for the experiment a) is the main camera image and b) is the rectified image for IoT devices, c) is for proposed selection by SCFM.

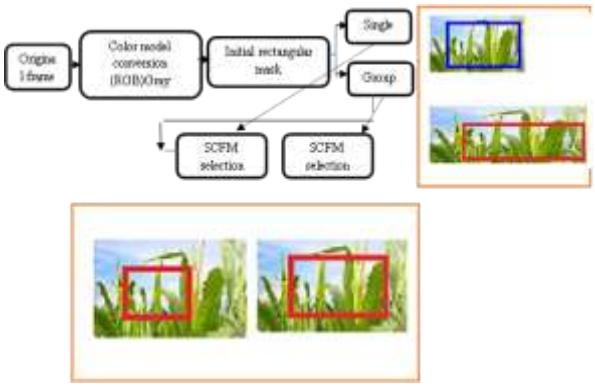


Fig. 12. Crop Diseases Estimation Model.

Fig. 12 is the original frame selected for data samples followed by color model RGB applying the rectangular mask for the region of selection SCFM is applied on trained IoT Network image [17]. The ROI (region of interest) is described by utilizing the rectangular masks (blue line).

$$ROI = [\, x\ y\ w\ h\,] \quad (14)$$

The rectangular region of interest is adjusted with x,y coordinates, and these are extended to height h and width w. the geometrical center of region explained by (red line).

$$ROI = [\, x^1\ y\ w\ h\,] \quad (15)$$

Commencing Sustainability these two points, as illustrated in Fig. 13.

Fig. 14 demonstrates the region of interest camera adjustment system by using 2 centers that are c1 and c2 the point x at camera 1 and x` at camera 2 are projected method [14]. To determine the p, p` objectives are inserted by backpropagation rays. The centers c1 and c2 are randomly modified by the coordinate system mechanism. In this mechanism, SCFM-RWSA optimization techniques help for tilting adjustment at classification step leaf disease, and corresponding fertilizer has identified. This explanation experimentally proved in MATLAB software.

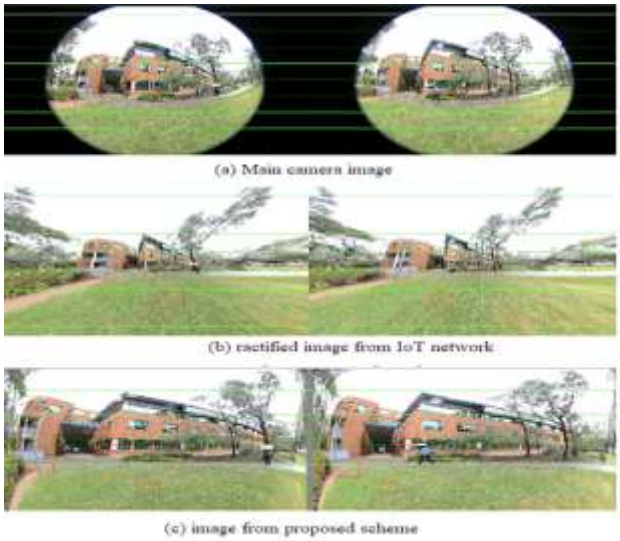


Fig. 13. 3D Camera Image Training.

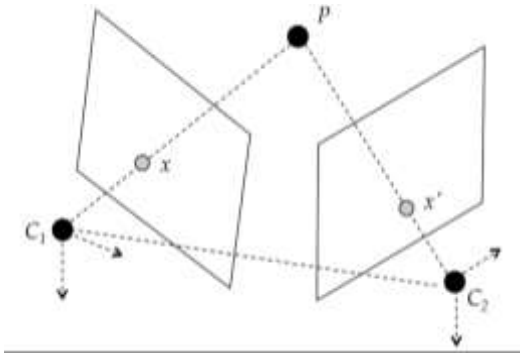


Fig. 14. Coordinate Management.

To validate the SCFM IoT preprocessing algorithm is used to calculate the disease diameter and type of disease. In this system, the RWSA classifier gives the advantage over IoT data by a real-time former crop protection system. This investigation explains that various trees like a palm tree, wheat, rice crops are taken as input and perform the two tasks of disease identification and classification. Final crop disease and our own fertilizer is the output of our research.

Graphical 3D rotation and final crop image outputs are observed by using Display and iteration methods with the help of camera projection. The graphical 3D view is balanced with the coordinate adjustment system [19,20].

IV. Result and Discussion

Using the existing method the proposed findings are compared, and also plant monitoring is done with day wise manner 1st day sowing the seeds 1-45 days watering process, 7 greenhouse and organic fertilizers adding 25 days for leaf color and leaf size analyzed. 45 day observe the results. Shown in Table II.

Fig. 15 demonstrates that every stage performance is analyzed with three algorithms. This is the best IoT based agriculture monitoring system for disease prediction. Here fertilizer type is estimated according to the corresponding disease.

Table III and Fig. 16 explains that manual, SCFM, manual SCFM comparison, n is the no of point to be analyzed mean function for average the probability, S.D. is the standard deviation t is time delay p is the distance using these parameters calculating the result which plant leaf is healthy, and color of leaf and all using this gives the fertilizers to plants [21] has a single main stem plus 2-3 tillers per plant. With better increasing circumstances and reduced plant density, the amount of farmers tends to rise. Filleting begins at the 3-4 point of the leaf, about when it is possible to see the first nodal roots. Final output is shown in Fig. 17 (also see Table IV).

Fig. 18(a) to (c) and 19(a) to (c) demonstrate that diseases finding outcome combination of RWSA-fuzzy-c and L.R. machine learning algorithms. Therefore crop diseases are easily identified, and corresponding fertilizers are suggested.

Dataset: This investigation Indiana pine data set is used for classification and preprocessing. This implementation is performed on Matlab code in this folder Accuracy_Data.mat, and below datasets are taken as reference

Dataset 1: https://www.quantitative-plant.org/dataset/plant-database

Dataset 2: http://helminen.co/plant-disease

Dataset 3:https://plantvillage.psu.edu/

Fertilizer: In this project Nitrogen Fertilizers are used as agents for anthracnose disease. Balanced amounts of plant vitamins, mainly nitrogen are also used. Desirable drainage of fields (in conventionally flooded vegetation) and nurseries is ensured. Fields are smooth. Weed hosts and plow below rice stubble, straw, rice ratoons, and volunteer seedlings which could serve as hosts of bacteria are removed.

TABLE II. Performance Estimation

NO OF DAYS	PROCESS
01	Seeds estimation
01-60	Watering process
15	Applying fertilizers
13,19,22,55	Organic fertilizers
35	Leaf color and size
55	Final results.

Fig. 15. Performance Analysis.

TABLE III. Comparison of Methods

Method	N	Mean	Standard deviation	TS	PS
Manual[10]	15	2.488	0.2672	-	-
Manual-SCFM[11]	15	2.1428	0.2343	28.181	<0.001
SCFM-Fuzzy[12]	15	2.2414	0.2313	29.121	<0.001
RWSA-FUZZY-LR	15	0.3470	0.04270	29.912	<0.001

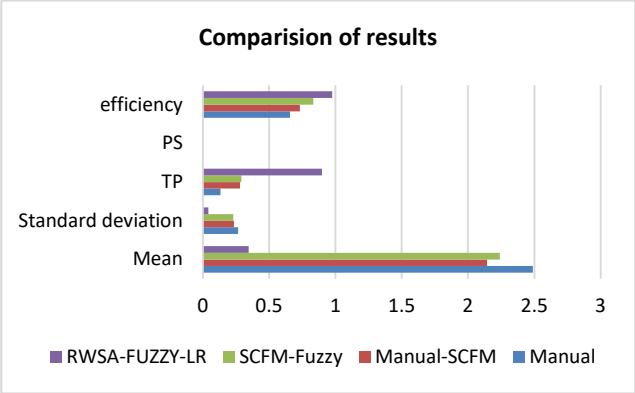
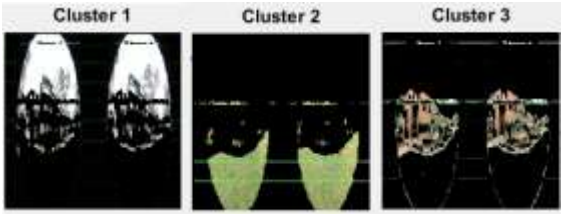


Fig. 16. Comparison of Work.

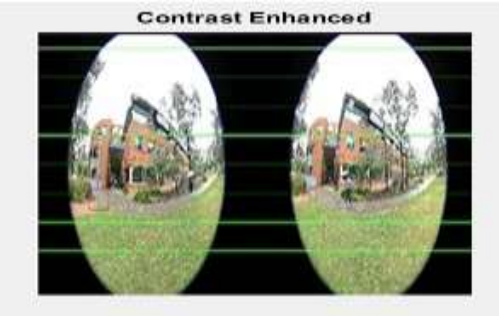
TABLE IV. T_P AND EFFICIENCY

Method	Mean	Standard deviation	T_P	PS	efficiency
Manual	2.488	0.2672	0.1342	<0.001	0.6567
Manual-SCFM	2.1428	0.2343	0.28181	<0.001	0.7312
SCFM-Fuzzy	2.2414	0.2313	0.29121	<0.001	0.8345
RWSA-FUZZY-LR	0.347	0.0427	0.92912	<0.001	0.9754

Fig. 17. Final Output.



(a)



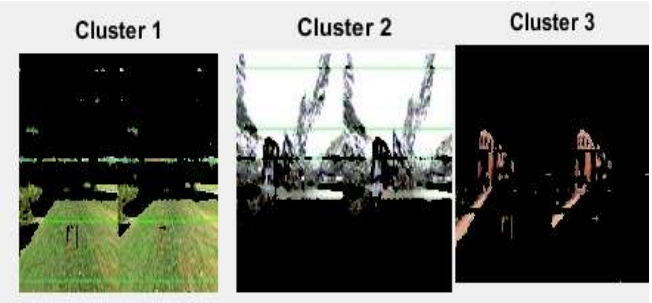
(b)

(c)

Fig. 18. (a) Cluster Finding Output of RWSA. (b) Input of our Proposed RWSA. (c) Diseases Finding.

Table VI and Fig. 20 explains that the comparison table between DWT and RWSA method here, all parameters are improved compared to the existed method [24].

Table V explains that results from the RWSA algorithm using this IoT technique finding an image of plant and diseases.



(a) Cluster Finding Output of RWSA.

(b)

Fig. 19. (a) Cluster Finding Output of RWSA. (b) Input of our Proposed RWSA (c) Diseases Finding.

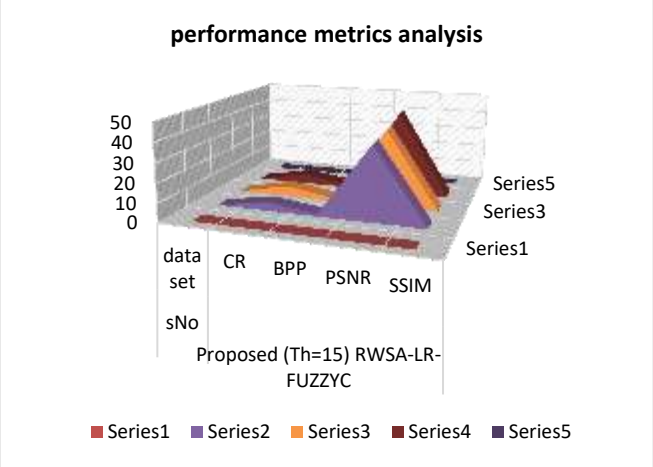


Fig. 20. Performance Matric Analysis.

TABLE V. RESULTS FROM ANALYSIS

Dataset	Threshold Value	CR	BPP	PSNR (db)	SSIM	Efficiency %
Dataset 1 https://www.quantitative-plant.org/dataset/plant-database	5	3.94	2.22	49.78	0.9997	58.34
	10	4.65	1.81	43.05	0.9970	73.33
	15	4.96	1.23	39.14	0.9961	79.85
	20	5.87	1.14	37.51	0.9859	85.96
	25	5.74	1.12	35.52	0.9851	88.26
Dataset 2 http://helminen.co/plant-disease	5	4.88	1.73	49.74	0.9975	72.72
	10	5.48	1.44	43.93	0.9968	86.13
	15	5.59	1.41	40.69	0.9957	89.48
	20	5.64	1.37	38.24	0.9906	90.57
	25	5.79	1.28	38.22	0.9887	92.23
Dataset 3 https://plantvillage.psu.edu/	5	4.79	1.85	49.91	0.9993	78.94
	10	4.89	1.71	45.12	0.9966	86.14
	15	5.25	1.46	44.18	0.9943	87.46
	20	5.78	1.42	40.27	0.9915	90.61
	25	5.89	1.37	39.99	0.9883	92.61

TABLE VI. COMPARISON WITH EXISTING VS. PROPOSED METHOD

Dataset	**GA Genetic algorithm**				**3-DWT (Th=15)**				**Proposed (Th=15) RWSA-LR-FUZZYC**			
	CR	**BPP**	**PSNR**	**SSIM**	**CR**	**BPP**	**PSNR**	**SSIM**	**CR**	**BPP**	**PSNR**	**SSIM**
Dataset 1	4.11	1.81	36.18	0.9940	4.64	1.72	37.18	0.9902	4.65	1.23	39.14	0.9961
Dataset 2	5.12	1.49	39.54	0.9941	5.59	1.46	40.68	0.9937	5.59	1.41	40.69	0.9957
Dataset 3	5.11	1.48	39.94	0.9942	5.25	1.50	40.73	0.9946	5.25	1.46	44.18	0.9943

V. CONCLUSION

In this investigation, a crop monitoring system with IoT platform has been designed, various methods have been implemented, but accuracy, disease finding with respect to fertilizers has not been designed. Therefore, a real and accurate IoT based agriculture monitoring system is necessary. In this work IoT crop monitoring system has been implemented for disease finding and classification model. Moreover, RWSA-FUZZYC and L.R. - IoT crop monitoring system achieved high accuracy and throughput, and this agriculture monitoring system gives better results for farmers. The threshold value 15 RWSA-SCFM method obtained PSNR=44.18dB, SSIM = 0.9943, BPP =1.46 and CR = 5.25. These results are more accurate compared to existed methods. So, the implemented method is better than conventional techniques.

REFERENCES

[1] Things: A Survey on Enabling Technologies, Protocols, and Applications," IEEE Commun. Surveys & Tutorials, vol. 17, no. 4, 2015, pp. 2347–76. c.

[2] P. G. Lopez , Alberto Montresor ,Dick Epema, AnwitamanDatta,TeruoHigashino, Adriana Iamnitchi, "Edge-Centric Computing: Vision and Challenges," SIGCOMM Comp. Commun. Rev. vol. 45, no. 5, 2015, pp. 37–42.

[3] T. G. Rodrigues, Hiroki Nishiyama, "Hybrid Method for Minimizing Service Delay in Edge Cloud Computing through V.M. Migration and Transmission Power Control," IEEE Trans. Computers, vol. 66, no. 5, 2017, pp. 810–19.

[4] J. Ren, HuiGuo, ChuguiXu, YaoxueZhang "Serving at the Edge: A Scalable IoT Architecture Based on Transparent Computing," IEEE Network, 2017. 5. ZHANG Yaoxue, REN Ju, LIU Jiagang, XU Chugui, GUO HuiAnd LIU Yaping"A Survey on Emerging Computing Paradigms for Big Data," Chinese J. Electronics, vol. 26, no. 1, 2017, pp. 1–12.

[5] JuRen, Yaoxue Zhang, Kuan Zhang, and Xuemin (Sherman) Shen"Exploiting Mobile Crowdsourcing for Pervasive Cloud Services: Challenges and Solutions," IEEE Commun. Mag., vol. 53, no. 3, Mar. 2015, pp. 98–105.

[6] Popovi´c, T.; Latinovi´c, N.; Peši´c, A.; Ze˘cevi´c, Ž.; Krstaji´c, B.; Djukanovi´c, S. Architecting an IoT-enabled platform for precision agriculture and ecological monitoring: A case study. Comput. Electron. Agric. 2017, 140, 255–265. [CrossRef].

[7] Huang, J.M.; Ong, S.K.; Nee, A.Y.C. Real-time finite element structural analysis in augmented reality. Advances in Engineering Software. Adv. Eng. Softw. 2015, 87, 43–56. [CrossRef].

[8] Daponte, P.; Vito, L.D.; Picariello, F.; Riccio, M. State of the art and future developments of the Augmented Reality for measurement applications. Measurement 2014, 57, 53–70. [CrossRef].

[9] C˘ olakovic´, A.; Hadžialic´,M. Internet of Things (IoT): A review of enabling technologies, challenges, and open research issues. Comput. Netw. 2018, 144, 17–39. [CrossRef].

[10] Chuang, C.L.; Yang, E.C.; Tseng, C.L.; Chen, C.P.; Lien, G.S.; Jiang, J.A. Toward anticipating pest responses to fruit farms: Revealing factors influencing the population dynamics of the Oriental Fruit Fly via automatic field monitoring. Comput. Electron. Agric. 2014, 109, 148–161. [CrossRef].

[11] Yang, F.; Wang, K.; Han, Y.; Qiao, Z. A Cloud-Based Digital Farm Management System for Vegetable Production Process Management and Quality Traceability. Sustainability, 2018, 10, 4007. [CrossRef].

[12] Kamilaris, A.; Gao, F.; Prenafeta-Boldu, F.X.; Ali, M.I. Agri-IoT: A semantic framework for Internet of Things-enabled smart farming applications. In Proceedings of the IEEE 3rd World Forum on Internet of Things (WF-IoT), Reston, VA, USA, 12–14 December 2016.

[13] Liao, M.S.; Chen, S.F.; Chou, C.Y.; Chen, H.Y.; Yeh, S.H.; Chang, Y.C.; Jiang, J.A. On precisely relating the growth of Phalaenopsis leaves to greenhouse environmental factors by using an IoT-based monitoring system. Comput. Electron. Agric. 2017, 136, 125–139. [CrossRef].

[14] Ferrández-Pastor, F.J.; García-Chamizo, J.M.; Nieto-Hidalgo, M.; Mora-Pascual, J.; Mora-Martínez, J. Developing Ubiquitous Sensor Network Platform Using Internet of Things: Application in Precision Agriculture. Sensors, 2016, 16, 1141. [CrossRef].

[15] Murphy, F.E.; Magno, M.; O'Leary, L.; Troy, K.; Whelan, P.; Popovici, E.M. Big Brother for Bees (3B)—Energy Neutral Platform for Remote Monitoring of Beehive Imagery and Sound. In Proceedings of the 6th International Workshop on Advances in Sensors and Interfaces (IWASI), Gallipoli, Italy, 18–19 June 2015.

[16] Díaz, M.; Martín, C.; Rubio, B. State-of-the-art, challenges, and open issues in the integration of Internet of things and cloud computing. J. Netw. Comput. Appl. 2016, 67, 99–117. [CrossRef].

[17] Tati´c, D.; Teši´c, B. The application of augmented reality technologies for the improvement of occupational safety in an industrial environment. Comput. Ind. 2017, 85, 1–10. [CrossRef].

[18] Velázquez, F.; Morales Méndez, G. Augmented Reality and Mobile Devices: A Binominal Methodological Resource for Inclusive Education (SDG 4). An Example in Secondary Education. Sustainability, 2018, 10, 3446. 1. [CrossRef].

[19] ElSayed, N.A.M.; Thomas, B.H.; Marriott, K.; Piantadosi, J.; Smith, R.T. Situated Analytics: Demonstrating immersive analytical tools with Augmented Reality. J. Vis. Lang. Comput. 2016, 36, 13–23. [CrossRef].

[20] Rashid, Z.; Melià-Seguí, J.; Pous, R.; Peig, E. Using Augmented Reality and Internet of Things to improve accessibility of people with motor disabilities in the context of Smart Cities. Future Gener. Comput. Syst. 2017, 76, 248–261. [CrossRef].

[21] Saikumar, K & Rajesh, V &Ramya, N &Shaik, Hasane& Santosh Kumar, Gurram. (2019). A deep learning process for spine and heart segmentation using pixel-based convolutional networks. Journal of International Pharmaceutical Research. 46. 1674-0440.

[22] Sushanth, G., &Sujatha, S. (2018). IoT Based Smart Agriculture System. 2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). doi:10.1109/wispnet. 2018.8538702.

[23] Abdullah Na, William Isaac, "Developing a human-centric agricultural model in the IoT environment," in 2016 International Conference on Internet of Things and Applications (IOTA) Maharashtra Institute of Technology, Pune, India 22 Jan - 24 Jan, 2016, 978-1-5090-0044-9/16, 2016 IEEE.

[24] Boopathi raja S Kalavathi Palanisamy "A Near Lossless Multispectral Image Compression using 3D-DWT with application to LANDSAT Images" International Journal of Computer Sciences and Engineering, 2017.

[25] G.Balakrishna and Moparthi Nageshwara Rao," ESBL: Design and Implement A Cloud Integrated Framework for IoT Load Balancing" International Journal Of Computers Communications & Control ISSN 1841-9836, e-ISSN 1841-9844, 14(4), 459-474, August 2019.

[26] G.Balakrishna and Moparthi Nageshwara Rao," Study Report on Using IoT Agriculture Farm Monitoring", Innovations in Computer Science and Engineering, Lecture Notes in Networks and Systems 74,https://doi.org/10.1007/978-981-13-7082-3_55.

[27] G.Balakrishna and Moparthi Nageshwara Rao " Study report on Indian agriculture with IoT "International Journal of Electrical and Computer Engineering.

Legal Requirements towards Enhancing the Security of Medical Devices

Prosper K. Yeng[1]
Department of Information Security
and Communication Technology
NTNU
Gjøvik, Norway

Stephen D. Wulthusen[2]
Department of Information Security
and Communication Technology
NTNU
Gjøvik, Norway
School of Mathematics
and Information Security
Royal Holloway, University of London
Egham, United Kingdom

Bian Yang[3]
Department of Information Security
and Communication Technology
NTNU
Gjøvik, Norway

***Abstract*—Over 25 million Americans are dependent on medical devices. However, the patients who need these devices only have two choices, thus the choice between using an insecure critical-life-functioning devices or the choice to live without the support of a medical device with the consequences of the threats presented by the disease. This study therefore conducted a state-of-the-art on security requirements, concerning medical devices in the US and EU. Food, Drugs and Cosmetic Act, HIPAA, Medical Device Regulations of EU and GDPR were some of the identified regulations for controlling the security of these devices. Statutory laws such as computer Fraud and abuse Act (CFAA), Anti-Tempering Act, Panel Code as well as Battery and Trespass to Chattel in the civil law, were also identified. In analyzing the security requirements, there are less motivations on criminal charges against cyber criminals in addressing the security issues. Because it is often challenging to identify the culprits in medical device hacks. It is also difficult to hold device manufactures on negligence of duty especially after the device has been approved or if the harm on patient was as a result of a cyber attacker. Suggestions have been provided to improve upon the regulations so that both the regulatory bodies and MDM can improve upon their security conscious care.**

Keywords*—*Information security; medical device; legal requirement; healthcare; privacy

I. Introduction

Medical devices play significant role in the sustenance of human life in our society. In addition, the connection of these devices to the internet has transformed the medical device management and thereby, increasing their flexibility of management and use.

Implantable medical devices fused with network communications, such as pacemakers, have been adopted for essential treatment of critical conditions such as tachycardia [8], [9]. Tachycardia condition makes one heartbeat faster than the average per minute [8], [9]. This can occur when the electrical signals in the upper chambers of the heart misfires resulting in increased heart rate [4, 5]. In such a condition, the heart is not able to fill with blood before contracting, and this reduces the blood flow to the rest of the body [8], [9]. Other related conditions include ventricular (a condition in which the electrical signals in these chambers fire wrongly) and sinus tachycardia which occurs when the heart's natural pacemaker transmit electrical signals faster than normal [8], [9]. Patients experience symptoms such as dizziness, shortness of breath, chest pain and heart palpitation. Sever issues includes unconsciousness and cardiac arrest. Implanted medical devices known as pacemakers are used in the management of these conditions [8], [9].

These network-enabled medical devices can also enhance the implementation of other functionalities. Such as continuous care which is not possible with medical devices not fused with communication networks [8], [9].

Much as medical devices are sustaining millions of lives, they are associated with some vulnerabilities. Recent studies showed vulnerabilities with potential risks to patients who are using devices with medium- or long-range wireless systems [10]. According to the FDA, cybersecurity is the process of preventing unauthorized access, unauthorized use, unauthorized modification, or misuse of information, which is accessed, stored or transmitted from a device to an external receiver [1], [14].

Cyber criminals can be heartless to an extend of taking undue advantage of these vulnerabilities to hack into medical devices with the intention to cause harm.There have been similar instances where cybercriminals hacked epilepsy support websites and posted animated images which caused pain and seizures to photosensitive epileptic patients [11], [12]. So, the communications to and from pacemaker can be compromised. This can lead to injuries or death [10], [16], [17]. More to this, security loopholes have also been discovered in some class II medical devices. Insulin pumps were assessed to have the potential of delivering excess insulin if the vulnerabilities found in them are exploited[6]. Additionally, the serial number was used to hack into an insulin pump such that the device could be disabled by the hackers [18]. The impact of such an attack could be life-threatening for people with diabetes.

Attack surface on medical devices increase as the number of devices connected to the internet increase [19]. This has increased the possibility of endangering patient lives since attackers can be able to access sensitive information and can infect devices with malware [20]. IMDs such as pacemakers, neurostimulators, implantable cardiac defibrillators (ICDs), and drug delivery systems have become target of attacks in

recent times [21].

In a vulnerability assessment in medical devices [10], Shodan (a search engine for IoT devices) was used to obtain a large collection of IP addresses that were scan with Nessus (a vulnerability scanner) to determine the existence of vulnerabilities. The study identified 1,604/16,078 (9.97%) of devices with vulnerabilities. In general, about 3,964 vulnerabilities were found in 1,604 devices. 345 devices had 'Critical' vulnerabilities, 411 with 'High' vulnerabilities, 1,468 with 'Medium', and 1,740 with 'Low' vulnerabilities. Dropbear SSH (a software package that provides a Secure Shell-compatible) Server was found to be one of the most common and critical vulnerabilities which hackers can execute malicious codes to disclose sensitive information in database. Other devices which were found to have vulnerabilities include some radios designed to communicate with the medical devices such as cardiac pacemakers, implantable neurostimulators, and implantable infusion pumps.

Additionally, vulnerabilities were identified in Magnetic Resonance Imaging (MRI) scanners and X-Ray machines. Furthermore, the study found devices with Electronic Health Records (EHR) software that have default community names of Simple Network Management Protocol (SNMP) of which hackers can gain ingress into the respective networks of these devices and can be able to access other network nodes [15].

With all the enormous benefits of network enabled medical devices, they are life-threatening security issues for the patients [22] ranging from network failures to hacking of medical devices. This raises serious concerns about the security and privacy of patients [12], [22], [23]. Various legal requirements including regulations, directives and laws were examined in this study towards enhancing the security of medical devices.

1) Research problem, objective and scope: The double-stress of a patient who has to battle with the effect of a disease as well as the fear of being harm due to medical insecurity call for more research in medical devices to overcome this challenge. The objective of this work is to therefore identify, assess and analyse the legal requirements in medical devices towards enhancing their usage safety for patients.

II. Background

A medical device per the World Health Organization, is an instrument, machine, object, or an apparatus that can be used for diagnosis, treatment, monitoring, and prevention of disease or illness [1], [2]. Similarly, in the EU, medical devices include "any instrument, software, or other tools, intended by the manufacturer to be used for diagnosis, prevention, monitoring, treatment, or alleviation of disease" [3]. Medical devices vary from each other based on their design, implementation and application. These devices can be made of software only, hardware only or a hybrid of both [3]. But most of the critical medical devices are made of both hardware and software to enable them to be more fit for vital use. Additionally, most of these medical devices are incorporated with communication technologies and networks to enhance their performance. Medical devices which are integrated with communication networks provides better ways of diagnosing, treating and monitoring of different kinds of medical conditions including heart related conditions and chronic diseases.

Such devices include wearable, connected-on-site equipment and implantable medical devices. These advanced medical devices have transformed diagnosis, treatment and monitoring of various medical conditions and have even increased life expectancy in the United States to about 10 years [1]. Many of such devices include vital sign monitoring devices, glucose monitoring, infusion pumps, electrocardiograms (ECG), implantable pacemakers, insulin pumps, blood pressure monitors, radiology equipment, ventilator machines embedded sensors, ECG sensors, acidometers and intensive care unit (ICU) equipment [1], [5]. Medical devices fused with communication technology have tremendously improved the efficiency of healthcare facilities. Currently, medical devices collect, process, analyze, measure, share and transfer biological signals in real-time.

Implantable medical devices (IMDs) including pacemakers and implantable cardioverter-defibrillators (ICDs) are developed to boost the physiological functioning of some organs such as the heart. Heart related problems could result in slow heartbeat rate, fast heartbeat rate and irregular rhythms in the heartbeat [6]. In 2001, about 25 million people in the US were recorded to be dependant on these devices for life-threatening functions [7]. Currently most of these devices are wirelessly made such that they can be able to communicate with remote equipment of about 5 meters away. ICDs and IMDs can now be remotely configured by doctors while avoiding the need for numerous invasions into patients. This may also reduce infecting sterilized operating rooms due to the need for the proximity of configuration equipment. Additionally, IMD devices transmit alerts to remote monitoring stations in which reports can be generated for the patient's physician to be analysed without causing interference to the patient' activities. But the adoption and usage of these devices require some legal considerations.

Legal requirements in this context include the laws and directives which are enforcing medical device security [51]–[53]. Laws are rules which are established by the appropriate bodies to control behaviours [51]–[53]. These can be categorized into regulatory law, statutory law, constitutional law and common or case law [51]–[53]. Statutory laws are enacted by governmental organs such as the legislation or the parliament [51]–[53]. Regulations are written to primarily implement specific aspects of the law [51]–[53]. Regulations and directives such as FDA, HIPAA, GDPR and EU MDR provides a framework for regulating medical device manufacturers and healthcare providers. Within the EU, when regulations are issued and implemented, all EU and their affiliate European Economic Area (EEA) members can directly apply the regulations without the need for the governments of the EU member states to pass legislation to implement the regulations [24], [25], [33], [35], [36]. On the contrary, directives are legal acts in the EU which are written to enable member state to obtain a desired result. Each member state is given the opportunity to define their ways and details of implementations of the directives [24], [25]. Essentially, a directive cannot be directly applied in member states in EU unless it is passed through legislation [24], [25]. Common Law, which is often used interchangeably with case law, refers to the precedents and authorities which have been set by previous court rulings, judicial decisions and administrative legal findings or rulings [53], [54]. In the U.S., constitutional law comes from the U.S. constitution, a state

constitution or local constitution, bylaws or charter [52], [53].

Statutory law is subdivided into criminal law, and civil law. Criminal law has various laws between individuals and organizations or among these parties. Criminal laws are deterrence in structure, with the primary objective to deter adversaries who are responsible for cyberattacks [24], [25], [51]–[53]. Some of the civil laws are contract law, employment law, family law and tort law [51]–[53]. Tort is a behaviour that causes harm to the complainant (in this context, the patient who is using the medical device) leading to legal liability for the involved person who committed the act (the malicious actor) [24], [25], [51]–[53]. Tort law therefore enables parties to seek redress in the event of injuries pertaining to physical, personal or financial injuries. Other related laws include Battery and Trespass to Chattels. Battery involve deliberate touching of the claimant which is tantamount to the physical invasion of the injured patients [27], [28]. Trespass to Chattels is violated when there is a deliberate interference with one's personal property which has resulted in the cause of an injury [27], [29].

Due to widely adoption of networked medical devices, legal requirements have become important in dealing with security-related challenges. This study therefore surveyed for the most common and recent regulations, laws and directives of medical devices in the US and EU towards enhancing the security of medical devices [24], [25], [51]–[53].

III. Related Work

Realizing the need to improve on the cyber security of medical devices, various researches have been conducted to strengthen the security of medical devices. In that light A.J. Burns et al. presented the legislative timeline and the evolving threats to information security in medical devices in the US with the aim to provide attention for future action [59]. Katherine Booth et al., also analyzed the legal gaps in medical devices in the US towards addressing medical device security and privacy issues [27]. These studies significantly contributed knowledge towards enhancing the security of medical devices.

Additionally, various studies [1], [31], [44], [45], [62] focused on the regulatory aspect. Daniel et al studied into how medical device regulation Perform in the United States and the European Union. This compared medical regulations in both US and EU, however, legal requirements of medical requirement is not limited to device regulations alone [45]. Additionally, Halperin et al developed a framework towards security and privacy measures in medical devices for the adoption of manufactures and regulatory bodies, having analysed the general operations of medical devices [62]. Additionally, Mariela Yaneva et al also identified some legal regulations of biomedical devices pertaining to EU [31]. Tahreem Yaqoob et al conducted a study into information security vulnerabilities in medical devices and other applicable regulations to provide suggestions towards enhancing the security and privacy of healthcare devices [1]. Jon et al work focused on vulnerable software in medical devices regarding patching and updating, manufacturers responsibilities towards assisting FDA processes to address security issues [44].

While these studies contributed to the body of knowledge in the context of medical device security, some of the studies [27], [59] focused their scope on only US and other studies focused on only regulations of the legal aspect [1], [31], [44], [45], [62].

IV. Method

A literature survey was conducted in Google Scholar, Science Direct, Elsevier and IEEE XPlore for legal requirements of medical devices. The most popular legal requirements of US and EU were identified and assessed towards enhancement of the security measures in medical devices. Keywords and phrases such as medical device, regulations, laws, directives and vulnerabilities were used in searching for the related literature. These words and phrases were combined with Boolean functions of AND, OR and NOT.

V. Findings of Legal Requirements

In the US, the Food and Drugs Administration (FDA) is the main regulatory body, responsible for regulating the development and certification of medical devices [10], [27]. Federal Communications Commission (FCC) [10], [27] and the Centers for Medicare and Medicaid Services (CMS) [10], [27] [10], [27] are other auxiliary agencies which are supporting the FDA in the regulations of medical devices. The FDA uses Federal Food, Drug, and Cosmetic Act (FD&C Act) in regulating the medical devices [10], [27], [54].

There are various categories of medical devices [10], [24] as depicted in Fig. 1 and Table I. Some of them do not present unreasonable risk of illness or injury while others could present unreasonable risk of illness or injury [10], [24] and are intended to be use in supporting or sustaining human life. So, regulatory classification was developed based on these risks that the devices pose to humans as shown in Table I. The level of controls required to ensuring the safety and effectiveness of the devices were also considered. The medical devices have hence been categorized into Class I, Class II and Class III. The Class I devices are basic and common medical devices which have low to medium risk, low complexity and consist of about 47% of the total medical devices [1]. The class I devices are basically not internet enabled and are exempted from regulatory controls based on their low security risk [10]. Example of class I devices include Lancet, and dental floss [10]. The cybersecurity issues are mostly around the class II and class III medical devices [10], [20]. The class II devices pose medium to high risk to patients. Class II devices are more complex and partially implantable [10]. They form about 43% of the total number of medical devices [1] and these devices include Syringe, Insulin pump and blood glucose meters (BGM) [1], [10], [24].

The class III medical devices consists of only 10% of the total medical devices and are categorized into the highest security level, requiring the most strict security measures [1]. They are fully implanted to regulate body functions. The class III medical devices include Artificial pancreas, Continuous glucose monitoring (CGM), pacemaker and Replacement Heart valves

The relevant regulations of FDA on medical devices and the processes therefore involve:

TABLE I. MEDICAL DEVICE CLASSES [10], [27]

Medical Device Class.	Attributes	Example Devices
Class I	Common, low risk, low complexity	Lancet, Dental Floss
Class II	More complex, greater risk to patient, partially implanted	Syringe, Insulin Pump, Blood Glucose Meter
Class III	Fully implanted, greater risk, regulate body functions	Artificial Pancreas, CGM, Replacement Heart Valves

- Medical device listing and establishment registration: The manufacturers and distributors of medical devices must register their organization with the FDA to be able to market their product. Organizations must provide full details of the medical devices being manufactured.
- Labeling: Labeling must be in accordance with information and description of the device usage.
- Medical Device Reporting (MDR): manufacturers/importers/healthcare facility must report events of device malfunctions or causes of serious injuries or death to the FDA. This will enable FDA to detect and correct issues.
- Quality System (QS) regulations: Indicates requirements relating to controls, facilities, and methods used in the entire medical device life-cycle. These indications include designing, purchasing, manufacturing, labeling and packaging, servicing, and installation of the devices. The FDA is responsible to ensure that the devices fulfill important specifications and requirements.
- Investigational Device Exemption (IDE) for clinical studies: This enables manufacturers to provide device-specific effectiveness and safety data to support Pre-market notification (510-k) or post-market approval (PMA)application.

FDA satisfies medical devices after going through a total product life cycle method which has two important phases thus pre-market notification/510-k approval and post-market approval (PMA). Manufacturers need to provide detailed information with evidence of the device use safety and effectiveness as shown in fig 1. FDA then validate the information in addition to sharing identified security vulnerabilities, monitoring and examination of connected medical device's effectiveness and safety.

Medium risk related devices are mostly routed through the 510-k approval process. Significant assurance of the medical device's safety and effectiveness are normally provided by the manufacturer who submits a 510-k application. Basically, the 510-k application is exempted from non-clinical and clinical data of showing the effectiveness and safety of the device. But the high risk devices goes through PMA, which involve a complete review of the device including the device's clinical and non-clinical trials and testing data.

Health Insurance portability and accountability act (HIPAA) privacy and security rules were passed for protecting personal health and medical records in the United States of America (USA). The HIPAA rules covers healthcare providers, health plans and healthcare clearing housing entities. HIPAA

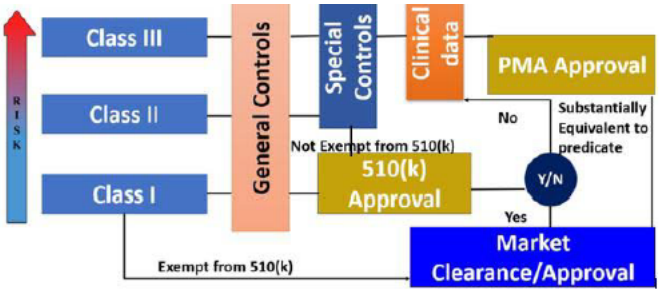


Fig. 1. FDA Medical Device Regulation Process [10]

privacy and security rules primarily protect personal identifiable health information (PHI) including names, diagnosis and identifiable numbers of medical device [27]. This rule therefore demands for appropriate privacy and security protection controls. However, the mandate of the HIPAA rules excludes the protection of pharmaceutical companies and medical devices [1], [27]. As HIPAA concentrates on the protection of PHI, it does not extend its mandate to include the protection of cyber-attacks against medical devices. The regulation of medical device manufacturers are not also covered by HIPAA regulations [27].

Cybersecurity issues should be addressed by the manufactures at the design and development stages. The process should involve [59], [60]

- Identifying assets, threats and vulnerabilities.
- Assessing the impact of the threats and vulnerabilities on device functionality and patient or user.
- Assessment of the likelihood exploitation of the threats and vulnerability.
- Assessing residual risk and risk acceptance criteria.
- cybersecurity documentation should be done to include.
  - Traceability matrix between security controls and their risks.
  - Hazard analysis, mitigation and design consideration.
  - Documented plan for validating software update in the life-cycle of the device.
  - Documentation of controls that have been implemented to assure the integrity of the device.
  - Instructions for the device use and specification on cybersecurity controls needed for the

intended use environment.

- Appropriate standards should be followed and documented.

The post-market management of cybersecurity in medical devices is to complement the premarket management, to form a comprehensive security measure. So the security measures cover the design, development, peoduction, distribution, deployment and maintenance stages [61]–[63]. As cyberseurity issues continue to evolve, it is not possible to put in measures to take care of all issues at one point in time. So after the device has been deployed on the market, the MDM, need to always document complaint handling, quality audit, corrective and preventive actions, software validation and risk analysis and serving, as specified in the quality system regulation. In addition, MDM need to [61]–[63]:

- Constantly identify vulnerabilities and risks and assessing their impact by monitoring cybersecurity information sources.
- Maintain software life cycle process such as monitoring third party software for vulnerabilities, design verification and validation for the software updates and patches.
- Using threat models and vulnerability handling process standards (e.g. ISO/IEC 30111:2013) to maintain safety and security.
- Adopting standard procedures (e.g. ISO/IEC 29147:2014) for vulnerability disclosure.
- Timely deployment of mitigation measures to address cybersecurity issues prior to exploitation.
- Other guidelines include, having a structure and systematic approach to risk and quality management, as provided in 21 Code of Federal Regulation, part 820.
- The MDM is to also follow procedures that are in line with the NIST framework for improving critical information cyber security (Identify, Protect, Detect,Report and recover).
- Maintaining safety and core functionality of the device to prevent patient harm.
- Adopting appropriate for managing cybersecurity risks.
- Assessing the exploitability of vulnerabilities.
- Assessing the severity of harm to patients.
- Assenting and controlling risk of patient harm.
- Mitigating and reporting vulnerabilities.

Aside the regulatory laws, statutory laws were also identified in the U.S. to have protection for medical devices. These include Computer Fraud and abuse Act (CFAA) and Anti-Tempering Act [1], [27], [51]–[53]. CFAA punishes cybercriminals who access medical devices or transmits code which resulted in causing harm [27], [53].Within the scope of this law, the medical device manufacturer(MDM) or hospital network is not charged with negligence of duty [27]. But the cyber-criminal under this behaviour is fined, imprisoned for not more than 10 years or both [27]. Under the Anti-Tempering Act, it is a criminal offence to temper with consumer products including medical devices [27]. This Act directly applies to cybercriminals in a breach scenarios but does not apply to MDM or hospital networks. In the context of common or case law, [27] there exist tort liability in which the cybercriminal can be liable to Trespass to Chattels or Battery. When a patient is injured through medical device attacks, the patient can take a civil cause of legal action against the malicious attacker, the device manufacturer and the hospital. The hospital can be charged if the compromised device was as a result of cyber attacks on the hospital's network. A medical device manufacturer or a hospital may be held accountable for negligence if they fail to comply with established cyber security measures [27], [53].

In the European Union (EU), Medical Device Directives (MDD) was responsible for regulating the marketing and safety of medical devices as far back as 1990 [1], [24], [32]–[36], [40]. But this has been changed to regulation 2017/745 of EU [1], [24], [31], [39]. EU also classified its medical devices but what is different is that, EU has four number of classes in accordance to their risk level and purpose. The classes are I (Is and Im), IIa, IIb and III with respective increases in the level of assessments.Before a medical device is advertised in any EU country, it must first go through the systematic regulatory assessment in order to obtain the Conformite Europe (CE) mark [1], [31]. CE mark means the device satisfied the safety criteria and can be sold without further controls.

The national competent authorities which is formed by EU member states, observe, appraise and nominates notifying bodies (NBs) to be responsible for these conformity processing [1]. Other vital responsibilities of this body are device certification, class designation, quality system verification, and assessment, and design profile reviews. The approval process of a device involves the selection of an NB by the manufacturer to grant certification of a new device for CE marking [30]. The NB then obtain technical details of the device based on its class [1]. The information is used to review the safety of the device [1]. Usually, devices in each class must declare its conformity to the EU directives and the specific conformity assessment plan [1]. Also designs of devices in the highest class have to be assessed however, devices in the lower class I are exempted from such regulations [1]. In Spite of that, these class1 devices must follow vital propositions of efficacy and safety in their design alongside with labelling and construction requirements. After a medical device is approved, there is post-market surveillance by competent authority through the authority of member-state [1].

As the devices are getting sophisticated, better regulations are much needed since the current directives have not catch-up with the technical and scientific developments pace in the healthcare domain. Currently devices are not thoroughly assessed in the pre-market phase except medium to high-risk devices which go through conformity assessment for the NB to decide on the needed controls of the device safety [1], [40].

The regulation (EU) 2017/745 which has recently been written [1], [39], will enable NBs to visit manufacturers on their site without prior notice towards ensuring safety, security, quality, and performance of the medical devices. The Medical Device Regulation of EU have incorporated some security

measures for device manufactures. These security measures are [39]:

- Under Regulation 17.2. Medical device manufactures (MDM) shall follow state-of-the-art development and manufacturing process, including the principles of development life cycle, risk management, including information security, verification and validation.
- In regulation 17.4, Medical Device Manufacturers(MDM) shall specify minimum requirements to run the medical device and software as intended and the specification should include hardware, IT networks characteristics and IT security measures including protection against unauthorized access.
- In regulation (39) of the MDR, MDM are to provide clear and easily accessible essential information to patients who are on implanted medical devices. Information that should be provided include information concerning how the implanted device can be identified, any necessary health risk warnings or precautions to be taken. Such warnings or precautions includes information as to whether or not the device is compatible with certain diagnostic devices or with scanners used for security controls.
- Under regulation 4.5, MDM are required to provide description of the arrangements that fulfil existing rules controlling the protection and confidentiality of personal data, such as [39]:
  1) organizational and technical arrangements that will be implemented to avoid unauthorized access, disclosure, dissemination, alteration or loss of information and personal data processed;
  2) a description of measures that will be implemented to ensure confidentiality of records and personal data of subjects; and
  3) a description of measures to be adopted towards mitigating potential adverse impact in the event of data security breach.
- Under Section 4.1, a signed statement must be provided by the natural or legal person responsible of the MDM satisfying that the medical devices is in conformity with the general safety and performance requirements and that precautions has been taken to protect the health and safety of the subject.
- In Section 4.3, MDMs are to provide and proof insurance cover or indemnification of subjects in case of injury, pursuant to Article 69 and the corresponding national law.

The general data protection regulations (GDPR) of EU's privacy-related regulation is concerned with the processing of personal data by a data processor or a data controller in EU. The GDPR defines personal data to include information which can be linked to an identifiable person [38]. Unlike the HIPAA regulation, the GDPR is application to all sectors that are processing personal information of the EU citizens. Biometric data, genetic data and PHI are classified under sensitive information. Explicit consent is required in order to process such data. The GDPR also applies to all healthcare organizations, health insurance companies, and medical device manufacturers [37].

Accordingly, there are no general Applicable laws as at now, which are serving the purpose of cybersecurity only in Norway [42], [43]. The cybersecurity regulations are fragmented into sector specific [42], [43]. In the context of common or case law, there exists a criminal code which is originally known as Penal code in Norway [42], [43]. This code is for handling criminal cases. On April 8, 2005, the penal code relating to cybercrime was amended and enacted to include various offences. The offensive provisions are [42], [43]:

1) Under Penal Code 151 b: [42], [43] Any person who is found guilty of destroying, damaging, or putting out of action any data collection or any installation for supplying power, broadcasting, telecommunication, or transport, causes comprehensive disturbance in the public administration or in community life in general shall be liable to imprisonment for a term not exceeding 10 years. If the aforementioned act was found to be negligent acts the person shall be punishable by fines or imprisonment for a term not exceeding one year.
2) In Penal Code 145b: "Any person who unlawfully disclose or make available a computer password or similar data, by which the whole or any part of a computer system is capable of being accessed, shall be sentenced for spreading of access data, to a fine or imprisonment not exceeding 6 months or both". If the act involves serious spreading of access data the culprit shall be sentenced to imprisonment not exceeding 2 years.

Also, Under section 204 of the Penal Code of 20 May 2005, some violations are punishable upto two years imprisonment or by fines. Some of these offensive activities include unauthorised access or hacking, Denial-of-service-attacks, phishing, infection of IT systems with Malware and possession or use of tools for committing cybercrime. Other punishable offences are identity theft, electronic theft and any activity that can have adverse effect on CIA of any IT system, infrastructure, communications network, device or data [41]–[43]. A summary of the findings are shown in Table II. where the legal requirements are listed with their respective origin.

A. Gap Analysis

In the European Union (EU),the GDPR and the EU Medical Device Regulations (Regulations 2017/745) [1], [39] have some intersections towards holding device manufactures to be responsible of negligence of duty in the event of device compromise [1], [24], [39]. However, there are gaps in the HIPAA privacy and security rules in the regulation of medical devices. HIPAA does not concern itself much with the security of medical devices [1]. Unlike the GDPR, which holds both hospitals and device manufacturers responsible for data protection in medical device regulations in EU, the HIPAA privacy and security rules are only limited to the healthcare entities such as hospitals and other healthcare providers. HIPAA provides heavy penalties for breaches against patient health information (PHI). MDM who deals with healthcare entities directly are

TABLE II. SUMMARY OF LEGAL REQUIREMENTS FOR MEDICAL DEVICES

#	Legal requirement	Origin
1	Food, Drug, and Cosmetic Act (FD&C Act) [10], [27], [54].	U.S.
2	Health Insurance portability and accountability act (HIPAA)	U.S.
3	General Data Protection Regulation (GDPR)	EU
4	Medical Device Regulation 2017/745 of EU [1], [24], [31], [39]	EU
5	Computer Fraud and abuse Act (CFAA) [1], [27], [51]–[53]	U.S.
6	Anti-Tempering Act [27]	U.S.
7	Trespass to Chattels [27], [53]	U.S. /EU
8	Battery [27], [53]	U.S. /EU
9	Penal Code [42], [43]	EU

covered by HIPAA but not when devices are directly sold to patients This does not adequately cover the protection of the entire medical devices against cyberattacks [1], [6], [24], [27].

In this shortfall of HIPAA, privacy concerns are not also addressed in medical devices. According to [1] safety and security issues are also affected in scenarios where devices are prone to safety and security risks. But FDA does not provide guidelines for MDM to explicitly deal with that [1].

Furthemore, FDA have some cybersecurty guidelines for controlling the security of medical devices and these guidelines dependent on NIST's recommended security framework for critical infrastructure [1]. Though the guideline is useful, it was not specifically developed for enhancing the security of medical devices. The severity of hazards pose by medical device malicious errors and non-malicious errors could be different from conventional IT systems [1]. Example, malicious error in water or power system could cause a substation to go off. But in the context of medical device, a malicious or non-malicious error could have direct harm on the patient ranging from pain to death in a short time [1]. Also, the Food, Drug and Cosmetic Act of U.S., have detailed description for safety controls for medical devices but specific security related controls are limited [1].

Additionally, quality and safety labelling of medical devices has been a requirement but cybersecurity labelling of medical devices have not been adopted [1], [27], [39]. This makes is difficult for patients to choose secure medical devices. Again, the FDA require hospitals and device users to report serious security issues within a time line. But this has been found to be violated due to lack of capacity and training in timely determination of security issues.

Within the confines of statutory and case laws, for a patient to establish a claim arising from harm of cyber attack, the patient must proof that the defendant deliberately interfered with his or her possessory interest without authorization [27], [45]–[49]. Also if there is an unauthorized access by the defendant which resulted in a harm to the involved patient, the defendant can be found liable in such scenario. The difficulty is that the patient or the plaintiff may not be able to provide justification for the intention of attacks. According to [1], a number of hospitals and MDMs have been fined for various offences including failure to report faults on medical devices [55], [56], failure to follow PMA regulations [55]–[57], safety issues with medical devices [27], [56], [58],and for selling unapproved medical devices [1], [57]. Apparently, this will deter others from committing related acts but security related offences were not seen.

VI. DISCUSSION

As threats to information security evolve, security requirements such as regulations, directives, statutory law and case laws are also revised accordingly. These requirements are usually updated to enhance their ability to mitigate current and foreseeable threats. This study was therefore conducted to identify the state-of-the-art legal requirements which are being used to control the security of medical devices. Medical devices serve critical functions in the sustenance of human life in the eHealth space [1], [2], [24], [27]. But the current laws that exist to safeguard these devices in terms of security and how adequate they are, need to be assessed. Regulations and their procedures, statutory law and case law or common law were identified and assessed in the study as shown in Table II.

With reference to Table II, in the U.S., the FDA is the main body that is regulating medical devices, using FD&C Act [1], [13], [24], [27], [44], [45]. In this regards the effectiveness of the security regulations were assessed. Also in the event of a device compromise, the responsible bodies or were also analysed. For example, who will be liable if a patient's medical device was hacked? Per the state-of-the-art studies, those who will be liable include the attacker, the MDM and the hospital if the medical device was compromised due to attacks on the hospital network [1]–[4], [27], [28]. In recent prosecutions of offenders of FDA regulations and HIPAA privacy and security rules in the U.S., those who were found liable are hospitals and MDA [1], [56]–[58]. None of the liabilities involve security issues left alone to charging a cybercriminal on the account of medical device compromised. Some of the legal structures have not fully addressed the threat of cyberattacks. For instance, it is sometimes difficult to identify and indict culprits of cyber criminals [27], [46], [47]. In some cyberattack instances, the adversaries conceal their identity, cover their tracks or at worse can divert the act on others through source spoofing [27], [47]. Much as it remains challenging to identify and get hold of the perpetrators behind cyberattacks, the criminal law remains insufficient as a deterrent measure [27], [47]–[49].

In comparing the medical device regulations of EU and that of FDA, the EU has comparative placed a higher responsibility for device manufacturers to be proactive in both pre-market and post-market release of the medical device [1], [39]. Literally, the EU ask their device manufacturers to take insurance cover for patients who are using their devices [1], [39]. In order for them not to pay claims,MDM in the EU will be encouraged to enhance security. The Insurance company of the medical device will also want to mitigate risk by charging the appropriate premium based on the severity of vulnerabilities in the medical devices. So the insurance company will also have interest in the level of security of the device. With all these

actors involve, the level of security in medical devices can be greatly improved.

Common law principles can descend on MDM on liabilities relating to negligence of duty to protect medical devices against cyberattacks [1], [27], [47]. If an MDM fails to implement acceptable cybersecurity measures then that MDM can be liable to negligent of duty of care. But the duty of care is relative in cybersecurity breaches [27]. Standard and guidelines changes as the threats in cyberspace changes. This complicates the identification and specification of duty of care. For instance, under Regulation 17.2 of EU MDR, MDM shall follow state-of-the-art development and manufacturing process, including the principles of development life cycle, risk management relating to information security, verification and validation [39]. The point in time where MDM becomes liable for negligence of duty in cyberattack of a medical device may be difficult especially in phases where standards and guidelines are undergoing changes [1], [27]. In some states in USA, where the patients' injury was directly caused by the acts of the adversary, the MDM was exonerated from acts of negligence liability [27]. Further to this, on the basis that an MDM was certified by FDA, through the PMA process, injured patients cannot hold the MDM liable [26], [50]. Based on these, there are uncertainties regarding negligence of duty actions against an MDM.

Furthermore, there is a gap on the share responsibility of regulators and bodies that certify medical devices. In the literature studies [1], [6], [24], [27], [39], [56], none of them blame the regulatory bodies in the event of cyberattack. But regulatory bodies need to be hold accountable for attacks on medical devices which they have approved. For instance, if a medical device was approved to be safe and secure by a regulatory body like the FDA when in fact it has some security loopholes, it could be that the regulatory body did not do due diligence. Notwithstanding, FDA and HIPAA were not primarily provided to safeguard against cyberattacks of medical devices and could lacks adequate regulatory safeguards [1], [27]. So the regulatory body may not be liable if the security assessments of the device was not part of their mandate [27]. The FDA and HIPAA need to improve upon their regulations to fully cover the security of medical devices such that the MDM and regulatory bodies can directly be responsible to vulnerabilities found in medical devices. In this way, the circle of efficiency maybe getting completed. Regulatory bodies would want to comprehensively assess a medical device for vulnerabilities in medical devices such that they will not be liable in the event of breaches. This would also compel MDM to want to put in the necessary measures to have their medical devices approved.

But, with this approach, there are also ethical hurdles that need to be cleared. If a potentially unsecure medical device is approved for use, patients can be vulnerable to attack[11]. On the contrary, if a device is not approved due to security reasons, that device may never be available for patients [23]. This implies that many more patients would be harmed since there would not be any effective treatments for the conditions [23], [27].

Comparing the legal requirements of U.S. and EU, the EU general data protection regulation(GDPR) highly complements the medical device regulation of EU. That is not the case, between the FD&C Act and HIPAA privacy and security Act OF U.S. The HIPAA has distanced itself a bit when it comes to medical device regulations [1], [6], [24], [27]. This has weakened the regulatory security controls in enforcing security measures in medical device. This is because aside the NIST's guidelines on critical infrastructure, FDA do not have tailored guidelines for controlling the security of medical devices [1], [6], [24]. A combined effort of FDA and HIPAA will greatly enhance the security of medical devices since HIPAA privacy rules will be extended to handle privacy issues of device manufacturers and while the HIPAA security rules handle the security concerns of the hospital and device manufactures [1].

A. Conclusion

Patients who are dependent on medical devices such as pacemakers and artificial pancreas are vulnerable to cyberattacks. This study therefore conducted a state-of-the-art on security requirements, concerning these devices in the US and EU. Food, Drugs and Cosmetic Act, HIPAA, Medical Device Regulations of EU and GDPR were some of the identified regulations for controlling the security of these devices. Statutory laws such as computer Fraud and abuse Act (CFAA), Anti-Tempering Act, Panel Code as well as Battery and Trespass to Chattel in the civil law, were also identified.

In analysing the security requirements, there are less motivations on criminal charges against cybercriminals in addressing the security issues. Because it is often challenging to identify the culprits in medical device hacks. It is also difficult to hold device manufactures on negligence of duty especially after the device has been approved or if the harm on patient was as a result of a cyber attacker.

Suggestions have been provided to improve upon the regulations so that both the regulatory bodies and MDM can improve upon their security conscious care.

However, this raises an ethical issue of balancing the practice of using a very secured medical devices which may take a long time to develop, versus causing more harm to patients who may not have the device to use due to stringent security regulatory processes. Future studies will analyse these ethical dilemmas to provide a balance point of enforcing security requirements while ensuring availability of the medical devices.

REFERENCES

[1] Yaqoob T., Abbas H., Atiquzzaman M. Security Vulnerabilities, Attacks, Countermeasures, and Regulations of Networked Medical Devices—A Review. IEEE Communications Surveys & Tutorials. 2019;21(4):3723-68.

[2] Syring G. Overview: FDA regulation of medical devices, Accessed June 06 From: https://www.fda.gov/medical-devices/overview-device-regulation/history-medical-device-regulation-oversight-united-states

[3] Fatema, N. and Brad, R., 2014. Security requirements, counterattacks and projects in healthcare applications using WSNs-a review. arXiv preprint arXiv:1406.1795.

[4] Tanev, G., Tzolov, P. and Apiafi, R., 2015. A value blueprint approach to cybersecurity in networked medical devices. Technology Innovation Management Review, 5(6).

[5] Yu, B., Kang, S.Y., Akthakul, A., Ramadurai, N., Pilkenton, M., Patel, A., Nashat, A., Anderson, D.G., Sakamoto, F.H., Gilchrest, B.A. and Anderson, R.R., 2016. An elastic second skin. Nature materials, 15(8), pp.911-918.

[6] Denning T., Borning A., Friedman B., Gill B. T., Kohno T., Maisel W. H., editors. Patients, pacemakers, and implantable defibrillators: Human values and security for wireless implantable medical devices. Proceedings of the SIGCHI conference on human factors in computing systems; 2010.

[7] Pope A., Bouxsein P., Manning F. J., Hanna K. E. Innovation and invention in medical devices: workshop summary: National Academies Press; 2001.

[8] Fisher J. D., Kim S. G., Furman S., Matos J. A. Role of implantable pacemakers in control of recurrent ventricular tachycardia. The American journal of cardiology. 1982;49(1):194-206.

[9] TOIVONEN L., VALJUS J., HONGISTO M., METSO R. The influence of elevated 50 Hz electric and magnetic fields on implanted cardiac pacemakers: the role of the lead configuration and programming of the sensitivity. Pacing and Clinical Electrophysiology. 1991;14(12):2114-22.

[10] Halperin D., Heydt-Benjamin T. S., Ransford B., Clark S. S., Defend B., Morgan W., et al., editors. Pacemakers and implantable cardiac defibrillators: Software radio attacks and zero-power defenses. 2008 IEEE Symposium on Security and Privacy (sp 2008); 2008: IEEE.

[11] Ertl B.: coping-with-epilepsy.com; 2007 [Available from: Hooligans Attack Epilepsy Patients During Epilepsy Awareness Month.

[12] @wired. Hackers Assault Epilepsy Patients via Computer. 2020.

[13] McMahon E., Williams R., El M., Samtani S., Patton M., Chen H., editors. Assessing medical device vulnerabilities on the Internet of Things. 2017 IEEE International Conference on Intelligence and Security Informatics (ISI); 2017: IEEE.

[14] Food and Drug Administration, 2015. Content of Premarket Submissions for Management of Cybersecurity in Medical Devices-Guidance for Industry and Food and Drug Administration Staff.

[15] McMahon E., Williams R., El M., Samtani S., Patton M., Chen H., editors. Assessing medical device vulnerabilities on the Internet of Things. 2017 IEEE International Conference on Intelligence and Security Informatics (ISI); 2017: IEEE.

[16] Cybersecurity and Hospitals. American Hospital Association Accessed on June 7 2020 From:https://www.aha.org/system/files/2017-12/ahaprimer-cyberandhosp.pdf

[17] Peterson A. Connected medical devices: The Internet of Things-that-could-kill-you. Washington Post. 2015 Aug.

[18] Kaplan D. Black Hat: Insulin pumps can be hacked. SC Magazine. 2011 Aug 4.

[19] Lake D, Milito RM, Morrow M, Vargheese R. Internet of things: Architectural framework for ehealth security. Journal of ICT Standardization. 2014 Mar 31;1(3):301-28.

[20] Sametinger J, Rozenblit J, Lysecky R, Ott P. Security challenges for medical devices. Communications of the ACM. 2015 Mar 23;58(4):74-82.

[21] Halperin D, Heydt-Benjamin TS, Fu K, Kohno T, Maisel WH. Security and privacy for implantable medical devices. IEEE pervasive computing. 2008 Jan 16;7(1):30-9.

[22] Marchang J., Beavers J., Faulks M., editors. Hacking NHS Pacemakers: A Feasibility Study. 12th International Conference on Global Security, Safety & Sustainability; 2019: IEEE.

[23] Sokolsky O., Lee I., Heimdahl M., editors. Challenges in the regulatory approval of medical cyber-physical systems. 2011 Proceedings of the Ninth ACM International Conference on Embedded Software (EMSOFT); 2011: IEEE.

[24] Pesapane F., Volonté C., Codari M., Sardanelli F. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. Insights into imaging. 2018;9(5):745-53.

[25] European Union Regulations, Directives and other acts, 2018. Available via https://europa.eu european-union/eu-law/legal-acts_en

[26] Medtronic, Inc. v. Lohr. US: Supreme Court; 1996. p. 470.

[27] Wellington K. Cyberattacks on medical devices and hospital networks: Legal gaps and regulatory solutions. Santa Clara High Tech LJ. 2013;30:139.

[28] Neal Hoffman, Battery 2.0: Upgrading Offensive Contact Battery to the Digital Age, 1 CASE W. RES. J.L. TECH. & INTERNET 61, 68 (2010).

[29] eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058, 1069 (N.D. Cal. 2000).

[30] Fawaz K, Kim KH, Shin KG. Protecting privacy of BLE device users. In25th USENIX Security Symposium ({USENIX} Security 16) 2016 (pp. 1205-1221).

[31] Yaneva-Deliverska M, Deliversky J, Lyapina M. Biocompatibility of medical devices–legal regulations in the European Union. Journal of IMAB–Annual Proceeding Scientific Papers. 2015 Feb 13;21(1):705-8.

[32] Pesapane, F., Codari, M. & Sardanelli, F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp 2, 35 (2018). https://doi.org/10.1186/s41747-018-0061-6

[33] European Economic Community (1993) 93/42/EEC - Council Directive concerning Medical Devices. Official Journal of the European Communities. Available via http://ec.europa.eu/growth/ single-market/european-standards/harmonised-standards/medicaldevices_ en

[34] European Economic Community (1990) 90/385/EEC - Council Directive on the approximation of the laws of the Member States relating to active implantable medical devices. Council Directive. Available via https://ec.europa.eu/growth/single-market/europeanstandards/harmonised-standards/implantable-medical-devices_en

[35] European Commission (1998) Directive 98/79/EC of the European Parliament and of the Council on in vitro diagnosticmedical devices. Official Journal of the European Communities. Available via https:// ec.europa.eu/growth/single-market/european-standards/harmonised-standards/iv-diagnostic-medical-devices_en

[36] Patients and Privacy: GDPR Compliance for Healthcare Organizations, Accessed on June 8th 2020 From:https://www.trendmicro.com/vinfo/dk/security/news/online-privacy/patients-and-privacy-gdpr-compliance-for-healthcare-organizations

[37] D. Cicco et al. (2018). Toward an Enhanced EU Cybersecurity Framework: Political Agreement Reached on EU Cybersecurity Act—Security—European Union. Accessed: Nov. 17, 2018. [Online]. Available: http://www.mondaq.com/uk/x/709760/Security/ Toward+An+Enhanced+EU+Cybersecurity+Framework+Political+ Agreement+Reached+On+EU+Cybersecurity+Act

[38] (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text With EEA Relevance). Accessed: Dec. 13, 2018. [Online]. Available: https://publications.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1/language-en

[39] (2017)REGULATION (EU) 2017/745 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. Accessed: June 08, 2020. [Online]. Avialable: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745

[40] Directive concerning Medical Devices. European Communities, (1993).

[41] Bastani F, Tang T. Improving security of wireless communication in medical devices. Massachusetts Institute of Technology. 2015.

[42] Cybercrimedata AS, Cybercrime Law, Norway. Accessed: June 09 2020, [Online].Available: https://www.cybercrimelaw.net/Norway.html#: :text=Penal

[43] ICLG.com, Norway:Cybersecurity 2020. Accessed: June 09 2020, [Online]. Available: https://iclg.com/practice-areas/cybersecurity-laws-and-regulations/norway

[44] Martinez JB. Medical Device Security in the IoT Age. In2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON) 2018 Nov 8 (pp. 128-134). IEEE.

[45] Kramer DB, Xu S, Kesselheim AS. How does medical device regulation perform in the United States and the European union? A systematic review. PLoS medicine. 2012 Jul;9(7).

[46] Fouretier A, Bertram D. New regulations on medical devices in Europe: what to expect?. Expert review of medical devices. 2014 Jul 1;11(4):351-9.

[47] Handler SG. New cyber face of battle: developing a legal approach to accommodate emerging trends in warfare. Stan. J. Int'l L.. 2012;48:209.

[48] Jensen ET. Cyber deterrence. Emory Int'l L. Rev.. 2012;26:773.

[49] Richard Clarke, War from Cyberspace, NAT'L INT., Nov.-Dec. 2009, available at http://nationalinterest.org/article/war-from-cyberspace-3278.

[50] Riegel v. Medtronic, Inc., 552 U.S. 312, 128 S. Ct. 999, 169 L. Ed. 2d 892 (2008).

[51] Whitman ME, Mattord HJ. Legal, ethical, and professional issues in information security. Principles of information security (4th ed.; pp. 133–147). Boston, MA: Course Technology, Cengage Learning. Retrieved from http://www. cengage. com/resource_uploads/downloads/1111138214_259148. pdf. 2012.

[52] WARREN E. Legal, ethical, and professional issues in information security. 2011:89-116.

[53] Whitman M. E., Mattord H. J.Legal, ethical, and professional issues in information security. Principles of information security. 6th Edition ed: CENGAGE Learning; 2017: 127-143.

[54] Case Law - Common Law.Accessed: June 09 2020, [Online].Available: https://www.hg.org/case-law.html

[55] Federal Food, Drug, and Cosmetic Act (FD&C Act),June 13 2020, [Online].Available:https://www.fda.gov/regulatory-information/laws-enforced-fda/federal-food-drug-and-cosmetic-act-fdc-act

[56] V. Pollard and M. Davar. (2017). FDA's Evolving Civil Money Penalty Authority: Simple Violations Can Lead to Major Costs. [Online]. Accessed On June 14 2020 Available: https://www.mastercontrol.com/gxp-lifeline/civil_money_penalty_authority_0609/.

[57] A. Brino, Grisly Medical Errors, Some Deadly, Lead to 700K in Fines for 10 California Hospitals, Healthcare Finance, HIMSS Media, Portland, ME, USA, 2020. [Online]. Available: https://www.healthcarefinancenews.com/news/grislymedical- errors-some-deadly-lead-700k-fines-10-california-hospitals

[58] B. Zimmermann, California Fines 9 Hospitals $500k+ for Patient Safety Issues, ASC Commun., Chicago, IL, USA, Accessed on June 14 2020 [Online].Available: https://www.beckershospitalreview.com/quality/californiafines- 9-hospitals-500k-for-patient-safety-issues.html

[59] Burns AJ, Johnson ME, Honeyman P. A brief chronology of medical device security. Communications of the ACM. 2016 Sep 22;59(10):66-72.

[60] Food and Drug Administration. Content of Premarket Submissions for Management of Cybersecurity in Medical Devices-Guidance for Industry and Food and Drug Administration Staff. Accessed on June 14 2020 [Online].Available: https://www.fda.gov/media/86174/download

[61] US Food and Drug Administration. Postmarket management of cybersecurity in medical devices: guidance for industry and food and drug administration staff. 2016. Accessed on June 14 2020 [Online].Available: https://www.fda.gov/files/medical

[62] Halperin D, Heydt-Benjamin TS, Fu K, Kohno T, Maisel WH. Security and privacy for implantable medical devices. IEEE pervasive computing. 2008 Jan 16;7(1):30-9.

[63] Ransford B, Molina-Markham A, Stewart Q, Fu K, Kramer DB, Baker MC, Reynolds MR. Security and Privacy Qualities of Medical Devices: An Analysis of FDA Postmarket Surveillance.

Fine-Tuning Pre-Trained Convolutional Neural Networks for Women Common Cancer Classification using RNA-Seq Gene Expression

Fadi Alharbi[1]
College of Computer and Information Sciences,
Jouf University, Sakaka, KSA

Murtada K. Elbashir[2]
College of Computer and Information Sciences,
Jouf University, Sakaka, KSA
Faculty of Mathematical and Computer Sciences,
University of Gezira, Wad Madani; Sudan

Mohanad Mohammed[3]
School of Mathematics, Statistics and Computer Science,
University of KwaZulu-Natal, Pietermaritzburg,
South Africa
Faculty of Mathematical and Computer Sciences,
University of Gezira, Wad Madani; Sudan

Mohamed Elhafiz Mustafa[4]
College of Computer and Information Sciences,
Jouf University, Sakaka, KSA
College of Computer Science and Information Technology,
Sudan University of Science and Technology,
Khartoum, Sudan

***Abstract*—Most of the recent cancer classification methods use gene expression profile as features because it can provide very important information regarding tumor characteristics. Motivated by their success in the computer vision area now deep learning has been successfully applied to medical data because it can read non-linear patterns in a complex feature and can allow the leverage of information from unlabeled data of problems that do not belong to the problem being handled. In this paper, we implement transfer learning, which refers to the use of a model trained on one task to perform classification on another task to classify five cancer types that most commonly affect women. We used VGG16, Xception, DenseNet, and ResNet50 as base models and then added a dense layer to reflect our five-class classification problem. To avoid training over-fitting that can result in a very high training accuracy and a low cross-validation accuracy we used L2-regularization. We retrained (fine-tuned) these models using a five-fold cross-validation approach on RNA-Seq gene expression data after transforming it into 2D-image like data. We used the softmax activation function with the prediction dense layer and adam as optimizer in the model fit for all four architectures. The highest performance is obtained when fine-tuning Xception architecture, which achieved classification accuracy = 98.6%, precision = 98.6%, recall = 97.8%, and F1-score = 98% on five-fold cross-validation training and testing approach.**

***Keywords*—Fine-tuning; RNA-Seq; gene expression**

I. Introduction

Every cell in multicellular organisms has the same genes and every gene is not transcriptionally active in every cell, therefore the patterns of gene expression differ from cell top another. These variations may play a major role in the difference between disease and health [1]. Therefore, different types of tissues or cells' transcriptomes comparison can reveal an understanding of what constitutes different cells and how changes in transcriptional activity may contribute to diseases. In humans, a small percentage of genetic code i.e. less than 5% of the genome is transcribed from the genome's DNA code into RNA molecules or just a messenger RNA molecule [2], [3]. RNA-Seq or DNA microarray can be used to measure the transcriptome of an organism [4]. The transcription of specific genes is measured by RNA-Seq, which converts long RNAs into a library of complementary DNA (cDNA) fragments, which generate the expression profile. The expression profile can provide very important information regarding tumor characteristics, which offer deep insight into cancer detection problem [5]–[8]. Finding the highly expressed genes in tumor cells but not in normal ones based on gene expression data is considered a problem that needs to be solved using computational techniques. The high dimensionality of the gene expression data that is associated with a small number of samples revealed other challenges to the use of computational techniques. The used computational techniques include the deep learning methods which are popularly used in computer vision problems [9], [10].

Recently deep learning has emerged and succeeded in machine learning applications because it can read non-linear patterns in a complex feature and can allow the leverage of information from unlabeled data of problems that do not belong to the problem being handled. Motivated by their success, now deep learning has been successfully applied to medical data [11], [12]. Transfer learning, which refers to the use of a model trained on one task to perform classification on another task has been successfully implemented in medical data classification and analysis after the introduction of the state-of-the-art deeper learning neural network models that improve the ability of deep learning substantially [13]. There are many state-of-the-art and on-the-shelf pre-trained models that can be used as a transfer learning approach. These state-of-the-art methods include VGG16 [13], Xception [14], DenseNet [15], and ResNets [16], which are convolutional neural networks (CNN) architectures that are trained on a

very large images dataset. Fine-tuning these architectures when applied to medical data is found to be one of the successful approaches because the characteristics of the medical data are not the same as the data in which these pre-trained models are trained on. In this paper, we compared the classification performances of VGG16, Xception, DenseNet, and ResNets after fine-tuning them to classify the common women cancer using RNA-Seq gene expression data. We first converted the gene expression data into 2D-image like data and then we fed the input convolutional layer of these architectures with these 2D-images like data. The results show that the proposed approach achieved high performance as measure by the accuracy, precision, recall, and F1-score using five-fold cross-validation training and testing approach.

II. Related Work

The methods that used deep learning approach for cancer classification based on gene expression data include the work of Rasool et al [17], Chen et al [18], Liao et al. [19], Kong and Yu [20], Lyu and Haque [21], Sevakula et al. [22], Danaee et al. [23]. Rasool et al. used deep learning and unsupervised features learning to detect cancer and analyses cancer types based on gene expression data. They learned a concise feature representation from unlabeled data using a sparse autoencoder. Chen et al presented a method based on deep learning known as D-GEX, which uses a multi-task multi-layer feedforward neural network to infer the expression of target genes from the expression of landmark genes. In their work, the performances of the deep learning method, Linear regression (LR), and k-nearest neighbor (KNN) regression are evaluated on microarray expression and RNA-Seq profile where they found that their deep learning methods outperform the other methods in terms of accuracy. Liao et al proposed a multi-task deep learning method to solve the few data problem of gene expression by leveraging the gene expression data of multi cancer and learn more representation for cancer that has a small number of cases. This way they enhanced the performance of diagnosing all types of cancer. Kong and Yu integrated external relational features information extracted from RNA-seq gene expression of the breast cancer into a deep neural network architecture using Graph-Embedded Deep Feedforward Networks, which enables the network layers to achieve spares connection and avoid over-fitting. They tuned their model's parameters using a grid search approach. Lyu and Haque converted the rows of the RNA-Seq gene expression data into 2D-images like data and then they trained a convolutional neural network using the obtained images like data for classifying multiple cancer types. Sevakula et al. used sparse autoencoders in combination with feature selection and normalization techniques on gene expression data and then they used a transfer learning procedure on their obtained features. They used the data of some tumor types to improve the features representation when classifying other tumor types. Danaee et al. extracted functional features from the gene expression profile using Stacked Denoising Autoencoder (SDAE) and then they used supervised classification to evaluate the performance of the obtained features to be used for cancer detection and identification. Also, they analyzed the SDAE connectivity matrices to identify a set of highly interactive genes.

III. Material and Methods

A. Dataset

Five RNASeq gene expression profile for different types of women cancers were downloaded from the genomic data commons (GDC) data portal. These types of cancers include breast (BRCA), ovarian (OV), colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD), and thyroid (THCA) cancer. We used TCGAbiolinks package in R to download these RNASeq gene expressions profile [24]. TCGAbiolinks has GDCquery function which uses GDC API to search and download the data and it has many arguments such as project, legacy, data.category, platform, data.type, experimental.strategy, sample.type, and workflow.type. These arguments are normally passed to the GDCquery to filter and determine the type of data that should be downloaded. The project argument determines a valid TCGA project data list that should be downloaded. Five project codes corresponding to our five types of cancer, which are TCGA-BRCA, TCGA-OV, TCGA-COAD, TCGA-LUAD, and TCGA-THCA were used as project argument. The legacy parameter is adjusted to "true", to get the unmodified data in the GDC data portal that is stored in the legacy repository. Consequently, to quantifying the gene expression data and to filter the data to be downloaded we adjusted data.type variable to "Gene expression quantification" and data.category has been set to "Gene expression". We used the data produced using the "Illumina HiSeq" platform. The file.type argument is set to "results" to filtering the legacy database, and since we are looking for counts data "RNA-Seq" protocol that was used to perform the laboratory analysis was chosen as experimental.strategy parameter to obtain the expression profiles. In this work, we are interested in the tumor samples only thus, "Primary solid Tumor" adjusted as sample.type argument to filter out the normal samples. The downloaded data is in a form of a matrix, where the columns represent the samples and the rows contain the genes, i.e. features (equivalently covariates). The five types of cancers have 2166 samples, along with 19947 common genes. To reduce the number of the genes, we constructed a symmetric square matrix of Spearman correlation known as Array-Array Intensity Correlation (AAIC) between samples to determine the highly correlated genes. The visualization of this matrix is shown in Fig. 1, where high correlated genes are depicted in dark color. A correlation cut off equal to 0.6 is used to remove the highly correlated genes. To ensure that we can infer the level of expression correctly without biases, we applied a normalization process on the obtained gene expression profile using TCGAanalyze_Normalization function [25]–[28]. Finally, the gene expression profile is filtered by selecting mean values higher than 0.25 across all samples. The final obtained gene expression profile after applying these preprocessing steps has 2166 samples with 14899 genes. The number of samples in each cancer type is as follows BRCA (1082), COAD (135), LUAD(275), OV (304), and THCA (370). These samples are transformed into 2D-images like data to be suitable for the convolutional layer of CNN architecture. The motivation to convert the data into 2D-images comes from many researches works e.g [3], [29].

To capture the linear and non-linear dependencies, we visualized our final obtained data in two-dimensional space using Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Entities (t-SNE). PCA and t-SNE are

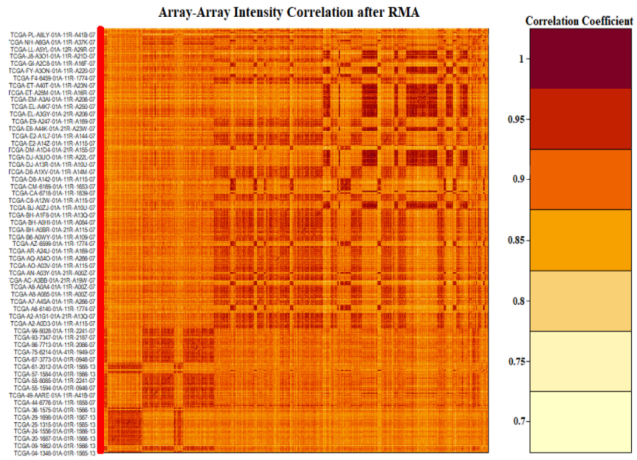


Fig. 1. Symmetric Square Matrix of Spearman Correlation or AAIC Visualization.

Fig. 3. Visualizing the Gene Expression Data using t_SNE.

linear and nonlinear projection methods, respectively. These two methods are used to capture the linear and non-linear dependencies. The obtained projection is depicted in Fig. 2 and Fig. 3. From the two figures, it is clear that the cancers types are overlapped in both the linear and non-linear projection.

Fig. 2. Visualizing the Gene Expression Data using PCA.

B. Problem Formulation

In this paper, we cast common women cancers classification based on gene expression data as a multiclass classification problem. The gene expressions for all the cases are transformed into 2D-images $X = (x_1, x_2, \ldots . x_N)$ that are associated with a ground truth class label $Y = (y_1, y_2, \ldots y_N)$. We are intending to develop a classification function $X \rightarrow Y$. The developed classification function should minimize a loss function using n training samples. We encoded the labels as a vector $y \in \{0, 1\}^M$, where $M = 5$ (the number of the woman common cancer types). We did an investigation using different loss functions and come up with a conclusion that the loss function that gives the highest performance is the categorical_crossentropy, which can be formulated mathematically as given in equation 1.

$$Loss = -\sum_{i=1}^{outputsize} y_i . log \hat{y_i} \quad (1)$$

Where $\hat{y_i}$, y_i, and outpusize represent i^{th} scaler value, the corresponding target value, and the number of scalar values in the model output, respectively.

Since our data is not large enough to train a CNN model from scratch, we used transfer learning because of its outstanding performance in the computer vision domain in general and in the medical data domain in specific [30]–[35]. We fine-tuned different models as a base model and then added a dense layer to reflect our five-class classification problem. To avoid training over-fitting that can result in a very high training accuracy and a low cross-validation accuracy we used L2-regularization. We compared the classification performance of the following models: ResNet50, DenseNet, Xception, and VGG16.

C. Obtaining the 2D-Images from the Gene Expression Data

We transformed our gene expression data into 2D-images by reshaping them into a square matrix of (123 By 123) to fit the convolutional layer of the used CNN methods. Transforming the gene expression data into 2D-images inspired by the work in [3], [29]. The number of the columns or features in the dataset (14899 genes) is not sufficient to be transformed into 123 by 123 matrix, therefore we appended columns of zeroes to the gene expression data. This kind of modification is normally applied to make the size of the data adjustable to the requirement.

D. Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNN) are deep learning models mostly used for image classification. The connectivity of the neurons in the CNN is similar to that of the animal visual cortex and they have special filters to capture the temporal dependencies in an image features and reduce them into an easier arrangement that can be processed without dropping important features to obtain high classification performance.

A sequence of layers makes the architecture of CNN and each layer can transform one volume of activation to another using a differentiable function. Normally, convolutional, pooling and fully-connected layers are stacked to build a CNN model. CNN takes an image that has n rows, m columns, and 3 color channels ((R, G, and B) as input while considering the special structure of an image into account. CNN uses a convolutional layer with three features map to represent the color channels and f × f local receptors or filter size. The features will be read using a stride. If a stride of 1 is used, the result will be a layer of 3× (m-f+1) × (n-f+1) hidden feature neurons. The convolutional operation, which multiplies the elements of the filter by the element of the image matrix element-wise is used to generate the features map. Sliding the filters across the input image matrix will generate the rest of the features. The mathematical formula of the convolutional operation is given in equation 2.

$$O(i,j) = \sum_{k=1}^{f}(\sum_{l=1}^{f} input(i+k-1, j+l-1)kernel(k,l)) \quad (2)$$

Where i runs from 1 to $m - f + 1$ and j runs from 1 to $n - f + 1$.

E. Transfer Learning

It is very challenging and expensive to acquire medical data and for gene expression datasets the small number of cases and a large number of dimensions can hamper the performance of deep learning significantly. On the other hand, deep learning models require a very large number of data for training to give good classification performance. To overcome this problem, we can use transfer learning to leverage information from other data to understand the distribution of our gene expression data. There are many state-of-the-art and on-the-shelf pre-trained models that can be used as a transfer learning approach. These state-of-the-art methods include VGG16 [13], Xception [14], DenseNet [15], and ResNets [16], which are convolutional neural networks (CNN) architectures that are trained on a very large images dataset. Fine-tuning these architectures, which means re-training them when applied to medical data is found to be one of the successful approaches because the characteristics of the medical data are not the same as the data in which these pre-trained model are trained on.

F. Experimental Setup

After trying many state-of-the-art CNN pre-trained architectures we selected the following models: ResNet50, VGG16, DenseNet, and Xception . These models are considered to be a breakthrough for CNN's progress as they have applied unique deep learning architecture. ResNet50 has 50 layers and is the first to introduce a residue model in CNNs architectures to ease the deeper architectures training and solve the degradation problem, which means that not all architectures are similarly easy to optimize [16]. In ResNet50, instead of learning unreferenced functions, the layers are formulated as learning residual functions with reference to the input layer. VGG16 uses a very small convolutional filter with a very deep architecture. DenseNet is one of the new on-the-shelf pre-trained CNNs for visual object recognition that has a similar architecture to ResNet with some essential differences. In DenseNet, each layer is connected to every other layer in a feed-forward fashion. With its structure, DenseNet reduced the problem of vanishing-gradient, make the feature propagation strong and promote its reuse, and uses a small number of features map, which makes it parameters efficient [36]. Xception is inspired by inception [37], where it replaces inception modules with separable convolutions. In all the architectures, We used softmax activation with the prediction dense layer and adam as optimizer when fitting the models. Also, we used L2 kernel and bias regularization for all the architectures. The categorical cross-entropy error function is used to perform the training where we used a five-fold cross-validation approach. We used 100 epochs for training in each architecture. To randomize the whole learning producers and ovoid over-fitting, we shuffled the training data in each epoch.

G. Performance Measures

Four measures are used to evaluate the different transfer learning architectures. These measures are the classification accuracy, precision, recall, and F1 -score. They are considered among the most frequent measures that are used to evaluate the performance of computational methods on medical data. The accuracy and F1-score are used to evaluate the comprehensive classification performance while precision and recall are used to evaluate the rate of recognition and sensitivity respectively. The mathematical formulas for these measures are as follows:

$$accuracy = \frac{\sum_i m_{ii}}{\sum_{i,j} m_{ij}} \quad (3)$$

$$recall_j = \frac{m_{jj}}{\sum_i m_{ji}} \quad (4)$$

$$precision_j = \frac{m_{jj}}{\sum_j m_{ji}} \quad (5)$$

$$F1 - score_j = \frac{2 \times recall_j \times precision_j}{recall_j + precision_j} \quad (6)$$

i and j stand for the different classes

IV. Results and Discussion

In this study, experiments are conducted to classify the five common women cancers: breast, ovarian, colon adenocarcinoma, lung adenocarcinoma, and thyroid cancer. As stated in the methodology, we used five-fold cross-validation, which is a very useful and rigorous validation method for estimating the performance of the classification model, especially with a small dataset. In the five-fold cross-validation approach, the training dataset is divided into five equal sets, four of these sets are used as training and the fifth one is used as a testing set. This process is repeated five times by removing one set to represent the testing set. We used a fine-tunned transfer learning approach in which we tried different architectures as a base model. We tried different activation functions in the prediction dense layer and different optimizer when fitting the model. From the results, we found that the softmax activation

function with the adam optimizer obtained the best results in all the used architectures. The architectures that obtained the best results are ResNet50, DenseNet, Xception, and VGG16. The performances of ResNet50, DenseNet, Xception, and VGG16 have been evaluated for each fold. The final classification performance is calculated as the average of the results of the five testing sets. Table I shows that Xception model has the highest performance in terms of precision, recall, F1-Score, and accuracy compared to DenseNet, ResNet50 , and VGG16 models. The three architectures are trained for 100 epochs. Fig. 4, Fig. 5, and Fig. 6 show the validation accuracy, the validation Loss, and the F-measure graphs respectively for the first fold of the Xception architecture.

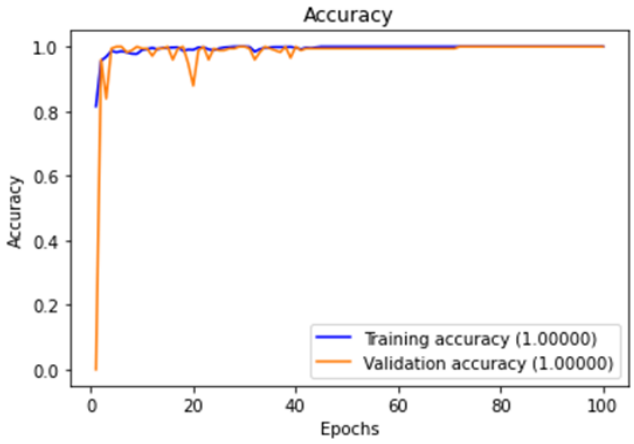


Fig. 4. Training Accuracy Curve of Xception Architecture.

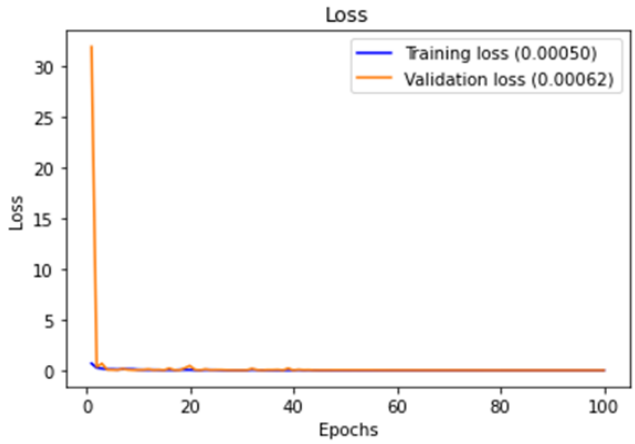


Fig. 5. Training Loss Curve of Xception Architecture.

Schematic Xception architecture diagram for cancer multi-class classification using transfer learning is shown in Fig. 7, where an overview of the layers that comprise the architecture of the base Xception architecture and the layers that we added are depicted. Different colors are used to depict the different layers.

Table I shows that fine-tuned Xception architecture achieved classification accuracy = 98.6%, precision = 98.6%,

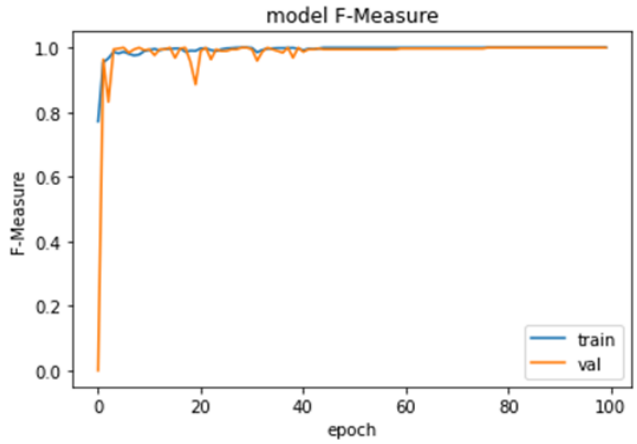


Fig. 6. F1-measure Curve of Xception Architecture.

TABLE I. CLASSIFICATION PERFORMANCES OF THE FINE-TUNED XCEPTION, DENSENET, RESNET50, AND VGG16 ARCHITECTURES.

Classification Method	Folds	Performance Metrics (%)			
		Precision	Recall	F1-Score	Accuracy
Xception	Fold1	98	97	97	99
	Fold2	98	96	97	97
	Fold3	99	99	99	99
	Fold4	99	98	98	99
	Fold5	99	99	99	99
	Average	**98.6**	**97.8**	**98**	**98.6**
DenseNet	Fold1	98	97	98	98
	Fold2	98	97	98	98
	Fold3	98	98	98	98
	Fold4	98	97	97	98
	Fold5	97	97	97	98
	Average	**97.8**	**97.2**	**97.6**	**98.0**
ResNet50	Fold1	95	96	96	97
	Fold2	98	97	98	98
	Fold3	99	99	99	99
	Fold4	98	96	97	97
	Fold5	94	94	94	96
	Average	**96.8**	**96.4**	**96.8**	**97.4**
VGG16	Fold1	94	93	93	95
	Fold2	90	85	86	90
	Fold3	96	95	96	97
	Fold4	94	92	93	95
	Fold5	97	96	96	97
	Average	**94.2**	**92.2**	**92.8**	**94.8**

recall = 97.8%, and F1-score = 98% on five-fold cross validation training and testing approach. The table shows that Xception achieved the highest performance in all the performance measures. DenseNet architecture follows Xception in terms of performance. It achieved classification accuracy = 98%, precision = 97.8%, recall = 97.2%, and F1-score = 97.6% on a test set. ResNet50 achieved classification accuracy =97.4%, precision = 96.8%, recall =96.4%, and F1-score = 96.8% on a test set. The lowest performance is achieved when using VGG16, which obtained classification accuracy = 94.8%, precision = 94.2%, recall = 92.2%, and F1-score = 92.8%.

The confusion matrices for the five folds of Xception are shown in Figure 8. Figure 8 also shows the overlapped confusion matrix, which is calculated as the summation of the five folds convolution matrices to reflect the general performance of the Xception model. The overlapped confusion matrix shows that Xception model classified THCA, OV, BRCA, and COAD better than LUAD cancer type in the multi-class classification

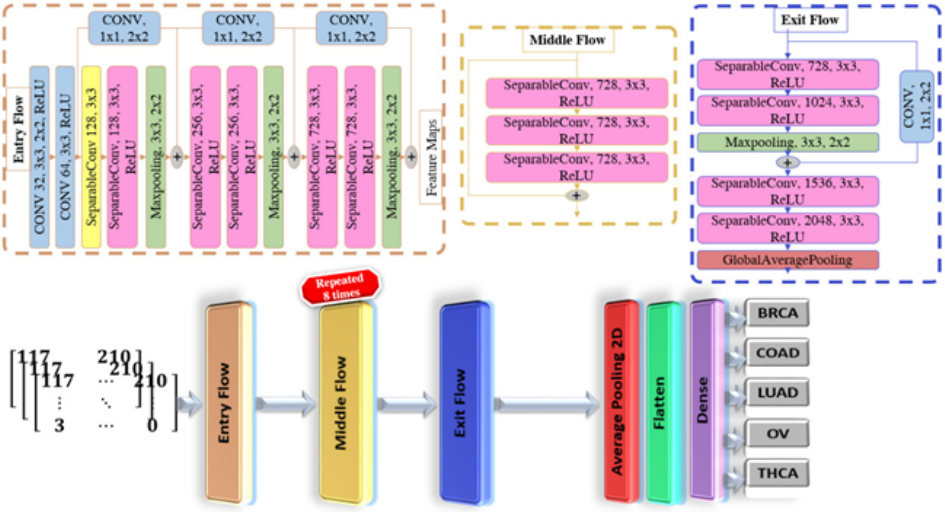


Fig. 7. Schematic Xception Architecture Diagram for Cancer Multiclass Classification using Transfer Learning.

task. This is because the dataset is imbalanced and the classifier does not have an equal number of instances for all the classes during training time.

V. Conclusions

In this paper, we used the fine-tuning transfer learning approach on RNA-Seq gene expression data to classify five cancer types that mostly affect women. These five types are breast (BRCA), ovarian (OV), colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD), and thyroid cancer (THCA). The RNA-Seq gene expression data for the five cancer types is downloaded from genomic data commons (GDC) data portal using TCGAbiolinks package in R. The downloaded data is in a form of a matrix, where the columns represent the samples and the rows contain the genes. The five types of cancers have 2166 samples, along with 19947 common genes. We used Spearman correlation to reduce the number of the genes by removing the highly correlated genes using correlation cut off equal to 0.6. To ensure that we can infer the level of expression correctly without biases, we applied a normalization process on the obtained gene expression profile using TCGAanalyze Normalization function. Finally, the gene expression profile is filtered by selecting mean values higher than 0.25 across all samples. The final obtained gene expression profile after applying these preprocessing steps has 2166 samples with 14899 genes. These samples are transformed into 2D-images like data to be suitable for the convolutional layer of CNN architecture. We fine-tuned four pre-trained models on the RNA-Seq gene expressing data, namely, ResNet50, DenseNet, Xception, and VGG16. Xception architecture shows the highest performance where it achieved classification accuracy = 98.6%, precision = 98.6%, recall = 97.8%, and F1-score = 98% on five-fold cross-validation training and testing approach.

Acknowledgment

The authors would like to thank Jouf University for all the support it provides

References

[1] J. Adams, "Transcriptome: connecting the genome to gene function," *Nat Educ*, vol. 1, no. 1, p. 195, 2008.

[2] M. C. Frith, M. Pheasant, and J. S. Mattick, "The amazing complexity of the human transcriptome.," *European journal of human genetics: EJHG*, vol. 13, no. 8, p. 894, 2005.

[3] M. K. Elbashir, M. Ezz, M. Mohammed, and S. S. Saloum, "Lightweight convolutional neural network for breast cancer classification using rna-seq gene expression data," *IEEE Access*, vol. 7, pp. 185338–185348, 2019.

[4] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[5] M. Maienschein-Cline, J. Zhou, K. P. White, R. Sciammas, and A. R. Dinner, "Discovering transcription factor regulatory targets using gene expression and binding data," *Bioinformatics*, vol. 28, no. 2, pp. 206–213, 2012.

[6] E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, *et al.*, "An integrative genomics approach to infer causal associations between gene expression and disease," *Nature genetics*, vol. 37, no. 7, pp. 710–717, 2005.

[7] J. Krammer, K. Pinker-Domenig, M. E. Robson, M. Gönen, B. Bernard-Davila, E. A. Morris, D. A. Mangino, and M. S. Jochelson, "Breast cancer detection and tumor characteristics in brca1 and brca2 mutation carriers," *Breast cancer research and treatment*, vol. 163, no. 3, pp. 565–571, 2017.

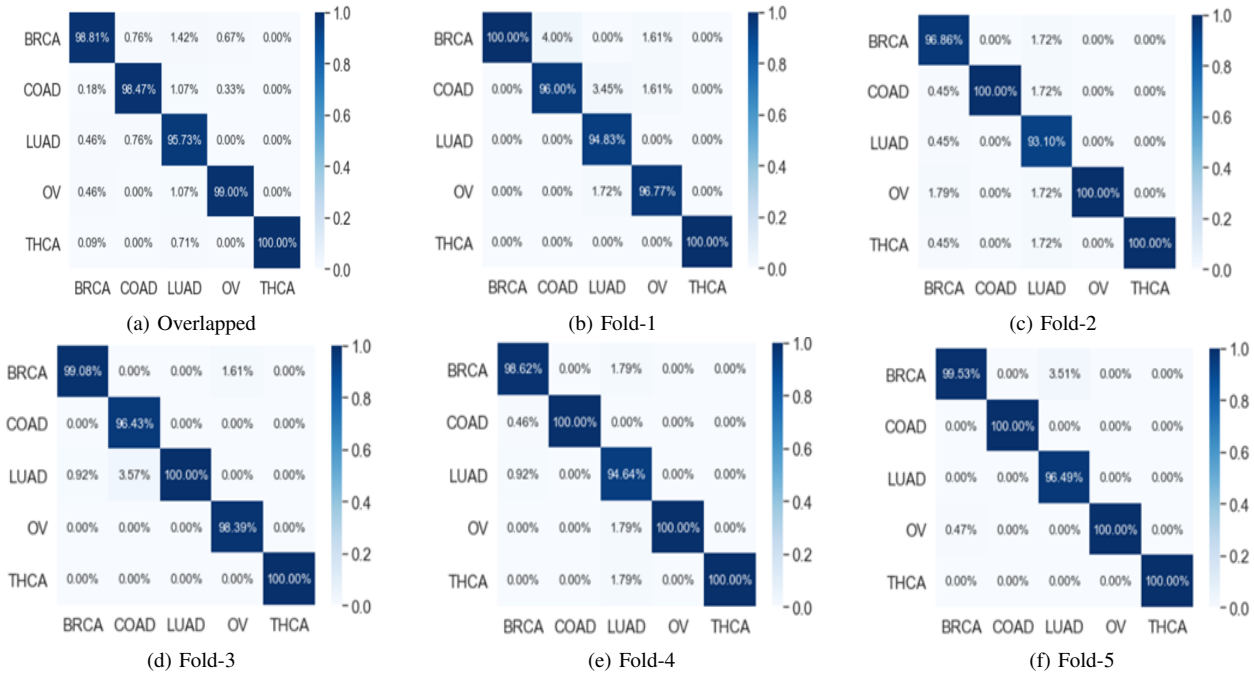
[8] M. Mohammed, H. Mwambi, B. Omolo, and M. K. Elbashir, "Using stacking ensemble for microarray-based cancer classification," in *2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, pp. 1–8, IEEE, 2018.

[9] H. Pan, B. Wang, and H. Jiang, "Deep learning for object saliency detection and image segmentation," *arXiv preprint arXiv:1505.01173*, 2015.

[10] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: a survey," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 84–100, 2018.

[11] H. Jia, Y. Xia, Y. Song, W. Cai, M. Fulham, and D. D. Feng, "Atlas registration and ensemble deep convolutional neural network-based prostate segmentation using magnetic resonance imaging," *Neurocomputing*, vol. 275, pp. 1358–1369, 2018.

[12] J. Zhang, Y. Xia, Y. Xie, M. Fulham, and D. D. Feng, "Classification of medical images in the biomedical literature by jointly using deep and handcrafted visual features," *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1521–1530, 2017.



(a) Overlapped (b) Fold-1 (c) Fold-2

(d) Fold-3 (e) Fold-4 (f) Fold-5

Fig. 8. The Overlapped and 5-fold Confusion Matrix Results of the Multiclass Classification Task: (a) Overlapped Confusion Matrix, (b) Fold-1 CM, (c) Fold-2 CM, (d) Fold-3 CM, (e) Fold-4 CM, and (f) Fold-5 CM.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[14] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[17] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, "Using deep learning to enhance cancer diagnosis and classification," in *Proceedings of the international conference on machine learning*, vol. 28, ACM New York, USA, 2013.

[18] Y. Chen, Y. Li, R. Narayan, A. Subramanian, and X. Xie, "Gene expression inference with deep learning," *Bioinformatics*, vol. 32, no. 12, pp. 1832–1839, 2016.

[19] Q. Liao, Y. Ding, Z. L. Jiang, X. Wang, C. Zhang, and Q. Zhang, "Multi-task deep convolutional neural network for cancer diagnosis," *Neurocomputing*, vol. 348, pp. 66–73, 2019.

[20] Y. Kong and T. Yu, "A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data," *Bioinformatics*, vol. 34, no. 21, pp. 3727–3737, 2018.

[21] B. Lyu and A. Haque, "Deep learning based tumor type classification using gene expression data," in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pp. 89–96, 2018.

[22] R. K. Sevakula, V. Singh, N. K. Verma, C. Kumar, and Y. Cui, "Transfer learning for molecular cancer classification using deep neural networks," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 16, no. 6, pp. 2089–2100, 2018.

[23] P. Danaee, R. Ghaeini, and D. A. Hendrix, "A deep learning approach for cancer detection and relevant gene identification," in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017*, pp. 219–229, World Scientific, 2017.

[24] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, *et al.*, "Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data," *Nucleic acids research*, vol. 44, no. 8, pp. e71–e71, 2016.

[25] D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit, "Gc-content normalization for rna-seq data," *BMC bioinformatics*, vol. 12, no. 1, p. 480, 2011.

[26] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments," *BMC bioinformatics*, vol. 11, no. 1, pp. 1–13, 2010.

[27] K. D. Hansen, R. A. Irizarry, and Z. Wu, "Removing technical variability in rna-seq data using conditional quantile normalization," *Biostatistics*, vol. 13, no. 2, pp. 204–216, 2012.

[28] W. Zheng, L. M. Chung, and H. Zhao, "Bias detection and correction in rna-sequencing data," *BMC bioinformatics*, vol. 12, no. 1, p. 290, 2011.

[29] B. Lyu and A. Haque, "Deep learning based tumor type classification using gene expression data," in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pp. 89–96, 2018.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[31] X. Liu, C. Wang, J. Bai, and G. Liao, "Fine-tuning pre-trained convolutional neural networks for gastric precancerous disease classification on magnification narrow-band imaging images," *Neurocomputing*, vol. 392, pp. 253–267, 2020.

[32] T. Tan, Z. Li, H. Liu, F. G. Zanjani, Q. Ouyang, Y. Tang, Z. Hu, and Q. Li, "Optimize transfer learning for lung diseases in bronchoscopy using a new concept: sequential fine-tuning," *IEEE journal of translational engineering in health and medicine*, vol. 6, pp. 1–8, 2018.

[33] G. Wimmer, A. Vécsei, and A. Uhl, "Cnn transfer learning for the automated diagnosis of celiac disease," in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6, IEEE, 2016.

[34] H. G. Kim, Y. Choi, and Y. M. Ro, "Modality-bridge transfer learning for medical image classification," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–5, IEEE, 2017.

[35] E. Ribeiro, A. Uhl, G. Wimmer, and M. Häfner, "Transfer learning for colonic polyp classification using off-the-shelf cnn features," in *International Workshop on Computer-Assisted and Robotic Endoscopy*, pp. 1–13, Springer, 2016.

[36] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

PlusApps: Towards a Privacy Risk Analysis for Android Plus Applications

Abdullah J. Alzahrani
Computer Engineering Department
College of Computer Science and Engineering
University of Ha'il, Hail, Saudi Arabia

***Abstract*—The Android platform leads the mobile operating system marketplace and subsequently has drawn the interest of malware authors and researchers. The significant number of proposed malware detection techniques, classification models and practical reverse engineering solutions are insufficient and there is a lack of perfection. Also, the number of Android apps has increased significantly in recent years, as has the number of apps revealing confidential data. It is essential to investigate the applications and make sure that none of them are leaking privacy data, and consequently a privacy leak analysis approach is needed. Therefore, this paper investigates plus apps behavior and data leakages with a machine-learning algorithm to determine the best features for differentiating plus apps from original apps. The result of the analysis discloses that the SVM classifier presents the greatest accuracy. Further investigation demonstrates that the classifier with the ranking algorithm that uses correlation coefficient (CorEvel) and information gain (InfGain) methods offers more exceptional precision than the other correlation algorithms. The result of this experiment proves that the ranking algorithm is able to decrease the dimension of features and produce an accuracy of 96.60%.**

***Keywords*—Android security; malware detection; permission analysis; privacy risk; plus application**

I. Introduction

Operating systems (OSes) such as Android used in various applications today can be prone to certain risks. Since its inception, Android has come a long way and is prevalently used today. Android has changed the world of smartphones. However, certain risks, such as privacy leakage, can influence the use of Android applications. Android applications can be attacked by malicious codes. According to Zhang et al. [1], malicious code is the "general term used for various hostile or intrusive software, such as viruses, worms, Trojans, spyware, botnets, Rootkits, and backdoors, among others". Malicious code can steal information and essential data from computer users, which greatly affects a user's privacy. When malicious code gets access to a user's personal data and computer, controlling illegal computer systems and cyber source may be possible. In this case, the computer and network credibility, integrity and availability can be destroyed. Meanwhile, Casey [2] noted that malicious code is usually not created using a robust software development lifecycle, with the essential testing and assessment phases needed to work out bugs. Because of this, attackers can crash a target application or OS, which then serves as a warning that a security issue exists. An administrator could not treat every OS or application error as an attack, but certain characteristics that a security issue exists can be looked into. Some of these include crash outcomes after opening an e-mail attachment and after viewing a certain web page in a web browser. There are several diverse Android applications that can be accessed in different parts of the world. But in this variety, there are those that are considered the most popular based on the number of downloads. The top five of the top 20 list of Android applications presented by Price [3] are shown in Table I. A plus application is an APK used to modify the features of the original app for Android. WhatsApp application versions will be investigated, both the plus and the original version from its existence.

TABLE I. The 5 Most Popular Android Apps in the Google Play Store.

No.	Android Application	Description	No. of Downloads
1.	WhatsApp	Instant messaging tool, under the ownership of Facebook since 2014	5.875 billion
2.	Facebook	Popular social networking site	5.478 billion
3.	Facebook Messenger	Instant messaging	3.756 billion
4.	Instagram	Photo and video-sharing social networking service	2.796 billion
5.	Subway Surfers	Game application	1.249 billion

As mentioned earlier, smartphones can be prone to leakage of a user's sensitive and personal data. Data leakage is known to be a serious threat to individuals and enterprise operations, as loss of sensitive information can result in significant reputational damage and financial losses and can be detrimental to the long-term stability of an organization [4]. Apparently, the concept of user awareness in the case of smartphone leakage does not only cover the technical aspects of the device, such as the functioning of Android applications. Alsaleh et al. [5] underscored that both human and technological factors are involved in the multi-dimensional problem of security threats in smartphones. There are also social factors representing the users in terms of user awareness and behavior for securing their smartphone and smartphone applications. For example, Alsaleh et al. [5] found that some smartphone users still chose to share their private data via instant messaging platforms, such as WhatsApp, even though they were aware of the privacy risk. This implies that they accepted the risks associated with their less protected sharing habits as the services offered by these platforms allow them to easily communicate with people they know. This implies that the smartphone's convenience and usefulness, alongside its features and applications, can sometimes affect the decision to behave in a risky manner. There is a lack of user understanding of privacy and security risks linked with installing smartphone applications. In a conducted survey, only 17% of users paid attention to the permissions in their applications, including those that grant

application access to the privacy-sensitive, particularly when an application is being installed. The results showed that only 3% of the survey respondents had a full understanding of the permissions screen [6].

Malicious code can wreak havoc on IoT and mobile devices [7]. The functioning of installed applications in mobile devices will also be negatively affected in the form of data breaches or privacy leakage. For example, data leakage can bring serious threats to organizations. Meanwhile, data leakages are privacy-sensitive data transmitted off the smartphone, through the applications in an unexpected manner. It is therefore important for the malware to be detected during leakage to avoid further or extreme damage. However, it is challenging to detect an application as malware, especially when information leakage occurs [7]. In addition, Cheng et al. [4] stressed that detecting internal data leaks is very challenging as the internal breaches typically involve users having legitimate access to the facilities and data. Doing such actions may also be successful without a trace, as the perpetrators are already knowledgeable about the organization and know how to bypass detection. The prevalent use of the Android OS and the fact that Android applications are often downloaded from third party sources makes it essential to accurately detect those that can be malicious [8]. Additionally, the popularity of Android applications opens the door to several threats and risks from malware applications. According to Singh et al. [9], these simultaneously increasing mobile malware apps can perform malicious activities, such as misusing the private information of users when sending messages and accessing their contacts and other information. Apart from these, confidential information stored in mobile devices can also be illegally exploited. Because of these threats, malware classification and identification becomes a crucial issue. However, permission mechanisms can still be considered great defense mechanisms in ensuring that certain applications cannot harm the user data. Because of this, Singh et al. [9] proposed that malware characterization is determined from the manifest file, allowing the user to enhance the efficiency of Android permissions. In this way, the user will be informed of the risks of Android permissions and applications.

Android's features somehow offer threats and risks to users. According to Fang et al. [10], Android security has been established in a "permission-based mechanism" that restricts the access of third-party Android applications to the critical resources on an Android device. However, the permission-based mechanism has been widely criticized, because of the "coarse-grained control of application permissions" as well as "difficult management of permissions by developers, marketers and even the end users" [10]. Some issues arising in Android security are incompetent administration, coarse granularity of permissions, insufficient permission documentation, incompetent permission administration, permission escalation attack, over-claim permissions and TOCTOU (Time of Check to Time of Use) attacks. Other approaches and initiatives concerning Android permissions were also studied [11].

The purpose of this research is to explore the privacy risks in Android plus applications. To overcome the issues surrounding privacy data leakage, a classification model of plus and original apps needs to be employed. This model examines all permissions used in Android plus apps via ranking algorithms and machine-learning approaches. It obtains features from the apps by analyzing plus and original apps and producing the best feature sets that are employed in the classification models. These models improve assessment of spotting data leakages of plus apps as these apps will be observed corresponding to these features sets. This paper presents an analysis of Android application behavior and data leakages using a variety of classification models. It utilizes feature selection techniques that need to nominate attributes that are engaged to construct the classification model to predict unknown samples. The proposed model examines the apps, differentiating plus apps from original apps, and defines the risk level of the apps using a ranking algorithm that uses correlation coefficient (CorEvel) and information gain (InfGain) methods, offering superior exceptional precision to other correlation algorithms.

The rest of this paper is organized as follows: in Section II, overview of Android platform and security; in Section III, the related work is presented; in Section IV, Android plus apps privacy risk analysis is explained; in Section V, the experiment and result is illustrated; in Section VI, the conclusion and future work are summarized.

II. Overview of Android Platform and Security

The Android Platform is defined as the platform for mobile devices that uses a modified Linux kernel and was introduced by the Open Handset Alliance. Applications running on the Android platform are written in Java programming language. The Java classes are compiled into what is known as "Dalvik Executables" and are operated on the "Dalvik Virtual Machine". Although Android is considered as an open development, it is not open for anyone to contribute, especially when a certain version is under development. All of these are undertaken behind closed doors in the Google office. A developer would need the Android SDK in order to create an application for the platform and this would include tools and APIs. The SDK will also be integrated into the graphical user IDEs (Integrated Development Environments) [12].

Android has developed several security mechanisms. Elovici et al. [13] noted that the Android software stack is established on the "Linux kernel" that is utilized for device drivers, memory management, process management and even networking. This is followed by the next level called the "Android native libraries", where several system components in the upper layers are using the said libraries. The libraries are incorporated into Android applications, which can be made possible through Java native interfaces. This is then followed by the Android "runtime" level, which is composed of the "Dalvik virtual machine and the core libraries" (Elovici et al. [13]). These core libraries are written in Java, while also offering substantial subsets of the Java 5 SE packages and some Android-specific libraries. The "application framework layer" is also fully written in Java and covers the Google tools and propriety tools extensions and services. The phone, web browser and email client, among others, are considered as the topmost application layer. Figure 1 shows the list of the security mechanisms embedded in Android.

Indeed, the security system of Android uses the Linux kernel and offers a set of security measures. It also permits a user-based permissions model, process isolation, secure IPC mechanisms and the ability to eradicate unnecessary or

Mechanism	Description	Security issue
Linux mechanisms		
POSIX users	Each application is associated with a different user ID (or UID).	Prevents one application from disturbing another
File access	The application's directory is only available to the application.	Prevents one application from accessing another's files
Environmental features		
Memory management unit (MMU)	Each process is running in its own address space.	Prevents privilege escalation, information disclosure, and denial of service
Type safety	Type safety enforces variable content to adhere to a specific format, both in compiling time and runtime.	Prevents buffer overflows and stack smashing
Mobile carrier security features	Smart phones use SIM cards to authenticate and authorize user identity.	Prevents phone call theft
Android-specific mechanisms		
Application permissions	Each application declares which permission it requires at install time.	Limits application abilities to perform malicious behavior
Component encapsulation	Each component in an application (such as an activity or service) has a visibility level that regulates access to it from other applications (for example, binding to a service).	Prevents one application from disturbing another, or accessing private components or APIs
Signing applications	The developer signs application `.apk` files, and the package manager verifies them.	Matches and verifies that two applications are from the same source
Dalvik virtual machine	Each application runs in its own virtual machine.	Prevents buffer overflows, remote code execution, and stack smashing

Fig. 1. Security Mechanisms Embedded in Android [13]

possibly insecure parts of the kernel in the operating system. Also, it can go further and render additional efforts in order to avert or prevent multiple system users from accessing and eventually exhausting each other's resources [14]. Fig. 2 shows and summarizes the five key security features of Android that should be carefully studied.

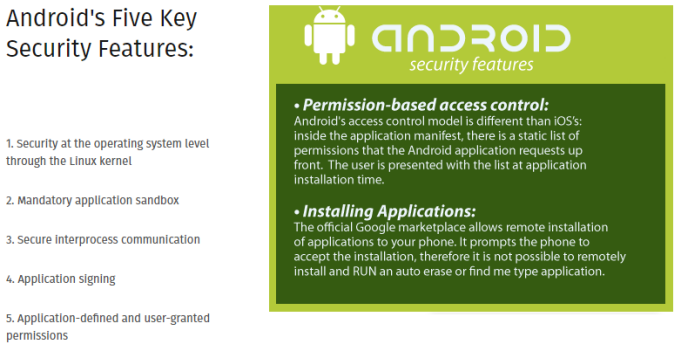


Fig. 2. Five Security Features of Android [14]

Data mining and machine learning are important measures for the security of Android. According to Dua et al. [15], both data mining and machine learning can offer unified reference for a certain machine learning solution towards cyber security issues. It can also supply a foundation for cyber security fundamentals and assess new challenges that detail the cutting-edge machine learning and data mining techniques. It is important for the managers to learn about these while considering the different challenges in data mining and machine learning for security. As Ahmad et al. [16] underscored, a large number of industries are already dependent on network connections, especially those that have sensitive business trading and security matters. In this case, communications and networks are extremely vulnerable to the challenges and threats or risks, such as hacking. Data mining security therefore needs to be applied.

III. Related Work

With the growth in worldwide sales of smartphones, there has been a significant increase in the number of malicious applications that misuse private data without a user's knowledge, publishing it on the online market. Given the massive evolution of the malware, security researchers are required to analyze smartphone applications to identify the intent of the software and to develop defense mechanisms. Previously, application disassemble was achieved by employing tools such as decompilers and runtime debuggers. These techniques require significant amounts of time and are capable of causes errors, depending on the proficiency of the analyst. In the fact an automatic analysis model [17] examines the downloaded applications without human involvement. A key technique in automatic analysis is performing reverse engineering on the application's disassembling, smali code, code decryption, pattern matching, static system call analysis, and Application Programming Interface (API) calls. Meanwhile, malicious code developers are improving their coding skills to find new ways for the malware to evade the detection techniques [18].

Android sensitive data leaks have recently been drawing attention. PlusApps seems to be the first to systematically study a technique to understand users who installed plus apps without knowing its bad side effects. All current Android privacy leakage detection techniques merely identify privacy leakage. Static Taint Analysis (STA) [19] [20] [21] aims to discover the potential sensitive data leaks with the support of deep analysis and program debugging. On the other hand, these methods generally present false positives and are unable to distinguish between user-intended and unintended operations because user intention and context information is absent. On the other side, Dynamic Taint Analysis (DTA) [22] monitors the sensitive data at runtime by using profiling code instrumentation to the original app code. This technique cannot be employed to automatically identify leaking sensitive data in application markets for the reason that privacy leakages are reported while the apps are executed, and dangerous propagation happens.

BLADE [23] identifies malware downloaded from the web by knowing whether it has the user's permission or not. Nevertheless, smartphone applications usually do not require end-user license agreements (EULA) or warnings, even though the user approves the data (e.g. text forwarding). Pegasus [24] spots abnormal behaviors that can be described as APIs and permissions of applications using the historical order and, similar to this paper, it concentrates on identifying malicious application behaviors that are not consistent with the GUI events. However, sensitive data leakages cannot be shaped as app usage of permissions or APIs, therefore several sensitive data leakages cannot be discovered by such techniques. Moreover, Pegasus validates application behaviors based on application-specific properties, which are complicated to indicate with no understanding of application code. Lately, VetDroid [25] improves Dynamic Taint Analysis by creating requirements for sensitive processes. However, the requirement primarily concentrates on the application rationality, not observing each function and the trigger condition for it.

AppIntent [26] investigates user-intended sensitive data transmission on the Android platform. Woodpecker [27] analyzes potential leakages which dissect the reachability of a critical permission from a public, unprotected interface.

Yajin et al. [18] presented inactive content leaks that altered applications to passively reveal application data. However, it does not examine system calls into the Android platform itself. Felt et. al [6] presented privacy and security risk techniques. The researchers conducted an Internet survey on 308 Android users and a laboratory study of 25 Android users. They found that only 17% of users pay attention to the permissions in the applications when an application is installed. Oglaza et al. [11] focused on the Identity Based Access Control (IBAC) models, which were used as permissions-management solutions on mobile devices. The researchers considered the results of a survey from Google in 2013 which showed that French users have on average 32 applications on their Android smartphones. The users would have to manage hundreds of permissions in order to protect their privacy. Apparently, IBAC can be complex aside from its scalability issues.

Some researchers approached data leakage detection by different methods from information flow analysis. Bayes-Droid [28] performs privacy enforcement by examining a comparison between the sensitive data with sinking-values. It can identify sensitive data leakage more precisely than taint analysis techniques, although it only focuses on EIFs and is not applicable to IIF detection. AGRIGENTO [29] utilizes a black-box variance analysis method to show data leakage exposure for obfuscated apps. Given that it only examines sources of sensitive data and network traffic, it cannot maintain locations of IIFs. In addition, it is not an efficient technique for privacy policy enforcement. Barbon et al. [30] uses a hybrid of data flow and quantitative assessment methods to identify IIFs. DAPA [31] also presented a method founded on abstract interpretation framework. Nevertheless, these methods are not relevant to IIFs other than control reliance.

Various researchers have studied permissions-based models, however PlusApp is focused more on the plus apps' behavior and data leakages. The proposed system has the ability to differentiate plus from original apps based on ranking methods and the application's misbehavior in using risky permissions techniques. This model employs Correlation Coefficient (CorEvel) and information gain (InfGain) methods that have the ability to rank the permissions.

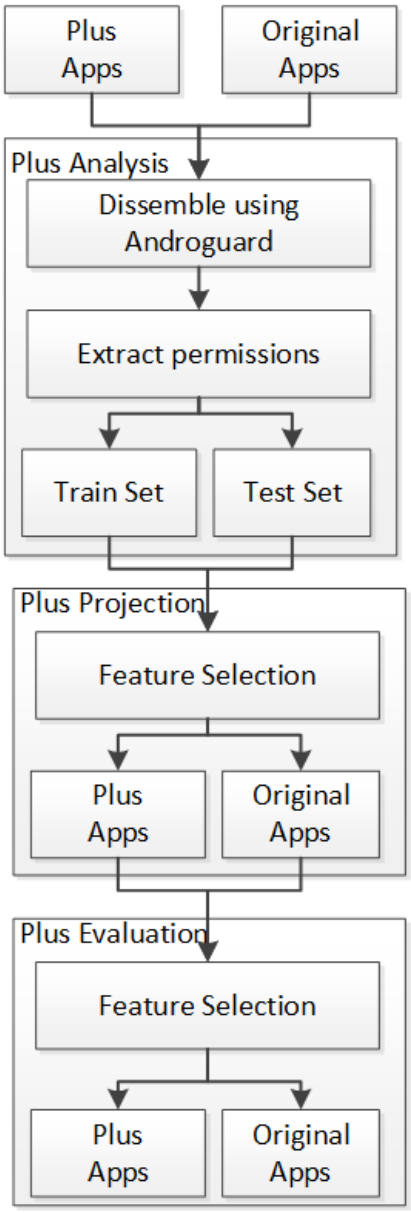


Fig. 3. PlusApps Model Architecture and Components.

IV. Proposed System

The contribution of this research is to examine how permissions are abused by attackers to steal data or damage the mobile device. This paper analyzes these permissions after removing permissions with the normal attribute to see if the proposed system could distinguish between plus and normal apps concerning these permissions. The approach of this research has four phases: plus data collection, plus apps analysis, plus APK projection and plus evaluation. Fig. 3 illustrates the entire workflow of the proposed approach with the corresponding components.

A. Data Collection

In the data collection phase, original and plus app samples are collected that need to be analyzed to be adequate for machine learning approaches. The data creation process began with the data cleansing that was performed to eliminate identical apps and decrease the data amount. Next was the disassembling of the Android applications, making use of reverse engineering tools to retrieve the source code of the apps. Finally the code was analyzed to get the used permissions and the permission occurrences. The first stage, the data cleansing, involves having unique samples and decreasing the features set to accelerate the training phase with highly accurate results. One substantial standpoint that effects data quality is data replication, which impacts the data mining outcomes, and hash techniques were used to identify the identical apps and label these apps by the MD5 hash names. Eliminating replicated data is required in order to have precise and reliable data because inappropriate features have a harmful impact on machine learning [32]. Consequently, replicated samples are excluded from the dataset. There were no duplicate samples found in the original apps, however, many plus samples were duplicated. The dataset was prepared using 454 samples of unique plus

applications. Also, 1,000 samples were downloaded from the Uptodown store. The original samples were separated such that 300 apps were included in the test set and the other 700 APK files assigned to the train set. From the 454 plus APK samples, 317 files were included in the test set and the remaining 137 files assigned to the train set.

B. Plus Apps Analysis

This model involves static analysis for plus application classification that implements machine learning techniques. It analyzes APK applications before executing to determine whether they are harmful applications or not. Many features that can be used towards distinct feature classes are obtained and employed to optimize the feature space. Androguard [33] is utilized for extracting features needed for recognizing plus and original apps. Plus analysis of components consist of two steps: dissembling APK files and feature extraction. An APK file contains all the program's source code, resources, assets, certificates and manifest files. However, the permissions from each APK file are also obtained using the Androguard tool. The distinct features are then used for classification in the first step, followed by the investigation with joint features. The distinct feature model generation requires features that are extracted from distinct categories. Different types of features are used. Permissions: the activities of an Android application must request permission to access sensitive user data. These permissions must be declared statically and a permission-based model is used to offer security for Android architecture. Permission total: this feature set is produced by calculating the set of used permissions by an application.

1) Feature Extraction : A feature in machine learning is a distinct measurable property or characteristic of a phenomenon being observed. This model takes APK files, originally in binary format, as input to the disassembler Androguard tool. One of the output files is the AndroidManifest.xml file that is readable, produced from the input .apk files using Androguard script. The permissions defined in manifest files, which can be extracted, then determines the number of used permissions in each APK file. Also, the amount of permissions used in APK files is a second feature for producing the proposed approach. Since Android APK files and their features have a great importance in Plus Android classification, the basic features are extracted. Moreover, some combined features based on statistical measurements were obtained. The number of dangerous permissions over the total number of permissions requested is an example of the ratio feature.

The total number of permissions of one application and the risk level of each permission are illustrated in Algorithm 1. This algorithm scans all dissembled APK applications and extracts all the features. In addition, it obtains permissions and the occurrence of each permission per application. These permissions are then labeled based on the level of protection and type of risk. These features and their values are saved in a ".csv" file.

C. APK Projection

1) Feature Selection: In this step, feature pruning is conducted to remove the attributes that result in misclassification. After eliminating the inappropriate attributes, joint features to both the modules $(P \cap O)$ are measured. Joint features are given higher priority over the other groups of attributes, such as the union of plus and original features $(P \cup O)$, differentiating original and plus features as they are considered to be irrelevant for the classification of plus samples [34]. Determining the significance of a feature and its attributes is known as feature ranking in machine learning, which has the aim of choosing the most revealing features and refining the performance of learned models [35]. This model utilized correlation Coefficient (CorEvel) and information gain (InfGain) methods

Feature selection is applied to produce the input data into the appropriate size to obtain a subset of k significance features from a set of n features. The nominated attributes are employed to construct the classification model to predict unknown samples. The permissions requested by the applications are classified according to their persistency. These methods are useful to the top 67 common permissions and distinguished 202 features to both plus and original train sets to minimize the feature space.

2) Plus Apps Classifier: From the above phases, the model obtains the requested permissions recognizable with each class label of samples, the total number of the request permissions, and the ratio of risky permissions to the level of risk. To differentiate plus apps from original apps, a classifier is constructed based on selected features that recognize risky applications. The classifier is implemented utilizing four classification algorithms: Support Vector Machine (SVM) [36], Naive Bayes (NB) [36], Decision Tree (J48) [37] and Random Forest (RF) [38], due to their good performance in predicting accuracy.

Support Vector Machine (SVM): SVM tries to find the optimum hyperplane that splits two or more data points from one class on one hand and others on the other hand. In order to separate two classes optimally, the top hyperplane is defined by maximizing margins of both classes. The margins are the distance between the hyperplane and the neighboring point in the classification that can be expressed using the Duality and

Algorithm 1 Total number of permissions of one application and risk level of each permission

Inputs
APKDataset (dissembled APK applications)
Labelled_Risk_Permission

Outputs
FeatureSet (appearance of each permission in Manifest.xml files)

PermiList=[]
$featureSet[] \leftarrow 0$
for e **do**ach Manifest.xml in APKDataset
  $perm \leftarrow 0$
  $currentFeature \leftarrow get\_Read(\text{“}Manfest.xml\text{”})$
  for e **do**ach permission in PermList
    $appearance \leftarrow count\#ofpermissionsin$
    $TypeofRisk \leftarrow compare(permission, Risk_Level)$
    `featureSet[perm] ← featureSet[perm] + appearance`
    $featureSet[perm] \leftarrow featureSet[perm] + perm + TypeofRisk$
    perm++
Return $featureSet[]$

Lagrange Multipliers [39].

Decision Tree (J48): The decision tree can be utilized for unravelling regression and classification issues by learning a training model that that can be used to predict the class with a tree structure [37]. The nodes are the features and leaves that determine class labels. The branches between nodes and leaves are connected with simple decision rules. Predicting a class label for an instance is dependent on the training objects which are all in the root of the tree, and then comparing the root attribute values with the instance's attribute recursively based on selected features. The selection of features is based on an empirical or arithmetic measure. For each recursive stage, the chosen top feature outstandingly decreases the indecision for classification. Thus, the decision tree algorithm naturally holds the function of feature selection.

Random Forest: Random Forest (RF) is an ensemble knowledge method that can be independently learned from a set of decision trees on reduced training sets [38]. To get improved predictive performance, a reduced training set is shaped by arbitrarily sampling with replacement of features. The last choice of classification is completed by choosing between all learned trees. This method performs better than a single tree on classification accuracy.

Naive Bayes (NB) is a supervised learning classifier based on Bayes' theorem that considers the "naive" assumption, finding the relation between every set of elements with equal impact to the target class. The NB classifier considers each feature as unique and does not cooperate with other features. Each class has independent and distinctive features that similarly distributes to the probability of a sample. Naive Bayes is straightforward to develop and is computationally quick, works well on large scale datasets and is not hypersensitive to noise [36].

V. Experiments and Results

This section presents the results of risky permission ranking, application evolution, the extracted explicit privacy leakages features based on level of protection and PlusApps' classification performance evaluation. To further evaluate the performance of the proposed approach, Weka version 3.8.4 software was used [40]. Weka offers resources to train and evaluate classification models for any given features set.

A. Risky Permissions Ranking

To distinguish between plus and original apps, Correlation Coefficient (CorEvel) and information gain (InfGain) methods have the ability to rank the permissions. This ranking uses to determine the plus and original apps and all the data are used to the permission ranking for the experiments. The top 30 risky permissions are illustrated in Table II. As described in the table, the Correlation Coefficient and information gain produced various orders of risky permissions. The results of ranking contain the same risky permissions, with five unique risky permissions in both methods.

Fig. 4 illustrates the existence rate of each top placed permission with CorEvel in original and plus apps. As is shown, the top risky permission differentiate plus apps from the original apps by the rate of occurrence. The top four risky

TABLE II. Top 30 Risky Permissions Ranked by CorEvel and InfGain.

Rank	Correlation Ranker			InfoGain
	Protection Level	Score	CorEvel	
1	Special	0.5854	KILL_BACKGROUND_PROCESSES	KILL_BACKGROUND_PROCESSES
2	Dangerous	0.3549	sticker.READ	sticker.READ
3	Special	0.3549	SYSTEM_ALERT_WINDOW	SYSTEM_ALERT_WINDOW
4	Special	0.3162	WRITE_USE_APP_FEATURE_SURVEY	WRITE_USE_APP_FEATURE_SURVEY
5	Dangerous	0.1696	READ_SETTINGS	READ_SETTINGS
6	Dangerous	0.1696	UPDATE_SHORTCUT	UPDATE_SHORTCUT
7	Signature	0.1584	BROADCAST_BADGE	BROADCAST_BADGE
8	Special	0.1576	WRITE_SETTINGS	WRITE_SETTINGS
9	Dangerous	0.1486	READ_EXTERNAL_STORAGE	BLUETOOTH
10	Dangerous	0.1412	USE_FULL_SCREEN_INTENT	BROADCAST
11	Normal	0.1334	BROADCAST_STICKY	BILLING
12	Dangerous	0.1334	BROADCAST	BROADCAST_STICKY
13	Normal	0.1334	BLUETOOTH	MAPS_RECEIVE
14	Dangerous	0.1334	BILLING	WRITE
15	Signature	0.1301	REQUEST_INSTALL_PACKAGES	CAMERA
16	Signature	0.1294	REGISTRATION	INSTALL_SHORTCUT
17	Dangerous	0.1191	WRITE	READ
18	Normal	0.1191	INSTALL_SHORTCUT	MODIFY_AUDIO_SETTINGS
19	Dangerous	0.1191	READ	READ_GSERVICES
20	Special	0.1191	MODIFY_AUDIO_SETTINGS	CHANGE_WIFI_STATE
21	Signature	0.1191	MAPS_RECEIVE	UNINSTALL_SHORTCUT
22	Dangerous	0.1191	READ_GSERVICES	READ_SYNC_STATS
23	Dangerous	0.1191	CAMERA	READ_SYNC_SETTINGS
24	Special	0.1191	CHANGE_WIFI_STATE	WAKE_LOCK
25	Normal	0.1175	FOREGROUND_SERVICE	WRITE_SYNC_SETTINGS
26	Dangerous	0.103	READ_PHONE_STATE	READ_PROFILE
27	Signature	0.103	UNINSTALL_SHORTCUT	RECEIVE
28	Normal	0.103	READ_SYNC_SETTINGS	VIBRATE
29	Dangerous	0.103	READ_PROFILE	WRITE_CONTACTS
30	Normal	0.103	READ_SYNC_STATS	RECEIVE_BOOT_COMPLETED

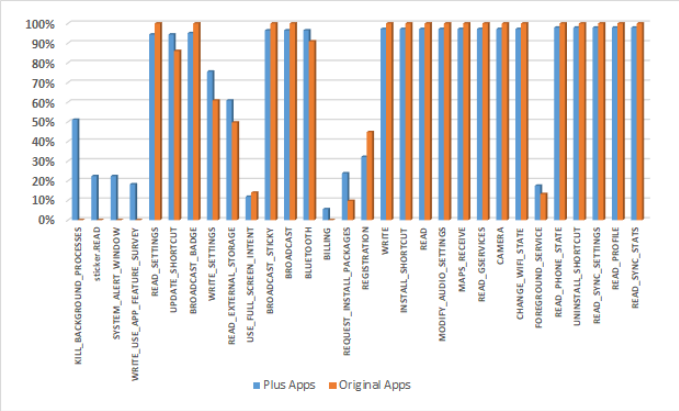


Fig. 4. Occurrence Percentage of the Top 30 Ranked Risky Permissions in Plus Apps and Normal Apps.

permissions are uniquely utilized in the plus apps, ranked by two ranking methods. 40% of the top 30 ranked permissions are dangerous permissions. The occurrence rate for original apps is higher than for plus apps. The top four occurrence rates for plus apps is above 15%, with the following permissions: KILL_BACKGROUND_PROCESSES, sticker.READ, SYSTEM_ALERT_WINDOW, and WRITE_USE_APP_FEATURE_SURVEY, with no occurrence rates for original apps. Most of the ranked permissions are similar due to applying the experiments on the WhatsApp application versions only. The result shows that the use pattern of system modification and data collection permissions is vastly different between the plus apps and the original apps, and many plus apps try to collect data using read and system modification permissions. These permission are consistent with the paper published by Shrivastava et al. [41], which implies that the number of dangerous permissions used to identify the untrustworthy applications and 68% of mobile threats are SMS abuse and data stealer accounts. Billing is also a risky permission that is more likely to be demanded by

plus apps, and CALL_PHONE and INTERNET are sensitive permissions. These latter two permissions are not placed in the top risky permissions demanded by plus and original apps.

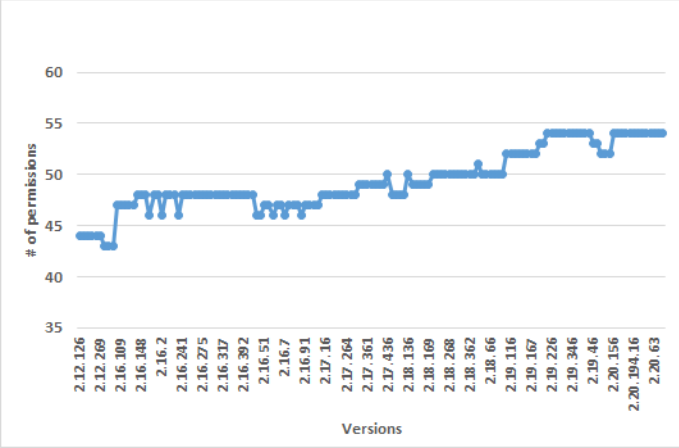


Fig. 5. Permission Evolution of WhatsApps Benign Apps.

Permissions are put into categories of protection levels as “normal”, “dangerous”, “signature” or “special”. The protection level affects whether runtime permission requests are required [42]. The 30 ranking results are almost consistent with the protection levels of Android APIs. The levels of protection are shown in Table II, based on Android 9 API level 28. The number of used permissions is increased in new versions of Android apps, and the occurrence of dangerous permissions in plus apps is significantly higher than the number of dangerous permissions in original apps. There are 6 normal permissions, 13 dangerous permissions, 5 signature permissions and 6 special permissions in ranked permissions. 17 of the top 30 ranked permissions are read, write and change or modification permissions that are categorized as dangerous, signature or special. Most of the plus apps induce users to gain permissions to the app that triggered these permissions to break the system privacy without the awareness of the users. For example, permission KILL_BACKGROUND_PROCESSES is recognized as “special” by Android. However, it is requested by 51% of plus apps in order to allow the plus apps to kill background processes of running applications to activate the suspicious behaviors. This study indicates that most of the ranked permissions are trying to access system resources.

With new development of Android APIs, the requested permissions of Android apps are increased with extra permissions. Fig. 5 presents the permission evolution of WhatsApp’s original apps that shows a significant increase. The permission evolution of WhatsApp’s plus apps are illustrated in Fig. 6, which are similar to the original app’s permission evolution, with the range of 40 to 57 permissions per version.

B. Evaluation Methodology

The evaluation methodology ensures that the PlusApps classifier modules perform well and are able to spot plus apps abnormal behavior intelligently. The development of the classification model consists of four steps: input data selection,

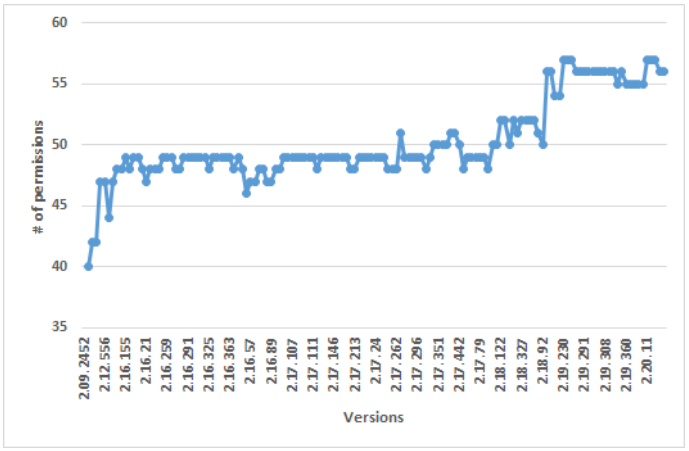


Fig. 6. Permission Evolution of WhatsApp’s Plus Apps.

data pre-processing and splitting (including selecting the features and addressing class), the setting of model parameters and model implementation. The first step in developing the classification models is input variable selection. The APK files vary for each type of dataset that contain plus and original apps. The second step is data pre-processing and splitting. Data pre-processing is an essential step to prepare the data, eliminate outliers and balance the features to the same range. Datasets are typically pre-processed before they can be used for training to speed up convergence. Data is normalized using a linear transformation, which is the method of re-scaling one or more attributes to the range of 0 to 1. Data pre-processing is the process of data conversion that necessitates each data instance is reproduced as a vector of real numbers. Consequently, data has to be converted to numeric data if they are definite attributes [39]. For app classification, they are typically classified to be normal or plus apps, being represented as 0 or 1, before they can be provided to the classifiers. After that the datasets are separated into two parts, which are a training and a testing set. There is no regulation to choose the data partition of training and testing datasets [43]. In many situations, the researchers applied various combinations of data division and adjusted corresponding to the problems.

TABLE III. SUMMARY OF THE 10-FOLD CROSS VALIDATION ACCURACY METRICS.

Accuracy	Kappa	Plus		Normal	
		TP	FN	TN	FP
95.8%	0.902	423	31	970	30

The third step is defining the model parameter and is extremely vital. The appropriate model parameters can enhance the J48, Naive Bayes, Random Tree and SVM classification accuracy performances. There are three types of parameters that should be considered for training the J48 algorithm, which have influence on the resulting decision tree, namely, minimum number of instances in a leaf, use of unpruned trees, confidence factor used in post-pruning and the subtree-raising operation in post-pruning. For the Naive Bayes classifier, all model parameters can be estimated with relative frequencies from the training set namely, class priors, useKernelEstimator

and feature probability distributions. Random Forest is a meta estimator that suits a number of decision tree algorithms on a variety of sub-instances of the data. The most important parameters of Random Forest use averaging to enhance the projecting precision and reduce over-fitting. The sub-instance size is managed with the max_samples parameter if bootstrap=True (default), otherwise the entire dataset is utilized to construct each tree. It then optimizes the random forest, which can be done through a random search using the RandomizedSearch. For the SVM classifier model, there are two important parameters that are considered in the RBF kernel function, namely, regularization C parameter and gamma parameter [44]. The tradeoff cost between minimizing the training error and the complexity of the model are verified by the C parameter, which identifies the non-linear drawing from the low-level space to some high-level dimensional space [44]. In this paper, a parameter search is performed to identify the finest values of parameter C, using trial and error approaches.

The last steps in developing the classification models are model implementation. For the J48 model, the algorithm uses attribute selection to decrease dataset size by eliminating irrelevant/redundant attributes. This algorithm discovers the minimum set of attributes and the resulting probability distribution of data classes, which should be near to the original distribution. In the Naive Bayes classifier, the NB trains until all the features are evaluated, and the one with the highest probability (score) the the winner. For Random Forest, an additional action is to improve the random forest using a random search. Optimization implies obtaining the best hyperparameters for a model on the dataset. The finest hyperparameters differ among datasets and perform model tuning. For SVM classification, the model is trained until the finest sets of parameters (C, Y) are taken. To train and test the generated models, a 10-fold cross-validation technique was performed. In this technique, the instances were divided into training sets (90%) and testing sets (10%), in which the testing set was not part of the training set [45]. Ten unique datasets were produced; in each a different 10% data partition was held out for testing and the rest of the data was used for training. The benefits of k-fold cross validation are that the influence of data dependency is reduced and the consistency of results can be increased. The model's performance was measured using values of true positive rate, false positive rate and AUC (the area under the ROC curve).

C. Performance Evaluation

To differentiate between plus and original application, classification techniques are usually employed in order to evaluate the proposed approaches . The confusion matrix is the best way of representing the classification result (Table V). Due to the two-class nature of the classifier, there are four measures as follows:

- True positive (TP): represents a plus application classified correctly as a plus version.
- False negative (FN): refers to a plus application classified incorrectly as an original version.
- True negative (TN) represents an original application classified correctly as an original version.
- False positive (FP) refers to an original application classified incorrectly as a plus version.

In addition, the performance of different classifier modules is measured using the standard metrics true positive rate, false positive rate and AUC (the area under the ROC curve). The standard metrics extract part of the information from the confusion matrix to produce a numeric value. The higher the true positive rate and the lower the false positive rate, the better the classification is.

1) True Positive Rate (TPR): TPR is the proportion between the plus applications classified correctly as plus version (Equation 1).

$$TPR(recall) = \frac{TP}{TP + FN} \tag{1}$$

2) False Positive Rate (FPR): FPR is the ratio between the number of misclassified original applications and the total number of original applications (Equation 2).

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

3) The ROC curves are used to visualize the relation between true and false positive rates of a certain classifier while tuning it, and on the other hand to compare the accuracy of several classifiers [43]. This metric has two limitations but is nevertheless very effective. The first shortcoming is that the ROC curve is based on the ratio of attack to normal data. It is used for the comparison of detection approaches that run on the same dataset, but the graph of the ROC curve is completely mislead when using it to compare different detection methods that run on different datasets. The second shortcoming is that it may be misleading and basically inadequate for understanding the strengths and weaknesses of a proposed method [43].

Data mining techniques were used to detect the behavior of plus apps based on permissions features. In this experiment, the analysis of different classifiers was used and applied on the dataset to verify whether the app was plus or normal. The results of the analysis were saved in ".csv" file that converted the arff extension in order to process it in Weka. The dataset consisted of 454 plus samples and 1,000 original samples. There were 202 feature vectors with a related label, and the last feature was classification either to plus or normal. The Random Forest classifier with a 10-fold cross validation was used for testing our model. Table III shows the results of evaluation metrics and reveals a high predictive performance.

The precision of recognizing 1,454 different plus and original apps is 95.8% and the Kappa measurement is 0.902, which indicates the performance of the Random Forest classifier using cross validation in this experiment. The False Positive Ratio (FPR) is 3% and 6% for plus and original apps respectively. Moreover, 30 original samples (6.8%) are incorrectly recognized as plus and 93.2% of original apps are correctly identified. 97% of plus apps out of 454 were detected and only 30 plus apps misclassified as legitimate apps. The number of plus apps correctly classified (TP) is 423 and the number of original apps precisely categorized (TN) is 970 apps. Table IV illustrates the results.

TABLE IV. AREA UNDER THE RECEIVER (AUC) AND FALSE POSITIVE RATIO FOR DIFFERENT CLASSIFIERS.

Classifiers	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	PRC Area
SVM	96.60%	5.00%	96.60%	96.60%	92.10%	95.80%	94.90%
NB	93.90%	11.40%	94.10%	93.80%	85.80%	95.50%	96.20%
J48	91.10%	19.30%	92.00%	90.70%	79.50%	85.70%	87.10%
RF	95.80%	5.60%	95.80%	95.80%	90.20%	98.10%	97.90%

TABLE V. THE CONFUSION MATRIX.

	Original	Plus
Original	TN	FP
Plus	FN	TP

1) Discussion and Comparison: Other machine learning algorithms were used with 10-fold cross validation in the experiment. These algorithms were Decision Trees (J48), Naive Bayes, Random Tree and Support Vector Machine (SVM). Table IV shows that all classifiers provided high detection accuracy. An SVM with an RBF kernel classifier accomplished the highest accuracy for the proposed approaches, followed by Random Forest with a high achievement of 95.8% detection rate. The Naive Bayes classifier has a close result with 93.9%, and the Decision Tree (J48) classifier demonstrated the lowest outcome with a 91.10% detection rate. Hence, due to the high similarity of the samples, the Decision Tree (J48) performed the lowest in the proposed model.

Most malware detection approaches consider dangerous permissions are malicious such as SMS permissions. In the paper presented by Wang et al. [21], the SMS-related permissions are always ranked very top risky permissions by the three defined ranking methods. The proposed model has not rated SMS related permission as unsafe permissions in plus application. Also, bill-related and system-related permissions are ranked risky permission in the proposed approach, but the paper published by Wang et al. [21] which not deem these permissions as malicious permissions due to the difference in the behavior of the Plus apps from the malicious application. Other ranked risky permissions are almost similar in both works.

VI. CONCLUSIONS AND FUTURE WORK

The Android platform is an open ecosystem that allow its developers to tailor some of its default features and settings. Many of these settings can be easily customized and use read, write or modified settings of the devices. A plus application is the modified version of an original application with extra features that could result privacy data leakages or misbehavior of these applications. It is crucial that mobile scientific analyzers pay close attention to the types of permissions that these Android applications can request. Therefore, it is important that the many plus applications which are identical to the original application with extra features are investigated in order to study the application's behavior and observe its privacy handling. This paper presents a risk analysis of Android plus applications that investigates the plus apps' behavior and data leakages using classification algorithms. It reveals classification modes to distinguish plus apps from original apps through identifying the effectiveness of Android permissions. The research has shown that it is possible to retrieve some artifacts from plus applications when users installed and used these types of plus applications. These artifacts included kill processes, sticker read, system alert and billing features. The research has also shown that some unique permissions can be used by plus applications that are considered risky features, which results in permissions abuse and data stealer accounts. The analysis reveals that the SVM classifier presented the highest accuracy and additional investigation demonstrates that the classifier with the ranking algorithm use Correlation Coefficient (CorEvel) and Information Gain (InfGain) methods.

REFERENCES

[1] B. Y. Zhang, X. A. Yan, and D. Q. Tang, "Survey on malicious code intelligent detection techniques," *Journal of Physics: Conference Series*, vol. 1087, p. 062, sep 2018.

[2] E. Casey, C. Daywalt, and A. Johnston, "Intrusion investigation," in *Handbook of Digital Forensics and Investigation*. San Diego: Academic Press, 2010, pp. 135–206.

[3] D. Price, "The 20 most popular android apps in the google play store." New York ,USA, 03 2020.

[4] L. Cheng, F. Liu, and D. D. Yao, "Enterprise data breach: causes, challenges, prevention, and future directions," *WIREs Data Mining and Knowledge Discovery*, vol. 7, no. 5, pp. 1–11, 2017.

[5] M. Alsaleh, N. Alomar, and A. Alarifi, "Smartphone users: Understanding how security mechanisms are perceived and new persuasive methods," *PLOS ONE*, vol. 12, 03 2017.

[6] A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner, "Android permissions: User attention, comprehension, and behavior," in *Proceedings of the Eighth Symposium on Usable Privacy and Security*, ser. SOUPS '12. New York, NY, USA: Association for Computing Machinery, 2012.

[7] Y. Kim, T. Oh, and J. Kim, "Analyzing user awareness of privacy data leak in mobile applications," *Mobile Information Systems*, vol. 2015, pp. 1–12, 12 2015.

[8] S. Rai, R. Dhanesha, S. Nahata, and B. Menezes, *Malicious Application Detection on Android Smartphones with Enhanced Static-Dynamic Analysis*, 01 2017, pp. 194–208.

[9] P. Singh, P. Tiwari, and S. Singh, "Analysis of malicious behavior of android apps," *Procedia Computer Science*, vol. 79, pp. 215–220, 2016, proceedings of International Conference on Communication, Computing and Virtualization (ICCCV) 2016.

[10] Z. Fang, W. Han and Y. Li, "Permission based android security: Issues and countermeasures," *Computers & Security*, vol. 43, pp. 205–218, 2014.

[11] A. Oglaza, R. Laborde, P. Zarate, A. Benzekri and F. Barrere, "A new approach for managing android permissions: learning users' preferences," *EURASIP Journal on Information Security*, vol. 2017, 07 2017.

[12] Technopedia, "Android platform." New York ,USA, 08 2011.

[13] A. Shabtai, Y. Fledel, U. Kanonov, Y. Elovici, S. Dolev, and C. Glezer, "Google android: A comprehensive security assessment," *IEEE Security Privacy*, vol. 8, no. 2, pp. 35–44, 2010.

[14] V. Code, "Android security: Guide to android os." Burlington, MA ,USA, 01 2020.

[15] S. Dua and X. Du, *Data Mining and Machine Learning in Cybersecurity*, 1st ed. USA: Auerbach Publications, 2011.

[16] B. Ahmad, J. Wan, and Z. A. Ali, "Role of machine learning and data mining in internet security: Standing state with future directions," *Journal Comp. Netw. and Communic.*, vol. 18, p. 10, 2018.

[17] M. Egele, T. Scholte, E. Kirda, and C. Kruegel, "A survey on automated dynamic malware-analysis techniques and tools," vol. 44, no. 2, 2008.

[18] Y. Zhou and X. Jiang, "Dissecting android malware: Characterization and evolution," in *2012 IEEE Symposium on Security and Privacy*, 2012, pp. 95–109.

[19] S. Arzt, S. Rasthofer, C. Fritz, E. Bodden, A. Bartel, J. Klein, Y. Le Traon, D. Octeau, and P. McDaniel, "Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps," *SIGPLAN Not.*, vol. 49, no. 6, pp. 259—269, Jun. 2014.

[20] M. Gordon, K. deokhwan, J. Perkins, L. Gilham, N. Nguyen, and M. Rinard, "Information-flow analysis of android applications in droidsafe," 01 2015.

[21] W. Wang, X. Wang, D. Feng, J. Liu, Z. Han, and X. Zhang, "Exploring permission-induced risk in android applications for malicious application detection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 11, pp. 1869–1882, 2014.

[22] W. Enck, P. Gilbert, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth, "Taintdroid: An information-flow tracking system for realtime privacy monitoring on smartphones," in *Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation*, ser. OSDI'10. USA: USENIX Association, 2010, pp. 393—407.

[23] L. Lu, V. Yegneswaran, P. Porras, and W. Lee, "Blade: An attack-agnostic approach for preventing drive-by malware infections," in *Proceedings of the 17th ACM Conference on Computer and Communications Security*. New York, NY, USA: Association for Computing Machinery, 2010, pp. 440—450.

[24] K. Z. Chen, N. Johnson, S. Dai, K. Macnamara, T. Magrino, E. Wu, M. Rinard, and D. Song, "Contextual policy enforcement in android applications with permission event graphs," 2013.

[25] Y. Zhang, M. Yang, B. Xu, Z. Yang, G. Gu, P. Ning, X. S. Wang, and B. Zang, "Vetting undesirable behaviors in android apps with permission use analysis," in *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*. New York, NY, USA: Association for Computing Machinery, 2013, pp. 611—622.

[26] Z. Yang, M. Yang, Y. Zhang, G. Gu, P. Ning, and X. S. Wang, "Appintent: Analyzing sensitive data transmission in android for privacy leakage detection," in *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, ser. CCS '13. New York, NY, USA: Association for Computing Machinery, 2013, pp. 1043—1054.

[27] M. C. Grace, Y. Zhou, Z. Wang, and X. Jiang, "Systematic detection of capability leaks in stock android smartphones." in *NDSS*, 2012.

[28] O. Tripp and J. Rubin, "A bayesian approach to privacy enforcement in smartphones," in *Proceedings of the 23rd USENIX Conference on Security Symposium*, ser. SEC'14. USA: USENIX Association, 2014, pp. 175—190.

[29] A. Continella, Y. Fratantonio, M. Lindorfer, A. Puccetti, A. Zand, C. Kruegel, and G. Vigna, "Obfuscation-resilient privacy leak detection for mobile apps through differential analysis," 01 2017.

[30] G. Barbon, A. Cortesi, P. Ferrara, M. Pistoia, and O. Tripp, "Privacy analysis of android apps: Implicit flows and quantitative analysis," in *Computer Information Systems and Industrial Management*. Cham: Springer International Publishing, 2015, pp. 3–23.

[31] G. Barbon, A. Cortesi, P. Ferrara, and E. Steffinlongo, "Dapa: Degradation-aware privacy analysis of android apps," in *STM*, 2016.

[32] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Engineering Bulletin*, vol. 23, p. 20, 2000.

[33] *Androguard*, available at: http://github.com/androguard/androguard./ accessed on Feb 14, 2020.

[34] A. M. Aswini and P. Vinod, "Droid permission miner: Mining prominent permissions for android malware analysis," in *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, vol. 5, no. 1, 2014, pp. 81–86.

[35] I. Guyon, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[36] S. Misra and H. Li, *Noninvasive fracture characterization based on the classification of sonic wave travel times*, 2020, pp. 243–287.

[37] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81—106, 1986.

[38] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5—32, 2001.

[39] C. Burges, "A tutorial on support vector machines for pattern recognition." *Data Mining and Knowledge Discovery*, vol. 2, no. 1, p. 121-167, 1998.

[40] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. Witten, and L. Trigg, *Weka - A Machine Learning Workbench for Data Mining*, 07 2010, pp. 1269–1277.

[41] G. Shrivastava and P. Kumar, "Intent and permission modeling for privacy leakage detection in android," *Energy Systems*, 10 2019.

[42] G. Developer, "Permissions overview," New York ,USA, 07 2020.

[43] S. Alsoghyer and I. Almomani, "Ransomware detection system for android applications," *Electronics*, vol. 8, p. 868, 08 2019.

[44] H.-L. Chen, B. Yang, J. Liu, and D.-Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 38, no. 7, pp. 9014–9022, 2011.

[45] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143.

Design of a Mobile Application for the Automation of the Census Process in Peru

Luis Alberto Romero Tuanama[1], Juber Alfonso Quiroz Gutarra[2], Laberiano Andrade-Arenas[3]
Facultad de Ciencias e Ingeniería
Universidad de Ciencias y Humanidades

***Abstract*—This study shows that the traditional census process in Peru has many shortcomings, including the loss of data and the long duration of the process. To solve this problem, a mobile application was designed to automate the census process in Peru. For the development, we will rely on the agile scrum methodology, Balsamiq Mockup and Adobe XD tools, for help us to make prototypes of this application. In the last census, many families were not registered due to lack of time or other factors, so we designed this prototype of a mobile application, which will help the census taker to make the data recording process faster. The result obtained is the proposal of a productive approach, optimizing the census process, through a mobile application, where each census taker will register the data of the families on the census in a faster way and this information will be taken directly to the database of the organization conducting the census, thus avoiding loss of data, saving time and money.**

***Keywords*—Automation; Balsamiq Mockup; census; scrum**

I. Introduction

In many countries around the world and the United States, the census process has had a very drastic change, they put aside all the paperwork and automated this process, now the inhabitants of that country can be registered using a web platform or a phone call, the use of a web platform allows the user to insert data more accurately about himself and his family. Census refers to the process of retrieving, compiling, and publishing demographic data, economic and social data of all people in a country or territory at one point. However, conducting a census isn't free of problems. It has become a delicate and challenging subject, especially in Nigeria [1].

On the other hand, in Peru the census process is "very outdated", result of that, data is inaccurate, a clear example would be the bonus that the government distributed because of the pandemic, which was distributed to the most "vulnerable" population in the country. Unfortunately, due to the inaccuracy of data, this bonus was not distributed properly, since there are families who met all the requirements of the bonus, but were not beneficiaries of it.

While some countries make extensive use of census data such as the U.S. That Use U.S. Census Explorer that provides information for intuitive analysis of census data, others under-exploit, null or limited the integration of census data with data from other sources for development planning and analysis [2]. Due to the loss of data within this process, the inhabitants are exposed to losing certain benefits provided by the state, a clear example would be the budget allocated for the execution of various public works, this loss of data generates that these budgets are lowered, or the citizens as individuals do not enter into a specific pattern. Researchers and health services often use the information on ethnic origin to plan services, ensure equality of access, and epidemiological studies [3].

One of the most known software is "Geographical Information System" which is used in Spain to analyze the impact of tourism, this software can be taken as a model to design an application for the census in Peru, because, this not only records the data of the geographical location where tourists come to the country, also sends you reports by specific areas as the approximate number of people who are by temperature measurements [4]. Also, the application that will be designed will not only have the registration of the families, it will also have a report section, where you will see the number of families that were registered and will be able to give a statistical report of the population growth compared to the previous census.

Another known software is CartoDruid that receives information about crops and plots from different geographical areas, the River Surveillance Agency (RSA) collects agronomic data on the type of crop inspected, its phenological status and the irrigation system used in each of them. Researchers use this tool to solve problems with the use of geographic information, both to be able to consult and edit them [5]. From this software we can get the most important thing that the handling of large-scale information divides by regions geographically, and can be of great help in Peru since there are places "hidden" where the census does not reach and that is an important data loss.

The objective is to design an application to automate the census process in Peru and to avoid data loss.

The points to be touched will be, in section II literature review, where it shows research similar to this work, in section III methodology, where you will see the tools that will be used for the development of the application and the scrum framework, then in section IV will go on to make the case study as proposed in the methodology , then section V presents the results and discussions, finally in section VI to make way for the conclusions.

II. Literature Review

In the revision of the literature there is a thesis and article regarding mobile applications for census.

This thesis by Julio Gonzales, describe the process of developing a mobile application and a web platform for the organization Techo Guatemala, explains the creation of a tool that facilitates the process of taking and managing data collected during the census, because it will be carried out in remote areas of Guatemala City, the mobile application conducts the census without internet connection and, subsequently, synchronizes the raised data with the web platform, this application for mobile devices is effective for lifting data and there is no need to invest in physical and expensive resources such as paper ballots [6].

The following research by Tarik Bourezgue deals with the use of tablets for the 2018 Algerian census: census data management and quality assessment, we talk about how tablets with mobile chips are used to collect information. rather than the traditional paper questionnaire. the author's idea is to deliver a tablet containing the census application to each packer through which they can collect census data and update the collected data to the census database over mobile networks, this tool serves to ensure that the ONS meticulously lists and presents census data with minimal difficulties and minimal risks to have accurate data to have accurate data [7].

III. Methodology

For the development of the methodological section below we describe in detail the steps to be followed for the development of the mobile application. For this development, we will use the agile Scrum methodology, as this provides us a development environment with more value towards the product. Also, the development tools are detailed, which are the programming language Kotlin, the database manager SQLite and the development environment IntelliJ IDEA, on the other hand, there is also talk about the design tools that are used, which are, Balsamiq Mockups for the initial design and Adobe XD for the final design.

A. Scrum

Scrum is a very useful framework that helps teams to work in complicated environments, where the requirements are very variable and the results must be given in a short period [8].

Scrum is composed of the team roles that are: the product owner who in most cases is the spokesperson for the company or client and is responsible for managing the backlog; the scrum master who is a leader with advanced knowledge of the scrum methodology, however, he does not have hierarchical authority over the group, it can be said that he is more like a guide who helps implement the methodology correctly, facilitating meetings and helping the product owner to prioritize the backlog; and the scrum team that can be composed of 3 to 9 members who must be well organized and committed to the project to perform the tasks agreed in each sprint [9].

Then we have the Scrum events that are predefined to collaborate as a group to improve their knowledge and reduce meetings, such events are:

1) The Sprint: It is given over 1 to 4 weeks during which time the team must perform the planned tasks [9].

2) Sprint Planning: Where it is determined which tasks will be performed and handed over at the next sprint [9].

3) Daily: It is a daily meeting of a maximum of 15 minutes where each member of the group quickly informs what they did yesterday, what they will do today, and if they had any difficulty in advancing the project [9].

4) Sprint Review: This is where the work is done according to the sprint is delivered to the product owner for review and to verify if it meets the requirements to be accepted or if some changes need to be made [9].

5) The Retrospective: It is the final team meeting where lessons are learned, for continuous improvement and use in the next development [9].

Finally, we have the scrum artifacts that provide primary information to have a better understanding of the project being developed, such artifacts are the product backlog, which is the list of tasks where the project requirements are described that are prioritized according to the value it gives to the business; the sprint backlog, which is a specific list of requirements chosen from the product backlog to perform in a sprint; and the increment that is the sum of the tasks that were performed since the last version of the product delivered [9]. Fig. 1 shows the process followed by the Scrum methodology

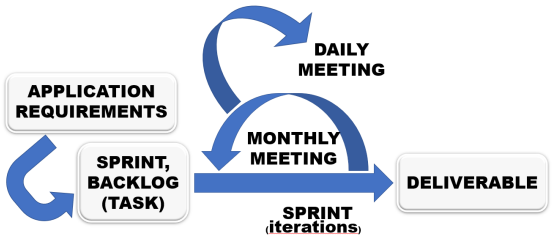


Fig. 1. Scrum Process

B. Development Tools

1) Kotlin: It is an object-oriented programming language, which develops both web and mobile applications, whether iOS or Android. Also, it is developed in the Java Virtual Machine (JVM) and its main advantage is that it reduces the

repetition of code, something that helps the programmer a lot because it saves resources and time [10].

2) SQLite: Database performance is a very important factor in the development of any mobile application. Some of these mobile applications need fast feedback, others process a large amount and variety of data, and without the support of modern database management systems, this is almost impossible [11]. SQLite is a database engine that has built-in SQL. Unlike most other SQL databases, SQLite does not have a separate server process. SQLite reads and writes directly to normal disk files [12]. It is the most suitable option in terms of development for mobile applications since it provides a large storage size, stability, and performance for development.

3) IntelliJ IDEA: IntelliJ IDEA is an integrated development environment, written in the Java programming language for computer software development and later also used for mobile applications developed specifically for Android. It is developed by JetBrains and is available with a community edition where the Intellij IDEA community provides support [13]. This IDE has a friendly and powerful environment which allows developers to work dynamically and accurately.

C. Design Tools

To design the mobile application, we used two design software, first, we used Balsamiq Mockups to get an initial view of the wireframe of the application, this helps us to know which elements will be used and where they will be positioned. For the final presentation of the design, we use Adobe XD, this software allows us to have a more detailed appearance thanks to its vectorial elements and gives us options of interactivity for the design of the application.

1) Balsamiq Mokups: It is a wireframe design tool for mobile applications and web interfaces, with its sketch appearance helps to create the initial stages of basic prototype development to have an initial design of the mobile application.

2) Adobe XD: Adobe XD is a tool for the design of mobile interfaces developed by Adobe Inc, it allows to create of a preliminary image of the final version of the interface using thousands of vectorial elements, this benefit both the developers and the user since it allows the interface to be very interactive.

IV. Case Study

A. Beginning Stage

1) Identification of requirements: We will identify the indispensable requirements for the development of the mobile application as shown in Table I.

TABLE I. Requirements

REQUIREMENTS
The application must allow the connection to the information platform using the identifier of each census taker; this is the ID number and password.
The application must have an initial menu, where you can add a new census, visualize the registered houses, census map and user profile.
The application within the new census option should allow you to fill out information on a form based on the census card format.
The application within the option to visualize registered houses must have the possibility to visualize in detail and modify the previously registered houses
The application within the census map option will allow you to visualize the places that have already been counted in real time and the location of the user.
The application within the user profile option will owe the data of the census taker and will be able to make change of password and mail.

2) User Stories: We defined the user stories, these user stories were written considering roles and requirements as shown in Table II.

TABLE II. User stories

USER STORIES
As a census taker I want to access the application through a login to manipulate its menu.
As a census taker I want to access the new census option in the main menu to register a new census.
As a census taker I want to fill out a form based on the census card format in order to register the housing census.
As a census taker I want to access the option to visualize houses in the main menu to see all the houses registered by the logged-in user.
As a census taker I want to access the option to visualize houses in the main menu to modify a house that is already registered.
As a census taker I want to access the census map option in the main menu to see all the homes listed around the user's location.
As a census taker I want to visualize my location to know exactly where the user is.
As a census taker I want to access the option my profile in the main menu to modify data such as the user's email or password.

3) Management of the backlog: The backlog was managed to prioritize the most or least important user as shown in Table III, also used the estimation method planning poker which will be discussed later.

B. Planning Stage

1) Poker planning estimate: Planning Poker is one of the most effective and well-known techniques in the world of agile development methodologies to estimate [14]. The agile dynamic was carried out with the scrum team and the final results were noted in Table IV.

2) Sprint Planning: In this subsection the planning of the sprints will be done, which were defined in 4 sprints, within them are detailed the stories to be made, the person in charge

TABLE III. MANAGEMENT OF THE BACKLOG

N°	Item	Priority	Estimate
1	As a census taker I want to access the application through a login to manipulate its menu.	HIGH	2
2	As a census taker I want to access the new census option in the main menu to register a new census.	HIGH	1
3	As a census taker I want to fill out a form based on the census card format in order to register the housing census.	HIGH	13
4	As a census taker I want to access the option to visualize houses in the main menu to modify a house that is already registered.	MIDDLE	5
5	As a census taker I want to access the option to visualize houses in the main menu to see all the houses registered by the logged-in user.	MIDDLE	2
6	As a census taker I want to access the census map option in the main menu to see all the homes listed around the user's location.	MIDDLE	13
7	As a census taker I want to access the option my profile in the main menu to modify data such as the user's email or password.	LOW	2
8	As a census taker I want to visualize my location to know exactly where the user is.	LOW	1

TABLE IV. POKER PLANNING ESTIMATE

N° Item	Poker planning estimate	Time
1	2	2 days
2	1	1 day
3	13	13 days
4	5	5 days
5	2	2 days
6	13	13 days
7	2	2 days
8	1	1 day
Total	39	39 days

of making the story, the estimation time, the acceptance criteria and the user for whom the story is intended.

a) Sprint 1: For the development of sprint 1, two user stories with high priority were taken, the people responsible for each story were also designated, the acceptance criteria were defined and finally the realization of this sprint gave an estimated time of 3 calendar days. All these aspects are discussed in more detail in Table V.

b) Sprint 2: For the development of sprint 2, 1 user story with high priority was taken, the person responsible for each story was also designated, the acceptance criteria were defined and finally the realization of this sprint gave an estimated time of 13 calendar days. All these aspects are discussed in more detail in Table VI.

TABLE V. PLANNING SPRINT 1

SPRINT 1	
USER HISTORY:	As a census taker I want to access the application through a login to manipulate its menu.
ORDERED:	Romero
N° ITEM:	1
ESTIMATED TIME:	2 days
USER:	Census taker
ACCEPTANCE CRITERIA:	* Text boxes to insert user and password. * Button that allows you to log in.
USER HISTORY:	As a census taker I want to access the new census option in the main menu to register a new census.
ORDERED:	Quiroz
N° ITEM:	2
ESTIMATED TIME:	1 day
USER:	Census taker
ACCEPTANCE CRITERIA:	* Access button to the new census form

TABLE VI. PLANNING SPRINT 2

SPRINT 2	
USER HISTORY:	As a census taker I want to fill out a form based on the census card format in order to register the housing census.
ORDERED:	Romero
N° ITEM:	3
ESTIMATED TIME:	13 days
USER:	Census taker
ACCEPTANCE CRITERIA:	* Text boxes based on the format of the census card. * CBOs containing information based on the census card

c) Sprint 3: For the development of sprint 3, two user stories with medium priority were taken, the people responsible for each story were also designated, the acceptance criteria were defined and finally the realization of this sprint gave an estimated time of 7 calendar days. All these aspects are discussed in more detail in Table VII.

d) Sprint 4: For the development of sprint 4 a user story with medium priority was taken, the people responsible for each story were also designated, the acceptance criteria were defined and finally the realization of this sprint gave an estimated time of 13 calendar days. All these aspects are discussed in more detail in Table VIII.

e) Sprint 5: For the development of sprint 5, two user stories with low priority were taken, the people responsible for each story were also designated, the acceptance criteria were defined and finally the realization of this sprint gave an estimated time of 3 calendar days. All these aspects are discussed in more detail in Table IX.

TABLE VII. PLANNING SPRINT 3

SPRINT 3	
USER HISTORY:	As a census taker I want to access the option to visualize houses in the main menu to modify a house that is already registered.
ORDERED:	Romero
N° ITEM:	4
ESTIMATED TIME:	5 days
USER:	Census taker
ACCEPTANCE CRITERIA:	* List of registered homes. * Button to modify the census record.
USER HISTORY:	As a census taker I want to access the option to visualize houses in the main menu to see all the houses registered by the logged-in user.
ORDERED:	Quiroz
N° ITEM:	5
ESTIMATED TIME:	2 days
USER:	Census taker
ACCEPTANCE CRITERIA:	* List of registered homes.

TABLE VIII. PLANNING SPRINT 4

SPRINT 4	
USER HISTORY:	As a census taker I want to access the census map option in the main menu to see all the homes listed around the user's location.
ORDERED:	Romero
N° ITEM:	6
ESTIMATED TIME:	13 days
USER:	Census taker
ACCEPTANCE CRITERIA:	* Box showing the map. * Button to locate.

TABLE IX. PLANNING SPRINT 5

SPRINT 5	
USER HISTORY:	As a census taker I want to access the option my profile in the main menu to modify data such as the user's email or password.
ORDERED:	Quiroz
N° ITEM:	7
ESTIMATED TIME:	2 days
USER:	Census taker
ACCEPTANCE CRITERIA:	* Text boxes to insert email and password. * Button that allows you to modify.
USER HISTORY:	As a census taker I want to visualize my location to know exactly where the user is.
ORDERED:	Quiroz
N° ITEM:	8
ESTIMATED TIME:	1 day
USER:	Census taker
ACCEPTANCE CRITERIA:	* Box showing the map with its exact location.

Fig. 2. Prototype User Login - User Menu

C. Development Stage

In the next sub-section the development stage is carried out, which consists of the realization of the initial prototyping, then the revision of the sprints, and finally the realization of the prototype for the final user.

1) Initial Prototype: For the development of these prototypes were used previously planned sprints, these were analyzed and made in order of priority which facilitates their development, and allows the scrum team to have the objectives met in the estimated time.

a) Initial Prototype-Sprint 1: The development of this initial prototype is based on all the aspects that were detailed in the planning of sprint 1, see Fig. 2.

b) Initial Prototype-Sprint 2: The development of this initial prototype is based on all the aspects that were detailed in the planning of sprint 2, see Fig. 3.

c) Initial Prototype-Sprint 3: The development of this initial prototype is based on all the aspects that were detailed in the planning of sprint 3, see Fig. 4.

d) Initial Prototype-Sprint 4: The development of this initial prototype is based on all the aspects that were detailed in the planning of sprint 4, see Fig. 5.

e) Initial Prototype-Sprint 5: The development of this initial prototype is based on all the aspects that were detailed in the planning of sprint 5, see Fig. 6.

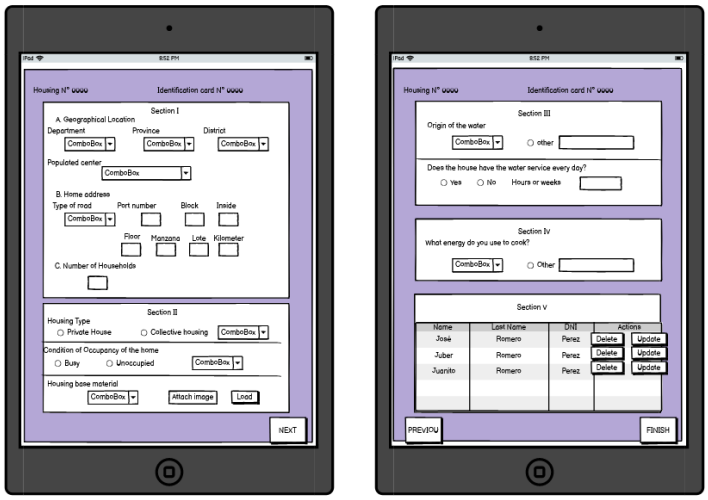


Fig. 3. Prototype New Census

Fig. 5. Prototype Map of Censored Homes

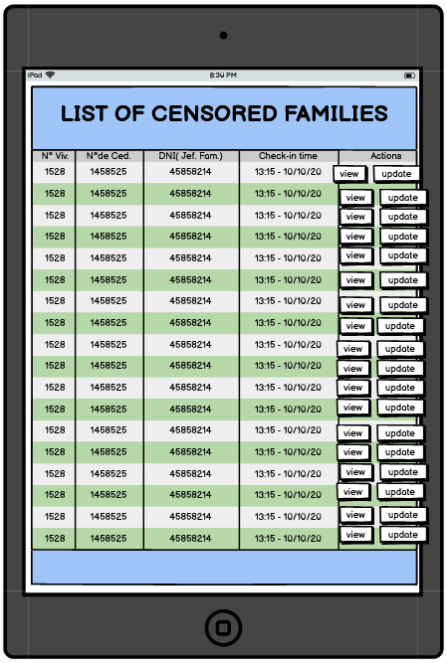


Fig. 4. Prototype List of Censored Homes

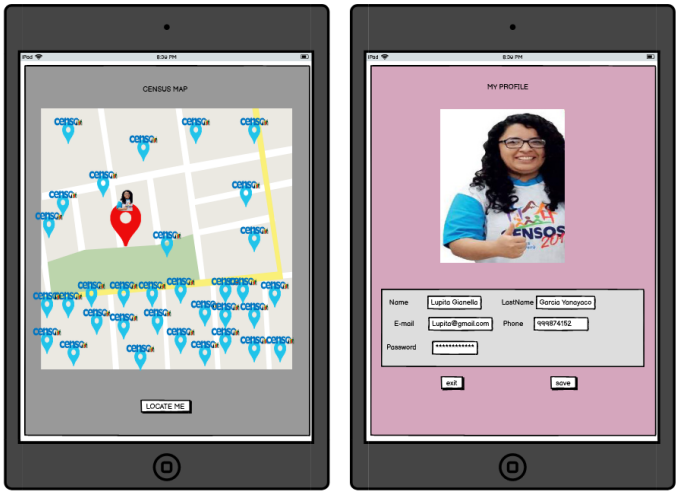


Fig. 6. Prototype Location and User Profile

2) sprint review: The Sprint Review is performed at the end of each sprint, this has a duration in hours equivalent to the time in days or weeks of the sprint. The product owner was responsible for reviewing what was done and identifies what was not done from the backlog; the Scrum team reports on the problems found and how they were solved, shows the product and its operation, if necessary or if the product owner asks for changes can be made. This meeting is of great importance to continue with the following Sprints [15]:

3) final prototype: For the development of these final phase prototypes, it was necessary to make a thorough review of the sprints where they were accepted or denied acceptance criteria, this helps in having a better version of the initial prototypes.

a) Login: This is a more detailed and stylized model for presentation to the end-user, this design comes from the Initial-Sprint 1 prototype. See Fig. 7

b) Menu: This is a more detailed and stylised model for presentation to the end user, this design comes from the Initial-Sprint 1 prototype. See Fig. 8.

c) Census Registration: This is a more detailed and stylised model for presentation to the end user, this design comes from the Initial-Sprint 2 prototype. See Fig. 9.

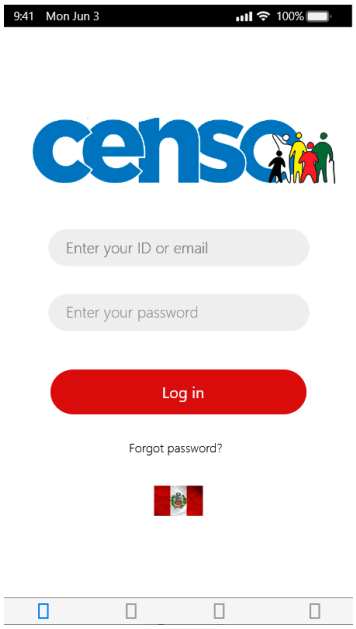


Fig. 7. Adobe xD Login

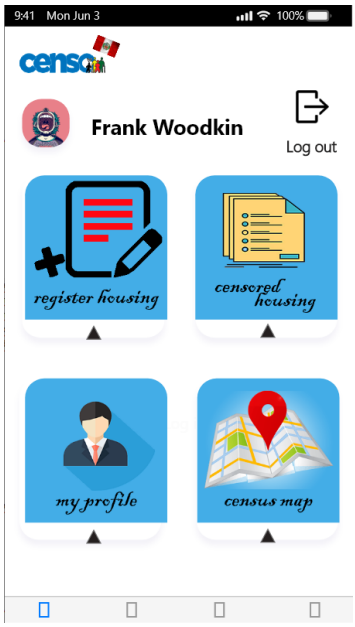


Fig. 8. Adobe xD Menu

d) List of Censored Homes: This is a more detailed and stylised model for presentation to the end user, this design comes from the Initial-Sprint 3 prototype. See Fig. 10.

e) User Profile: This is a more detailed and stylised model for presentation to the end user, this design comes from the Initial-Sprint 5 prototype. See Fig. 11.

f) User Location: This is a more detailed and stylised model for presentation to the end user, this design comes from the Initial-Sprint 4 prototype. See Fig. 12.

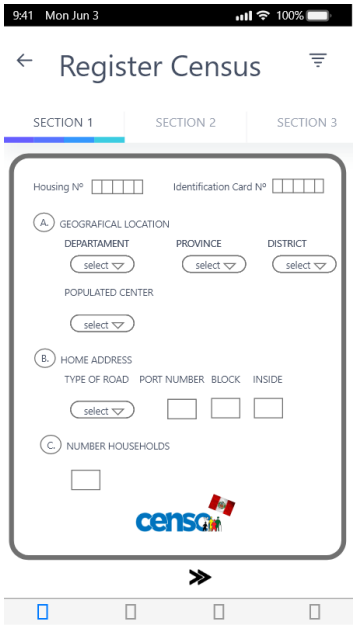


Fig. 9. Adobe xD Census Registration

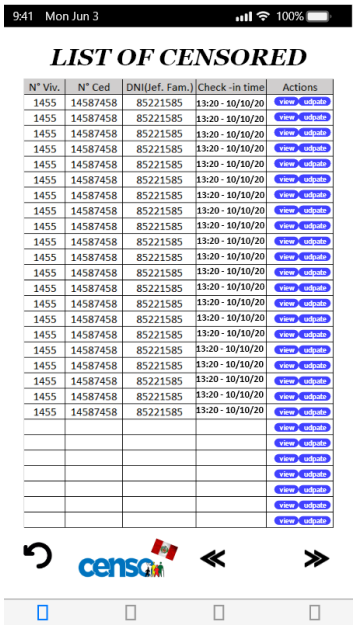


Fig. 10. Adobe xD List of Censored Homes

V. Results and Discussions

A. About the Case Study

For the case study, the design of the mobile application for the automation of the census process in Peru was carried out to make sure that the end-users have a non-complex interaction with the functionalities provided by the application. As a first tool for the design of prototypes, Balsamiq Mockups was used, since by its easy use it is possible to develop different types of prototypes for both web and mobile, comparing it with other research works we can see that most of them use this tool

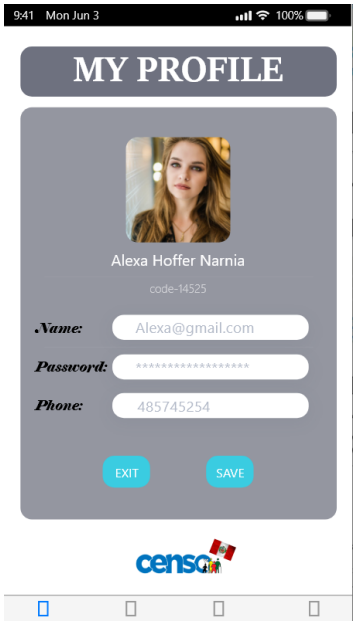


Fig. 11. Adobe xD User Profile

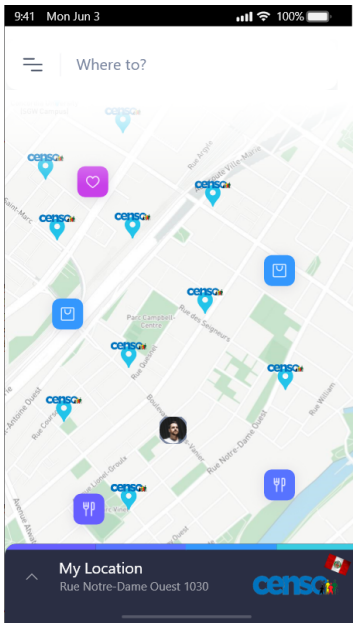


Fig. 12. Adobe xD User Location

[16], however, we believe that Balsamiq mockups do not show an interface that comes close to what you want in a more professional design, for this reason, it was decided to use the Adobe xD a tool to make the design of the final prototype because it showed an interface that allows a better idea of the creation of the application, can also be used as a base for implementing it in the Intellij IDE and only make the programming of the mobile application.

The Census is the largest data collection and processing process that exists in a country and serves as a basis for decision-making in the public and private sector, as well as a reference framework for egress and income calculations, living conditions or any other type of research that can be carried out, on the other hand the data obtained may be inaccurate, incomplete or have few groups and/or key areas [17] [18].Because this process is very ”outdated”, this leads to data being collected being lost or confused somewhere in the process.

1) As-is: Fig. 13 shows the current census process.

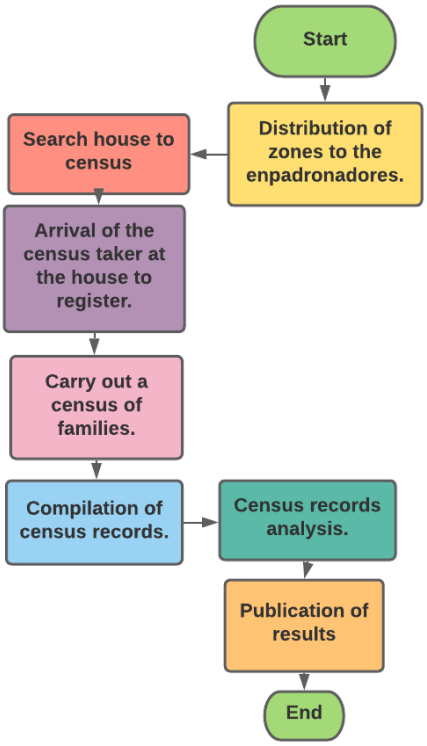


Fig. 13. Census Process without the Mobile Application

2) To-be: Fig. 14 shows the census process with the implementation of the mobile application.

As you can see, the change is very drastic, the improvement results in a shorter and more effective process, this means that the automation of this process is one of the best decisions that can be made, the process must be put aside. traditional census and Implement this new process because it avoids the loss and mixing of data, which benefits future projects that are carried out.

B. About the Methodology

For the methodology, Scrum was implemented for the development of this work, since it allows to manage technology projects and keep them in order, giving more value to the product through its deliverables and the constant participation of the client. The estimation and the correct priority of each user history give benefits to the Scrum team since thanks to these it is easier to achieve their objectives. In comparison with other methodologies such as RUP which is a traditional methodology and is used more for the analysis, implementation, and documentation of projects [19], but its

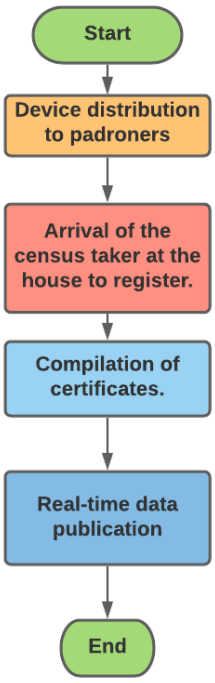


Fig. 14. Census Process with the Mobile Application

main disadvantage is that it is not subject to change in the full stage of development as scrum, also, scrum offers the client if the product is meeting the requirements that were raised from the planning, a rule that cannot be given in RUP, because you have to wait until the end of the phase to make changes, all of this is further detailed in Table X.

TABLE X. COMPARISON RUP - SCRUM

METHODOLOGY RUP	METHODOLOGY SCRUM
More documentation.	Knowledge needed to achieve a goal.
Changes are in a phase.	Involve from the beginning and everyone is given a role.
It is modeled by use case.	Deliverable on time.
It is closed in architecture guided by risk.	Reduces the cost of change at all stages.
In small projects, the costs of dedicating the necessary team of professionals may not be covered.	Visible, transparent by the specialist team.

VI. CONCLUSIONS AND FUTURE WORK

Finally in this research we managed to carry out the design of an application to automate the census process in Peru with the use of the Adobe Xd tool, since I give us a more realistic prototype to what mobile application will be like in its development. The Scrum framework was of great importance, as it allowed a development with more value towards the product and thus in the future it is hoped to complement this research article with the development of all the functionalities in order to make the mobile application a reality. And as future research work, we will expand by surveying everyone involved in the census process, on the other hand we want to complement the article with the development of this mobile application, as it is presented as a useful tool for the problem posed. It is hoped that this mobile application can continue to be developed by implementing new functionalities and giving it future changes according to the needs of the user and the organization.

REFERENCES

[1] O. O. Oluwagbemi, M. Keshinro, and C. K. Ayo, "Design and implementation of a secured census information management system," 2011.

[2] T. Snyder, "Us census explorer: A gui and visualization tool for the us census data api," 2019.

[3] R. Ryan, S. Vernon, G. Lawrence, and S. Wilson, "Use of name recognition software, census data and multiple imputation to predict missing data on ethnicity: application to cancer registry records," *BMC medical informatics and decision making*, vol. 12, no. 1, p. 3, 2012.

[4] J. I. G. Cruz, "Analysis of territorial impact of third tourist boom in canary islands (spain) through the application of a geographic information system (gis)," *Cuadernos de Turismo*, no. 36, p. 469, 2015.

[5] L. Piedelobo, D. Ortega-Terol, S. Del Pozo, D. Hernández-López, R. Ballesteros, M. A. Moreno, J.-L. Molina, and D. González-Aguilera, "Hidromap: a new tool for irrigation monitoring and management using free satellite imagery," *ISPRS international journal of geo-information*, vol. 7, no. 6, p. 220, 2018.

[6] E. V. Orozco, S. N. Campos, and I. S. Campos, "Use of new technologies for the 2020 population and housing census round," *Statistical Journal of the IAOS*, vol. 36, no. 1, pp. 83–88, 2020.

[7] T. Bourezgue, "Using tablets for the 2018 algerian census: Census data management and quality assessment," *Statistical Journal of the IAOS*, vol. 33, no. 3, pp. 777–784, 2017.

[8] B. G. Tavares, C. E. S. da Silva, and A. D. de Souza, "Analysis of scrum practices for risk treatment," *Product: Management and Development*, vol. 14, no. 1, pp. 38–46, 2017.

[9] J. F. Andry, R. E. Riwanto, R. L. Wijaya, A. A. Prawoto, and T. Prayogo, "Development point of sales using scrum framework," *Journal of Systems Integration*, vol. 10, no. 1, pp. 36–48, 2019.

[10] R. Coppola, L. Ardito, and M. Torchiano, "Characterizing the transition to kotlin of android apps: a study on f-droid, play store, and github," in *Proceedings of the 3rd ACM SIGSOFT International Workshop on App Market Analytics*, 2019, pp. 8–14.

[11] N. Obradovic, A. Kelec, and I. Dujlovic, "Performance analysis on android sqlite database," in *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*. IEEE, 2019, pp. 1–4.

[12] M. J. A. Ghali and S. S. Abu-Naser, "Its for data manipulation language (dml) commands using sqlite," 2019.

[13] A. Arcuri, J. Campos, and G. Fraser, "Unit test generation during software development: Evosuite plugins for maven, intellij and jenkins," in *2016 IEEE International Conference on Software Testing, Verification and Validation (ICST)*. IEEE, 2016, pp. 401–408.

[14] S. Casanova, "Estimación ágil con la técnica planning poker," *línea]. Disponible en: http://samuelcasanova. com/2016/01/estimacion-agil-con-latecnica-planning-poker/.[Accedido: 07-oct-2016].*

[15] A. N. Cadavid, J. D. F. Martínez, and J. M. Vélez, "Revisión de metodologías ágiles para el desarrollo de software," *Prospectiva*, vol. 11, no. 2, pp. 30–39, 2013.

[16] J. A. Álvarez-Bermejo, A. Codina-Sánchez, and L. J. Belmonte-Ureña, "Application architecture to efficiently manage formal and informal m-learning. a case study to motivate computer engineering students," *Dyna*, vol. 82, no. 190, pp. 113–120, 2015.

[17] A. Ponce, "Técnicas de procesamiento de datos en censos y encuesta," 2005.

[18] N. Wardrop, W. Jochem, T. Bird, H. Chamberlain, D. Clarke, D. Kerr, L. Bengtsson, S. Juran, V. Seaman, and A. Tatem, "Spatially disaggregated population estimates in the absence of national population and housing census data," *Proceedings of the National Academy of Sciences*, vol. 115, no. 14, pp. 3529–3537, 2018.

[19] L. J. Long, U. Eaganathan, and N. A. B. Sabri, "Adopting information system security services in online clothing marketing system using rup methodology under php," in *Journal of Physics: Conference Series*, vol. 1228, no. 1. IOP Publishing, 2019, p. 012062.

Augmented Reality Electronic Glasses Prototype to Improve Vision in Older Adults

Lilian Ocares Cunyarachi[1]
Facultad de Ciencias e Ingeniería
Universidad de Ciencias y Humanidades

Alexandra Santisteban Santisteban[2]
Facultad de Ciencias e Ingenieria
Universidad de Ciencias y Humanidades

Laberiano Andrade-Arenas[3]
Facultad de Ciencias e Ingenieria
Universidad de ciencias y Humanidades

***Abstract*—In this article, we focus on the elderly who suffer from low vision. We seek to design Augmented reality electronic glasses to help the elderly who suffer from vision problems, which causes a limitation when performing their daily activities, so this can affect their development in society causing serious physical and emotional damage, so we used a set of scientific articles that analyzed the percentage of visual impairment. Technology has demonstrated on numerous occasions that it can be a great ally for the health and well-being of the elderly. In this work, the objective is to design electronic glasses to help the elderly improve their vision. The methodology used is Design Thinking, which thanks to the phases of this methodology helps us to understand, collect information about the problem and give a solution, the result obtained is a prototype of electronic glasses in which it will benefit adults who suffer from low vision. As for the case study, we will show the design of the mobile application and the detailed development of the prototype.**

***Keywords*—Augmented reality; design thinking; electronic glasses; low vision; seniors**

I. Introduction

The present research refers to the benefits of electronic glasses in elderly people, in the branch of ophthalmology as well. The low vision problems are a limitation of the visual capacity that seriously affects people, consequently it prevents the patients to be able to carry out their daily activities in such a way that they do not manage to interact with the society [1].

The importance of research is to design electronic glasses, in order to improve the vision of people who suffer from problems of eye loss so that we can solve the problems of vision in older adults, consequently over the years suffer eye damage which does not allow them to see clearly, so these glasses will help people to be able to have better near vision [2]. This way the electronic glasses have implemented a camera with high resolution hd generating great luminosity that manages to capture the environment and all that where the person directs the intelligent glasses, in such a way that the image managed to process an algorithm in which it was adjusted and modified to detail each one of the images identifying separately according to the degree of affection of each one of them [3]. To close the idea the advantage of these electronic glasses will allow to project videos and to take pictures that are received through a connection via bluetooth or wifi of the connector HDMI that comes incorporated [4]. According to the World Health Organization [5], the estimated number of people with this visual disability is 253 million, of which 36 million are blind and 217 million have moderate-severe visual impairment [6]. It is important to note that 81% of people with blindness and visual disability are over 50 years old, which confirms that the vast majority of people who suffer from vision loss are adults [7].It is important to note that a detailed study was conducted in Peru, through a National Specialized Survey on Disability (Enedis). It was determined that the highest disability is in the visual limitation with a 48.3% [8]. As a result, the WHO and Enedis statistics on visual disability in Peru were known, with a high percentage of the elderly population [9].

The objective of our work is to be able to help older people with low vision by designing electronic augmented reality glasses that allow the capture of high quality images in real time, to improve visual perception and thus daily autonomy.

This paper is structured as follows: Section II will describe in detail the methodology used for the design. Section III will show the case study, Section IV will show the results and discussion and finally, Section V will present the conclusions.

II. Methodology

The methodology to be used will be Design Thinking, since it is a process originally oriented to the creation or improvement of products, which allows us to solve the problems reducing the risks and in such a way that increases the possibilities of success, focusing on the needs of the user, from then on the information about the topic or problem is understood and collected, then it is observed, the point of view of the team is given, a solution is devised, prototypes are made and finally the test in such a way that it is possible to connect the knowledge of different disciplines to arrive at a technically viable and economically profitable solution [10]. Design Thinking is an agile methodology for digital innovation, in such a way that it extracts the benefits of one and other methods that complement the challenges posed. The conformation of the concept to the agile evolution opens better opportunities and can be used in different agile iterations [11]. In Fig. 1 both the problem space and the solution space are mentioned, in the same way there is space to explore multiple options through the divergent phase.

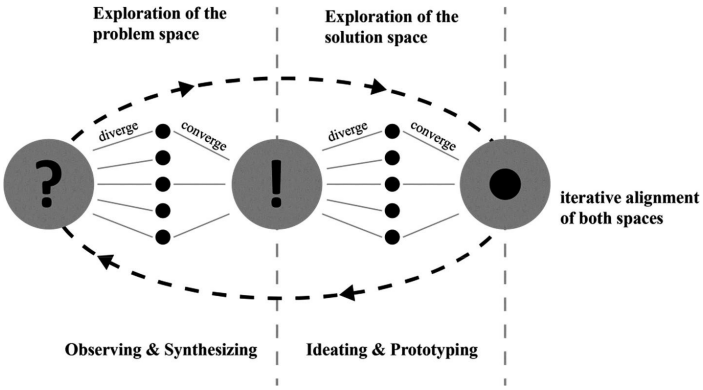


Fig. 1. Problem and Solution Space in Design Thinking

Phases of Design Thinking

1) Understand: The first phase is to understand, collect and analyze the information about the topic or problem to be solved, in order to understand the life of the users, as well as the different problems and needs they have.

2) Observe: The second phase collects and processes information on user needs through client interviews and the research.

3) Point of View: The third phase collects the information from the first two phases that is shared at the meeting and becomes a framework called a point of view.

4) Ideate: In the fourth phase, the development of an interdisciplinary team crosses through different processes that have been observed in detail and from then on a solution is designed with opportunities for change during implementation.

5) Prototype: This is the central phase of Design Thinking, a prototype is designed that is converted with ideas into action. The purpose of this phase is to convert the idea or solution into a digital or physical prototype.

6) Test: The last phase of the test is a thorough evaluation of the requirements and needs of the user. All these processes are highly iterative and can be repeated from any phase in which a completed product is delivered.

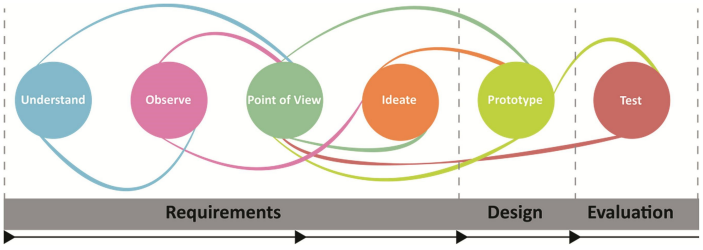


Fig. 2. Adapted Design Thinking Process Model

Fig. 2 refers to the proposed framework, the six processes of Design Thinking is adapted in the three phases as one of them is the requirement, Design and Evaluation. In the first four phases of Design Thinking, the focus is on gathering the necessary information to start a real implementation.

III. Case Study

With the methodology mentioned below, the processes will be explained Detailed according to the phases, for which electronic glasses will be designed to help older adults with low vision.

1) Understand: As a first step, the problem will be identified by adding a fact that globally, it is estimated that approximately 1.3 billion people live with some form of low vision at distance or near. With respect to distance vision, 188.5 million people have a moderate visual impairment, 217 million have a moderate to severe visual impairment and 36 million are blind [12]. On the other hand, 826 million people have an impairment of near vision [13].The vision problems in the older adults are frequent and the consequences that can be generated, arriving to affect seriously the vision, since it loses the capacity of visualization by different reasons as a result of the cataracts, glaucoma, presbyopia, macular degeneration, among others, consequently they affect the visual health which generates difficulties that cause in their states of spirit when being limited by the ocular problems. Electronic glasses make it easier for people to comfortably perform common multi-distance tasks such as reading, watching TV, doing crafts, interacting with people, viewing posters while moving around and much more [14].

2) Observe: As a second phase we collect information on the needs in older adults and on the visual pathological diseases that are mostly lacking in older people [15] that we will mention below:

Cataract: It is defined as an opacity in the lens and is a major cause of visual loss. The lens begins to deteriorate at 40 to 45 years losing its ability to accommodate what is known as presbyopia or tired eyesight, affecting virtually the entire population and requiring the use of glasses to focus on close objects [16].

Macular Degeneration: It is a frequent cause of severe and irreversible decrease in vision in elderly patients. The disease causes lesions in the central portion of the retina and in the macula which is responsible for the central vision necessary for reading or driving [17].

Glaucoma: Glaucoma is an irreversible disease of the optic nerve that is produced by an increase in intraocular pressure. The incidence of glaucoma increases with age and is therefore a risk factor for the disease, as are high myopia and the presence of a family history of glaucoma [18].

Palpebral Ptosis: In the third age, the musculature is losing tone and sometimes it generates a sagging of the eyelids that

can affect the vision [19].

Ocular Dryness: With age, it is common for the mucous membranes to dry out and this occurs with the lacrimal duct, causing problems of ocular dryness that can be very annoying and cause complications such as blepharitis, when the Meibomio glands of the eyelashes become obstructed, causing inflammation and oily secretion that can become infected. The treatment requires a daily and specific palpebral hygiene according to the affectation [20].

3) Point of View: From a general point of view, to address the relationship of ICT with adults we will use two perspectives: as active users who use them to do tasks of all kinds, like the rest of the population or as users receiving services to support their daily lives and thus provide the same opportunities to enjoy the benefits of electronic glasses and improve their lifestyle.

4) Ideate: The step to this phase begins by creating solutions to the problems found as we have been describing in the previous phases. We focus on our efforts to be able to understand and precise the information obtained. Next, we will present tools that will help us solve the needs and desires of the user.

Android Studio: It is a specific role for the development of applications, is a software in which provides different tools and services for different users to create and innovate with new applications, starting from the code to the design of the user interface. This tool is very simple for that reason facilitates users to understand the operation quickly. In this development environment we will find the necessary tools to create applications. It should be noted that Google revealed the new environment for the development of Android Studio app, one of the key parts of Android is its powerful code editor that is included with integrated elements such as Smart Editing ie provides a more decipherable code [21].

5) Prototype: In this phase we will show the ideal conceptual design of the glasses based on indicators with integral components as shown in Fig. 3. The screen mounted on top of the glasses, also includes peripheral components of the mounted camera, with a global positioning system (GPS) tracker, gyroscope, step detection sensors and Bluetooth headset. A rechargeable battery provides electrical power to the device. The Bluetooth headset facilitates the communication mechanism between the remote caregiver and the user. The implemented sensors collect detailed information in real time from the environment to the user to deliver it to the remote caregiver. The integrated camera captures the actual image showing the user's field of vision. Other essential raw data such as GPS coordinates and step detector data make it easier and better to locate the user for the remote caregiver.

The leds are implanted in the frame of the glasses and can blink in different combinations to provide navigation commands. The position of the indicators on the frame is shown in Fig. 4 to provide different combinations of the

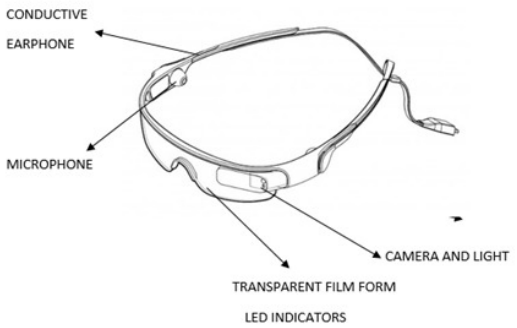


Fig. 3. Design of the Electronic Glasses

blinking indicator, in which seven indicators are implanted in the frame of each lens of the glasses. These indicators can blink in up to three different colors that come in red, green and finally yellow so that they generate navigation commands based on the traffic light image. The indicator combinations are integrated with a different blink pattern and are designed to transmit expressive navigation signals. Usability experiments evaluate the efficiency of these visual signals and qualify the different combinations in conjunction with patterns. It should be noted that the combination means that several indicators at once with an integrated pattern comes to make a group of adjustments by the frequency of blinking with the intensity of light [22].

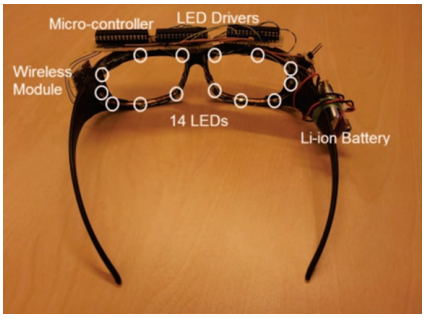


Fig. 4. LED Indicators Implanted in the Frame of the Glasses

In this study, an innovative concept is introduced to design a noble device to help elderly people with simple daily navigation tasks. The prototype of intelligent glasses based on indicators has been developed to provide users with visual signals that do not interfere with their vision of the physical environment. The diagram of the system and the integral components of the system are shown in Fig. 5. A lithium-ion battery generates a stable 3.3 V and provides electrical power to all components. The Android application is developed to establish a command pattern and activate individual or combined indicators to flash. The application connects to the Bluetooth serial port profile (SPP) module on the device.

To realize the usability in the built prototype of the electronic glasses, two Android applications are developed. Two applications communicate with the device as described above. The difference is the user interface of the application to set

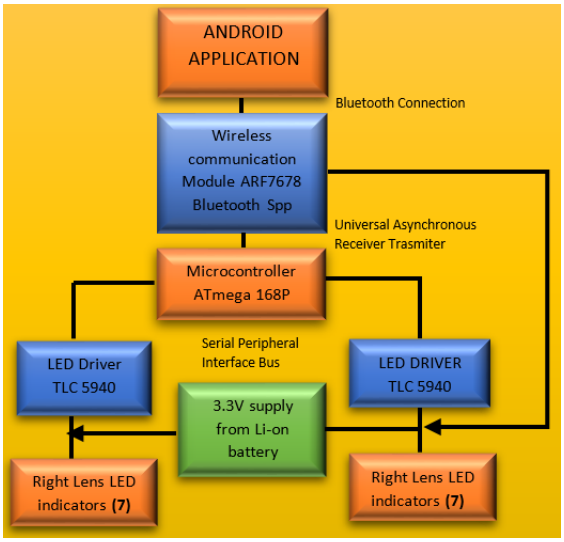


Fig. 5. Circuit Diagram Illustrates Component Connectivity

the blink pattern for individual indicators and the combination of indicators as navigation signals [23].In the first Android application, different indicators are set and sent to the device through the Bluetooth connection with colors, brightness and duration of the blink time. Fig. 6 shows the application's graphical user interface. In the second Android application, the navigation buttons are included in the graphical user interface on the other hand, Fig. 7 shows the user interface of the Android application to guide users in which specific visual signals can be assigned to each navigation button in the configuration menu [24].

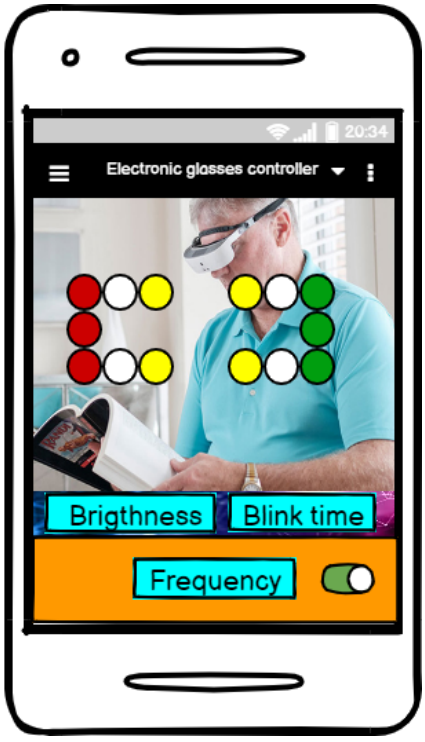


Fig. 6. Android Application

Fig. 7. Visual Commands

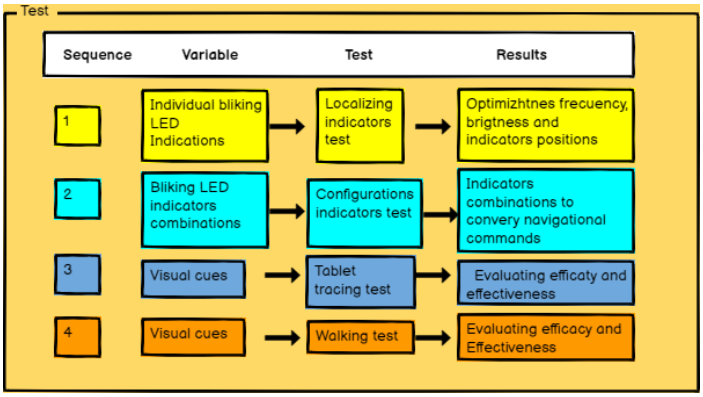


Fig. 8. Sequence of the Four Step Experiments

6) Test: In this last phase the usability of the prototype is tested, for real life situations in the users. We use the research method of design science with a qualitative and quantitative approach. Hevner et al.(2015) provided the guidance for the development of a design science artifact through iterative usability testing. The guide suggests early testing and usability experiments with real users, to evaluate and improve the system. For this purpose, we designed four-step experiments with different user groups [25].In Fig. 8, it shows the sequence of the experiments and specifies the factors and variables measured in the experiments. The first experiment is the location indicator test and aims to evaluate the visibility of the individual LEDs in the frame. The indicators are blinking individually. The result showed accuracy in recognizing individual blinking indicators for user display. In addition, the results clarify the satisfactory blinking pattern of the blinking indicators, including the optimized brightness and frequency.

IV. Results and Discussion

A. About the Case Study

For the case study, the design of electronic glasses to help the elderly was carried out so that users obtain electronic glasses that allow them to capture images in high quality and in real time to improve visual perception and with it, daily autonomy. the objective is to facilitate the elderly in their daily activities, so the equipment is designed so that they continue to have both hands free and good mobility. Balsamiq Mockups was used as the first tool for the design of prototypes, since it is a tool that allows to design in a more effective and simple way that is used to make mobile applications and interfaces and thus be able to obtain an ideal prototype [26]. However, for the creation of the mobile application we also use Android Estudio, which is a software that offers different tools and services so that different users can create and innovate with new applications, starting from the code to the design of the user interface. This tool is very simple for that reason it facilitates the users to be able to understand the handling quickly and to be able to make the creation of our mobile application.

B. About the Methodology

Design Thinking is a model of how innovations in insecure environments can be approached in an agile and radical way. Design Thinking has a series of tools that are used throughout the process of creating innovative products and services, depending on the stage you are at. Design Thinking can always be used because it is based on problem solving from the user's perspective [27]. The Design Thinking methodology was chosen because it is considered very suitable for proposing IT solutions in process models where prototypes are used to clarify the requirements to the satisfaction of the user or client. Some of these process models are used as prototypes for their solution. They allow a lot of interaction with users such as the creation and evaluation of the prototypes created.

– Advantages

One of the advantages of using this Design Thinking methodology is that it encourages the immediate solution of problems that may arise. It is for this reason that the user or client is placed as the center for the creation process, in which the development is clear to overcome challenges by providing guidance on the techniques and tools to be used to solve a challenge both in a company and in other environments. It is necessary to emphasize that this methodology is focused on people, this means that it is a customer-centered method whose important potential is to solve real problems that people face. Adding another advantage to this is that it focuses on adapting the solution to the product and service on the needs of the user in that sense you get good results.

– Disadvantages

The disadvantage of the Design Thinking methodology is that it cannot be used for all types of projects.

– Comparison

Design Thinking is not only used by organizations or entrepreneurs, in such a way that large and prestigious companies have discovered its benefits and have incorporated them into their daily work. Design Thinking places clients as the main focus for the creation process as opposed to other approaches that try to move from thought to action. Design thinking invites the consumer to assume a more active role in the design of the required product, therefore, it implies that the agents are in dialogue between the client and the developer. The first approach of a team in innovation, it is important to make it in a guided way and focused on a specific problem to apply the Design Thinking methodology in this way a process is proposed where the phases can be taken up again and repeated without major restrictions, while in Scrum, you can work with a constant flow of projects that must respond to the priorities of the business, this is how there is a series of events or meetings to follow in each Sprint. This is where there can be a contrast, however, some elements can be taken from Scrum such as prioritization, or organizational aspects can be adopted that contribute to teamwork.

V. Conclusions

In conclusion, the design of the electronic glasses will help the elderly to be able to comfortably perform the usual tasks, thanks to these electronic glasses the elderly will no longer escape from their reality and will not have any interruption or external stimulus that will prevent them from enjoying this new technological experience. On the other hand, the use of the Design Thinking methodology was fundamental, since it is focused on the user and the problems that may arise, which made the development of the prototype design and application possible. With this design of electronic glasses, it is intended in the future to have the implementation of the software in order to be able to implement it in Peru in that way giving a better quality of life to the elderly since the use of electronic devices is increasingly accepted, for the benefit of elderly patients with low vision, allowing people with partial visual disability to see more clearly.

References

[1] R. Peleg-Adler, J. Lanir, and M. Korman, "The effects of aging on the use of handheld augmented reality in a route planning task," *Computers in Human Behavior*, vol. 81, pp. 52–62, 2018.

[2] M. L. Bianco, S. Pedell, and G. Renda, "Augmented reality and home modifications: a tool to empower older adults in fall prevention," pp. 499–507, 2016.

[3] S. Gopalakrishnan, S. C. Suwalal, G. Bhaskaran, R. Raman *et al.*, "Use of augmented reality technology for improving visual acuity of individuals with low vision," *Indian Journal of Ophthalmology*, vol. 68, no. 6, p. 1136, 2020.

[4] H. P. van der Aa, H. C. Comijs, B. W. Penninx, G. H. van Rens, and R. M. van Nispen, "Major depressive and anxiety disorders in visually impaired older adults," *Investigative ophthalmology & visual science*, vol. 56, no. 2, pp. 849–854, 2015.

[5] E. R. Vieira, F. Civitella, J. Carreno, M. G. Junior, C. F. Amorim, N. D'Souza, E. Ozer, F. Ortega, and J. A. Estrázulas, "Using augmented reality with older adults in the community to select design features for an age-friendly park: A pilot study," *Journal of Aging Research*, vol. 2020, 2020.

[6] C. Y. Huang, J. B. Thomas, A. Alismail, A. Cohen, W. Almutairi, N. S. Daher, M. H. Terry, and L. D. Tan, "The use of augmented reality glasses in central line simulation:"see one, simulate many, do one competently, and teach everyone"," *Advances in medical education and practice*, vol. 9, p. 357, 2018.

[7] M. Ibarra Yancan de Villegas and J. Montes Pariona, "Efectividad del programa educativo "salud ocular" en el autocuidado en cuidadores del adulto mayor postoperado de catarata en el instituto nacional de oftalmología 2018," 2018.

[8] Y. Zhao, M. Hu, S. Hashash, and S. Azenkot, "Understanding low vision people's visual perception on commercial augmented reality glasses," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 4170–4181.

[9] F. Mostajeran, F. Steinicke, O. J. Ariza Nunez, D. Gatsios, and D. Fotiadis, "Augmented reality for older adults: Exploring acceptability of virtual coaches for home-based balance training in an aging population," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.

[10] C. Wrigley, E. Nusem, and K. Straker, "Implementing design thinking: Understanding organizational conditions," *California Management Review*, vol. 62, no. 2, pp. 125–143, 2020.

[11] K. Gurusamy, N. Srinivasaraghavan, and S. Adikari, "An integrated framework for design thinking and agile methods for digital transformation," in *Design, User Experience, and Usability: Design Thinking and Methods*, A. Marcus, Ed. Cham: Springer International Publishing, 2016, pp. 34–42.

[12] T. Fonseca Borges, L. M. Gil Morales, and Y. Ruiz Vega, "Impacto social de la atenciA visual en personas geritricas," *Conrado*, vol. 15, pp. 425 – 428, 12 2019.

[13] J. Ramke, A. B. Zwi, A. Palagyi, I. Blignault, and C. E. Gilbert, "Equity and blindness: closing evidence gaps to support universal eye health," *Ophthalmic epidemiology*, vol. 22, no. 5, pp. 297–307, 2015.

[14] G. T. Liu, N. J. Volpe, and S. L. Galetta, "3 - visual loss: Overview, visual field testing, and topical diagnosis," in *Liu, Volpe, and Galetta's Neuro-Ophthalmology (Third Edition)*, third edition ed., G. T. Liu, N. J. Volpe, and S. L. Galetta, Eds. Elsevier, 2019, pp. 39 – 52.

[15] C. Creuzot-Garcher, C. Binquet, S. Daniel, L. Bretillon, N. Acar, A. de Lazzer, L. Arnould, C. Tzourio, A. M. Bron, and C. Delcourt, "The montrachet study: study design, methodology and analysis of visual acuity and refractive errors in an elderly population," *Acta ophthalmologica*, vol. 94, no. 2, pp. e90–e97, 2016.

[16] I. V. Malov, Y. V. Bantsykina, V. M. Malov, E. B. Eroshevskaya, and O. V. Pavlova, "Cataract problems in patients with high myopia (clinical case)," *Aspirantskiy Vestnik Povolzhiya*, vol. 19, no. 5-6, pp. 67–71, 2019.

[17] R. Ratnapriya, O. A. Sosina, M. R. Starostik, M. Kwicklis, R. J. Kapphahn, L. G. Fritsche, A. Walton, M. Arvanitis, L. Gieser, A. Pietraszkiewicz *et al.*, "Retinal transcriptome and eqtl analyses identify genes associated with age-related macular degeneration," *Nature genetics*, vol. 51, no. 4, pp. 606–610, 2019.

[18] J. I. Orlando, H. Fu, J. B. Breda, K. van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee *et al.*, "Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Medical image analysis*, vol. 59, p. 101570, 2020.

[19] F. De Marchi, L. Corrado, E. Bersano, M. Sarnelli, V. Solara, S. D'Alfonso, R. Cantello, and L. Mazzini, "Ptosis and bulbar onset: an unusual phenotype of familial als?" *Neurological Sciences*, vol. 39, no. 2, pp. 377–378, 2018.

[20] J. Yang, "Composition for the prevention and/or improvement of dry eye syndrome, dryness of the eye, and stiff neck and shoulder," May 14 2020, uS Patent App. 16/740,843.

[21] T. Hagos, M. Zechner, J. DiMarzio, and R. Green, "What's in an android application," in *Beginning Android Games Development*. Springer, 2020, pp. 45–54.

[22] Y. Kashimoto, A. Firouzian, Z. Asghar, G. Yamamoto, and P. Pulli, "Twinkle megane: Near-eye led indicators on glasses in tele-guidance for elderly," in *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. IEEE, 2016, pp. 1–6.

[23] S. K. Rao, R. K. Rao, and R. K. Rao, "Image and augmented reality based networks using mobile devices and intelligent electronic glasses," Jun. 28 2016, uS Patent 9,380,177.

[24] S. Machida and S. Kouji, "Control device for variable focus lenses, control method for variable focus lenses, and electronic glasses," Jan. 10 2017, uS Patent 9,541,774.

[25] W. Diyatmika, J. P. Chu, B. T. Kacha, C.-C. Yu, and C.-M. Lee, "Thin film metallic glasses in optoelectronic, magnetic, and electronic applications: a recent update," *Current Opinion in Solid State and Materials Science*, vol. 19, no. 2, pp. 95–106, 2015.

[26] H. Plattner, C. Meinel, and L. Leifer, *Design thinking research: Taking breakthrough innovation home*. Springer, 2016.

[27] E. Knight, J. Daymond, and S. Paroutis, "Design-led strategy: how to bring design thinking into the art of strategic management," *California Management Review*, vol. 62, no. 2, pp. 30–52, 2020.

Clustering-Based Hybrid Approach for Multivariate Missing Data Imputation

Aditya Dubey[1]
Department of Computer Science and Engineering
Maulana Azad National Institute of Technology
Bhopal, India

Akhtar Rasool[2]
Department of Computer Science and Engineering
Maulana Azad National Institute of Technology
Bhopal, India

***Abstract*—In the era of big data, a significant amount of data is produced in many applications areas. However due to various reasons including sensor failures, communication failures, environmental disruptions, and human errors, missing values are found frequently These missing data in the observed data make a challenge for other data mining approaches, requiring the missed data to be handled at the preprocessing stage of data mining. Several approaches for handling the missing data have been proposed in the past. These approaches consider the whole dataset for making a prediction, making the whole imputation approach to be cumbersome. This paper proposes the procedure which makes use of the local similarity structure of the dataset for making an Imputation. The K-means clustering technique along with the weighted KNN makes efficient imputation of the missed value. The results are compared against imputations by mean substitution and Fuzzy C Means (FCM). The proposed imputation technique shows that it performs better than other imputation procedures.**

***Keywords*—Clustering; imputation; KNN; missing at random; multivariate**

I. Introduction

Since the age of big data began, the collection of data from various sources, and the resultant amount of data has risen to the greatest extent [1]. Multivariate datasets are prevalent in several real-world applications, such as electrical system analysis, meteorological or economical strategy planning, security control, and plenty more. In several application areas, multiple sensors are deployed to produce datasets, and they typically have one target to generate the data as activity occurs. For example, in a power grid application several sensors diagnosing the state of power transformers, produce the data by monitoring the state of gases over time [2]. In the era of IoT, a vast number of sensors are utilized for generating the multivariate environmental conditions, for example, the air or water pollution [3]. In biomedical, numerous devices can also be fitted in working areas to track the health and overall well-being of senior citizens, which also ensures that adequate medicine is suggested. Important information and the facts can be obtained deriving these datasets [4]. Preprocessing is one of the steps for analyzing the data. One major issue handled in the preprocessing step is missed value. Unfortunately, the raw dataset generated by the sensor network typically includes missing values due to the rough working conditions or uncontrolled variables such as adverse weather conditions, malfunctions of the infrastructure, or unstable signals. The problem of missing data is quite prevalent in many applications. The incomplete dataset is inadvertently and uncontrolled by the researcher. The outcome is that the data observed cannot be evaluated due to the incompleteness of the datasets.

Several studies have suggested strategies for handling incomplete values in the dataset. The methods for dealing with the missed values may be categorized into three types. The first approach is to disregard the complete record which contains any missing value. Additionally, replace the missed value with zero or mean of the attribute [5]. The major downside of these strategies is that they decrease the efficiency of estimation. By excluding any usable data in those cases having any missing values. This could degrade the expected result. The second is to determine the values that use combinations of the Expectation-Maximization method. The third approach is imputation, which requires the process of completing the incomplete values in the dataset by some potential values, depending on the details in the dataset.

Missing data imputation strategies can be categorized into two types depending on the method of approximating the missing values. The first type, mathematical or predictive methods are used to estimate missing values. These approaches are remarkably simple, replacing each missed value with the mean or mode value of the variable, as well as more complex methods focused on advanced statistical techniques. The second is the imputation based on machine learning that utilizes the dataset knowledge to model the calculation of missing data. It involves a number of methods K nearest neighbor [6], MLP imputation [7], auto-associative neural network imputation [8], SOM imputation [9], recurrent neural network [10] and multi-task networks [12].

The paper is summarized as: Missing value issue is formulated and presented in Section 2. Section 3 describes the literature of the research area by focusing on major imputation procedures. The proposed technique is subsequently implemented in Section 4. In Section 5, the proposed technique is used on the benchmark datasets. The last section includes the conclusion and future work. The analytical results demonstrate that the proposed method works better than other conventional imputation techniques. This research paper has the following contributions:

- The proposed technique considers the local data similarities and introduces a local imputation model that uses a clustering technique for estimating the missing values. In other terms, the complete mechanism includes clustering is performed first and then imputation is done.

	Att_1	Att_2	Att_3	Att_4	Att_5
x_1	?	100	2.2	1	2.7
x_2	1	30	?	0	2.9
x_3	?	90	1.8	1	3.0
x_4	2	30	1.3	0	3.5
x_5	?	20	0.3	0	2.8

Fig. 1. A Dataset Containing Missing Values

- A top KNN distance weighted imputation enhancing the prediction accuracy is utilized to predict the missing value in each cluster.
- The proposed method is implemented and verified using datasets of the UCI dataset repository [11]. For big data analysis, the proposed method is implemented and verified on the MATLAB platform.

II. Problem Formulation

This paper is aimed to focus on making accurate predictions of missing data. For illustrating this problem more specifically, figure 1 depicts a data set having many missing values in some of the attributes $Att_1, Att_2, ..., Att_5$. Let $X = \{x_1, x_2, x_3, ..., x_N\}$ denotes the data collected from N different data sources and j^{th} attribute from i^{th} source can be denoted as X_{ij} for i=1,2,3,...,N, j=1,2,3,...,M. In this paper, the missing value is denoted as '?'. Additionally, to represent whether the value of X is missing or not, an indicator matrix H is used.

$$H_{ij} = \begin{cases} 1, \text{ if value of } X_{ij} \text{ is present} \\ o, \text{ otherwise} \end{cases} \quad (1)$$

III. Literature Review

This section provides a short overview of research on the basic algorithms used for imputation.

Abdella et al.'s research were focused on the use of neural network combined with the genetic algorithms to make imputation of the missing values in the dataset [13]. For training the neural networks Multi-Layer Perceptron and Radial Basis Functions are utilized. Li et al. employed a soft computation-based clustering technique to efficiently handle inaccurate and unconstrained dataset in addition to this fuzzy clustering algorithm handles missed data [14]. Liao et al. provided a fuzzy k-means clustering technique using a sliding window to impute missing data so that the quality of the dataset can be improved [15]. Pelckmans et al. suggested a method not to rebuild missing data but by utilizing support vector machines, the effects of the incomplete dataset over the result and anticipated cost are simulated [16]. The procedure consists of assuming certain models for the covariates of missing data and using the maximum likelihood method to get the predictions for these models. The benefits of this technique are that even if missing data is present among the input variables, classification rules can be derived from the observed values. While the result is that the proposed technique is designed for better classification accuracy rather than improving the imputation accuracy for missing data. Lim et al. suggested a hybrid neural network technique that utilizes the ARTMAP and fuzzy c-means clustering for classifying the patterns utilizing the incomplete training and testing dataset [17]. Fuzzy ARTMAP has the drawback of higher susceptibility to arrange the training data. For fuzzy ARTMAP it is very crucial to select the vigilance parameters because it can be hard to determine the optimal value of the vigilance parameters.

Hathaway et al. proposed a clustering method relying on the dissimilarity of missing data [18]. A benefit of this technique is that for missing data, fuzzy c-means is considered to be a reliable clustering method. The support vector regression technique used by Feng et al. was also used to predict the missing data of DNA microarray gene expression using an orthogonal coding scheme [19]. Comparison research to the earlier established techniques for their imputation, such as KNN and BPCA, showed that the SVR procedure was efficient for imputation. One major benefit of the SVR model is that it takes less time for computation, but the hybrid SVR clustering process produces more sensible outcomes for the dataset having outlier data. Timm et al. observed that the dataset having missing values is an important issue in the data analysis [20]. The class-specific possibility of missing values was implemented to properly disseminate the incomplete data points to the clusters. Farhangfar et al. proposed an extensive analysis of representative imputation methods [21]. They stated that the usage of a low-quality single imputation method resulted in prediction accuracy comparable to the accuracy of utilizing some other advanced imputation method.

Li et al. presumed that missed data are defined in terms of intervals and introduced a novel fuzzy c-means based procedure for handling missing data based on nearest neighbor intervals concept [22]. The drawback of this technique is that the number of clusters is not selected on a theoretical basis, so further procedure is required to examine this issue. Nuovo performed a comparative study on imputation done using fuzzy c-means against the imputation done using case deletion [23]. The methods are compared using a mentally retarded patient dataset in psychological research environmental conditions. The research shows that imputation methods especially the FCM based method provide efficient imputation of data and prevent the deletion of missing data which causes a reduction in the strength of research. The fuzzy c-means imputation makes more improved predictions than that of the regression imputation technique and expectation-maximization technique. One main drawback includes that FCM imputation utilizes a weighting parameter, which values to 2, which requires to be adjusted according to the dataset type. L. Zhao et al. make use of the local similarity imputation procedure for imputing the missing value in the cyber-physical system. The procedure includes the usage of a two-layered stack encoder together with the concept of KNN [24]. Q. Ma et. al in his research addresses the issue of the insufficient complete data subset to make an imputation from its clustered neighbors [25]. Ordered sensitive imputation for clustered missing data makes use of previously imputed value to be used for the next iteration of imputation. Two regularized learning algorithms have been utilized in the method proposed by A. Wang et. al. for imputing the missing value in microarray experiments on gene expression [26]. The RLLSimpute L2 regularized local least square is trained on the target gene and its neighbors so that the missing value can be

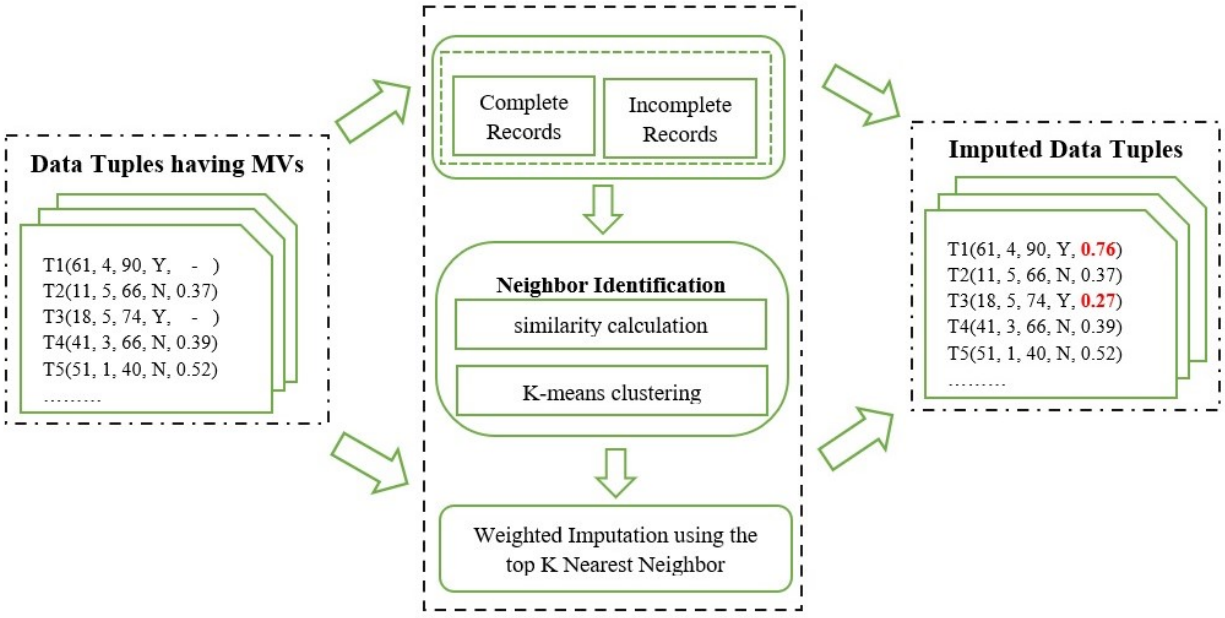


Fig. 2. Block Diagram of Proposed Method

imputed.

A. Types of Missing Data

Missing data have three kinds of missing patterns.

1) Missing completely at random (MCAR)- The missed data does not have any dependability on any other data. In other words, the probability of missing data is equal to all the units [27], [29].
2) Missing at random (MAR) - The missed data depends upon the available data. These available data can be utilized for estimating the missed data.
3) Missing Not at Random (MNAR)- The analysis of missed data relies upon other missing data which causes the missed data to become unpredictable.

IV. Proposed Work

A basic issue in missing data imputation is the retrieval of lost value by utilizing the available dataset information. Clustering is one of the most common data mining methods which arise to fix this problem [28]. The general goal of this clustering is to split available data into multiple desired clusters by recognizing the similarity of objects. The principle is to minimize the intra-cluster similarity and maximize the inter-cluster dissimilarity. Figure 2 describes the imputation procedure, in which the first step is to partition the dataset into complete and incomplete records according to the missing data.

K-means approach is used as a clustering approach consisting of a four-step procedure. The first step is the random selection of a fixed number of cluster centroids. The second step is the assignment of each record to a certain cluster having the closest centroid. The third step is the recalculation of the cluster centroid. The last step states to iteratively repeat the procedure from step second if the algorithm does not reach the termination condition.

The last step in the imputation is to use the cluster information and provide value for each non-reference attribute having an incomplete object. Objects falling in the same cluster are considered as the nearest neighbor of missing values having higher similarity and based on the nearest neighbors missing values are imputed.

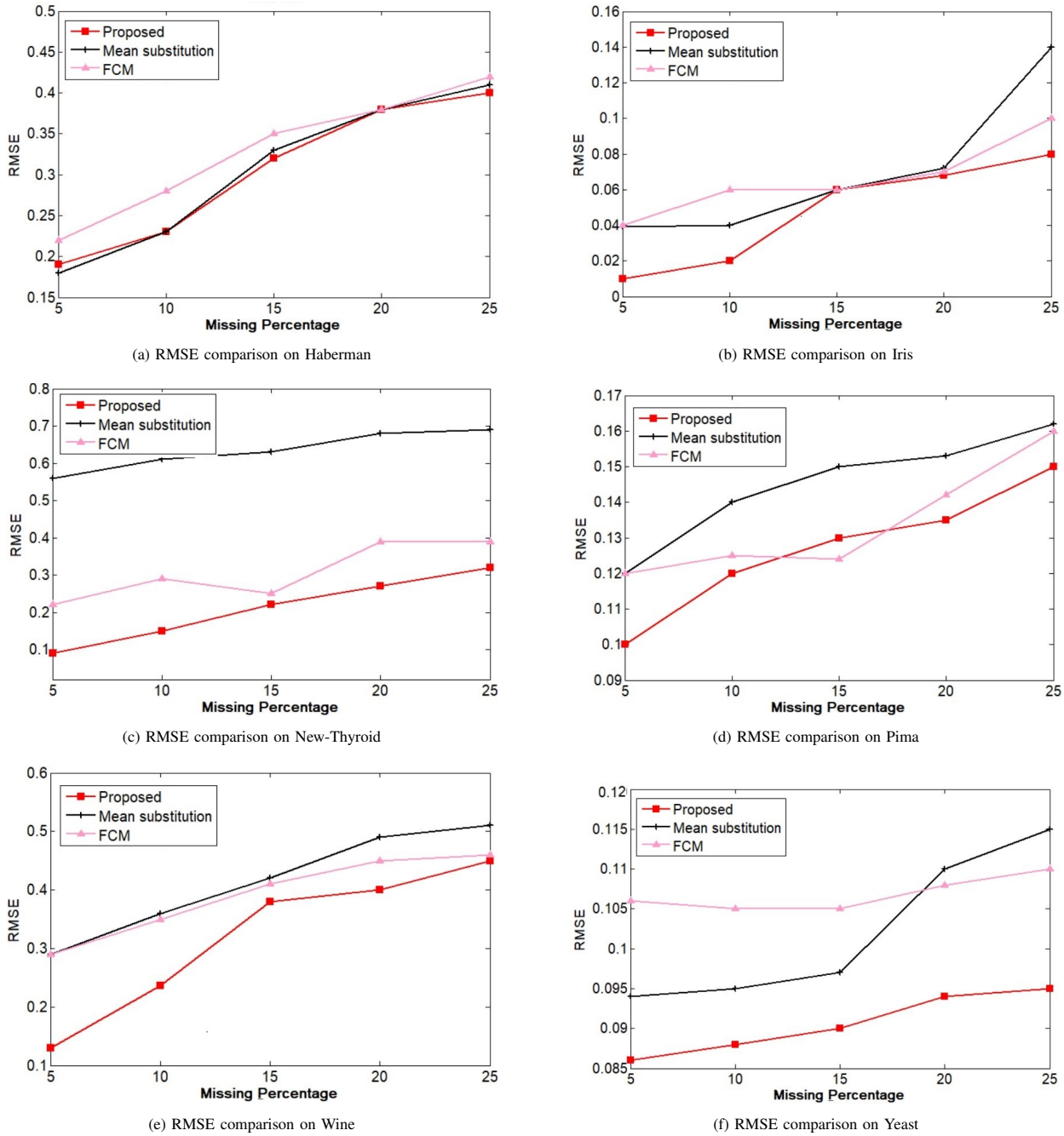
TABLE I. Summary of the Datasets

Name of dataset	Number of Instances	Number of Attributes
Haberman	306	3
Iris	150	4
New-Thyroid	215	5
Pima	768	9
Wine	178	13
Yeast	1484	8

V. Experiments and Analysis

A. Experimental Design

For illustrating the efficacy of the proposed method, extensive experiments are conducted on six UCI datasets [11]. To decide how the technique generalizes, experiments with more datasets, dealing with various numbers, and different types of missing value patterns are required. For performing experiments, some percentage of the data is deleted so that they have 5%, 10%, 15%, 20%, 25% of missing ratio in the dataset. Table I demonstrates the datasets used in this paper.



(a) RMSE comparison on Haberman

(b) RMSE comparison on Iris

(c) RMSE comparison on New-Thyroid

(d) RMSE comparison on Pima

(e) RMSE comparison on Wine

(f) RMSE comparison on Yeast

Fig. 3. RMSE comparison of Mean substitution, FCM and Proposed technique on (a) Haberman (b) Iris (c) New-Thyroid, (d) Pima (e) Wine (f) Yeast

B. Performance Evaluation

Several efficiency procedures have been used for estimating the predictive performance and comparison of distinct models [30]. In every definition that comes ahead, D_i is the actual value, P_i is the predicted value, and error $E_i = D_i - P_i$. Root Mean Square Error (RMSE) represents a root mean square deviation of the predicted values. Since the negative and positive signed errors that have been recorded do not compensate each other, RMSE provides an overview of the error during prediction. RMSE points out that the overall prediction error is greatly influenced by big errors that are considerably more costly than tiny errors. RMSE is also subject to scale shift and data transformations happened at the pre-processing. The less the RMSE, the stronger will be the prognosis. For n number of missing values, RMSE is defined as-

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} Err_i^2} \tag{2}$$

Fig. 3(a)-3(f) represents a comparison of experimental results that imply that the proposed technique is advantageous

in terms of imputed error. Ten different sets of the same missing percentage are created, all three imputation procedures are applied on these sets, thereafter the average of that ten RMSE is calculated. The lower the RMSE, the better will be the imputation. As the missing percentage increases, it becomes a challenge for each technique to make accurate imputation resulting in the increment of RMSE. For the small percentage of missingness, the techniques may have a relatively small difference in their performance but as the missing percentage increase, there exists a clear difference in the performance of each technique. For comparing the performance of the proposed technique, mean substitution and FCM-based imputation are utilized. Experimental results demonstrate that the proposed technique performs much better as compared to two other imputation technique.

VI. Conclusion

In this paper, a local similarity-based hybrid imputation approach utilizing the K-means clustering has been implemented which increases prediction accuracy. Experimental results on natural datasets show that the proposed method dominates over other existing prediction approaches. The analysis of accurate parameters and operational framework for the efficient implementation of the proposed technique has to be done for maximal technique execution. The kind of missing pattern that is MCAR, MAR or NMAR also plays an important role in the imputation procedure. In future, different missing pattern are experimented with the proposed imputation procedure. In addition, more experiments are also required with the larger number of attributes and records. At last, it can be concluded that the proposed technique makes an accurate prediction with higher efficiency.

References

[1] L. A. Kurgan, K. J. Cios, M. Sontag and F. J. Accurso, “Mining the cystic fibrosis data”, Next Generation of Data-Mining Applications, 2005, pp. 415–444.

[2] J. Barnard and X. L. Meng, “Applications of multiple imputation in medical studies: From aids to nhanes”, Stat. Methods Med. Res. vol. 8, no. 1, 1999, pp. 17–36.

[3] K. J. Cios and G. Moore, “Uniqueness of medical data mining”, Artif. Intell. Med., vol. 26, no. 1/2, 2002, pp. 1–24.

[4] A. Dubey and A. Rasool, ”Data Mining based Handling Missing Data”, Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2019, pp. 483-489.

[5] A. Dubey and A. Rasool, ”Time Series Missing Value Prediction: Algorithms and Applications”, Information, Communication and Computing Technology (ICICCT), vol. 1170, 2020, pp. 21-36.

[6] G. Batista and M. C. Monard, “Experimental comparison of k-nearest neighbour and mean or mode imputation methods with the internal strategies used by c4.5 and cn2 to treat missing data”, University of Sao Paulo, 2003, pp. 1–97.

[7] P. Sharpe and R. Solly, “Dealing with missing values in neural network-based diagnostic systems”, Neural Computation Applications, vol. 3, 1995, pp. 73–77.

[8] D. Pyle, “Data preparation for data mining”, morgan kaufmann publishers inc. San Francisco, vol. 22, no. 2, 1999, pp. 115–170.

[9] T. Kohonen, “Self-organizing maps”, Springer notes 3rd edn., 2001, pp. 347–371.

[10] Y. Bengio and F. Gingras, “Recurrent neural networks for missing or asynchronous data”, Adv. Neural Inf. Process Syst., vol. 8, 1995, pp. 395–401.

[11] https://archive.ics.uci.edu/ml/datasets.

[12] T. Kohonen, “Self-organizing maps”, Springer notes 3rd edn., 2001, pp. 347–371.

[13] M. Abdella and T. Marwala, “The use of genetic algorithms and neural networks to approximate missing data in database”, Comput. Inform. vol. 24, 2005, pp. 577–589.

[14] D. Li, J. Deogun, W. Spaulding and B. Shuart, “Towards missing data imputation: a study of fuzzy k-means clustering method”, Rough Sets Curr. Trends Comput., 2004, pp. 573–579.

[15] Z. Liao, X. Lu, T. Yang and H. Wang, “Missing data imputation: a fuzzy k-means clustering algorithm over sliding window”, Fuzzy Syst. Knowled. Discovery, vol. 14, 2009, pp. 133–137.

[16] K. Pelckmans, J. D. Brabanter and J. Suykens, “Handling missing values in support vector machine classifiers”, Neural Networks, vol. 18, 2005, pp. 684–692.

[17] C. P. Lim, J. H. Leong and M. M. Kuan, “A hybrid neural network system for pattern classification tasks with missing features”, IEEE Trans. Pattern Anal., vol. 27, 2005, pp. 648–653.

[18] R. Hathaway and J. Bezdek, “Clustering incomplete relational data using the non-euclidean relational fuzzy c-means algorithm”, Pattern Recogn. Lett., vol. 23, 2002, pp. 151–160.

[19] X. Wang, A. Li, Z. Jiang and H. Feng, “Missing value estimation for dna micro-array gene expression data by support vector regression imputation and orthogonal coding scheme”, Bmc Bioinform., vol. 7, no. 32, 2006, pp. 1–10.

[20] Timm, C. Doring and R. Kruse, “Different approaches to fuzzy clustering of incomplete datasets”, Int. J. Approx. Reason., vol. 35, 2004, pp. 239–249.

[21] A. Farhangfar, L. Kurgan and W. Pedrycz, “A novel framework for imputation of missing values in databases”, IEEE Trans. Syst. Man. Cybernet., vol. 37, no. 5, 2007, pp. 692–709.

[22] D. Li, H. Gu and L. Zhang, “A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data”, Expert Syst. Appl., vol. 37, 2010, pp. 6942–6947.

[23] A. D. Nuovo, “Missing data analysis with fuzzy c-means: a study of its application in a psychological scenario”, Expert Syst. Appl., vol. 38, 2011, pp. 6793–6797.

[24] L. Zhao, Z. Chen, Z. Yang, Y. Hu and M. S. Obaidat, “Local similarity imputation based on fast clustering for incomplete data in cyber-physical systems”, IEEE Systems Journal, vol. 12, no. 2, 2018, pp. 1610–1620.

[25] Q. Ma, Y. Gu, W. C. Lee and G. Yu, “Order-sensitive imputation for clustered missing values”, IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 1, 2019, pp. 166–180.

[26] A. Wang, Y. Chen, N. An, J. Yang, L. Li and L. Jiang, “Micro array missing value imputation: A regularized local learning method”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 3, 2019, pp. 980–993.

[27] Q. Ma, Y. Gu, W. C. Lee and G. Yu, “Order-sensitive imputation for clustered missing values”, IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 1, 2019, pp. 166–180.

[28] A. Ng, M. Jordan and Y. Weiss, “On spectral clustering: Analysis and an algorithm”, Advances in Neural Information Processing Systems, vol. 14, 2002, pp. 849–856.

[29] Mellenbergh, Gideon J, Missing Data, In Counteracting Methodological Errors in Behavioural Research, 2019, pp. 275- 292.

[30] Chai T, Draxler R, Root mean square error (rmse) or mean absolute error (mae), Geosci Model Dev Discuss, 7, 2014, 1525-1534.

Design of a Mobile Application for the Learning of People with Down Syndrome through Interactive Games

Richard Arias-Marreros[1]
Facultad de Ciencias e Ingeníeria
Universidad de Ciencias y Humanidades

Keyla Nalvarte-Dionisio[2]
Facultad de Ciencias e Ingeníeria
Universidad de Ciencias y Humanidades

Laberiano Andrade-Arenas[3]
Facultad de Ciencias e Ingeniería
Universidad de Ciencias y Humanidades

***Abstract*—The research work is focused on people with Down syndrome since it is the most common genetic disorder worldwide, also, these people have cognitive and visual-motor disabilities, however, the Peruvian state does not use even 1% of the budget allocated to the educational sector of these people, so, by not receiving the education they can not develop their skills. Therefore, a prototype of a mobile application was designed for the learning of people with Down syndrome through interactive games, implementing the Scrum methodology for the development of the application prototype, using the Troncoso method for the teaching of reading and writing, but the teaching of visual-motor coordination was added to the method and concerning the design of the prototype, the Balsamiq tool was used because it was the most appropriate. And so the objective of developing the prototype of the application was achieved. Having as a result that people with Down syndrome can read or write basic words and differentiate the hemispheres of their body, through the unlimited attempts of the exercises, in each level or type of learning. In this way, with the teachings received, these people will have a better quality of life, being able to integrate into society, and be more independent when performing daily activities.**

***Keywords*—Application of games; cognitive disabilities; Down's Syndrome; scrum methodology; Troncoso Method**

I. Introduction

Currently, Down syndrome is the most common genetic disorder worldwide and is one of the main causes of intellectual disability that affects people, generating a major problem in the functionality and independence to perform daily activities [1]. It can also be said that the visual-motor cognitive ability is one of the most sensitive, and it is a fact that this cognitive ability can be improved in people with Down syndrome [2].

In Europe, in the period 2011 to 2015, there is an estimated annual 7,800 born with Down syndrome, which corresponds to a predominance of 9.8 per 10,000 births [3]. Similarly, in Peru, according to Conadis, in 2015 there will be a total of 141,731 people with disabilities, of which 8,800 suffer from Down's syndrome, 6.21% of the total number registered, and the number is increasing every year [4].

A characteristic of people with Down syndrome is that they may have difficulty learning numbers compared to letters. They need personalized learning because their learning rate is slower [5]. However, these people not only face a motor and intellectual disability but also barriers that can hinder their full and effective participation in society [6]. Besides, people with Down syndrome have different levels of IQ. Some of them don't even know how to speak and interact, but some are smart and can learn well [5].

Because of the above, there are several approaches to teaching different skills to students with learning and behavior problems, for example, one method to teach people with Down syndrome is the method of Troncoso that is designed to recognize written words as a whole rather than divide them into letters and its philosophy is to recognize the meaning of the symbols that are written [7]. However, the use of information technology in special education can improve the ability in the learning process, beyond being entertainment is a teaching tool that will help you in the development of cognitive and visual-motor skills [6].

The application will help people with Down syndrome because they do not have much support in the educational sector. After all, in-state schools there are no people trained to teach, apart from that, the use of technology as a learning tool will help the development of cognitive ability, since you would be learning, but at the same time playing.

Also, this game will be a complement to conventional motor therapies, which aim to provide support to professionals in the area and generate performance reports for further analysis and thus monitor the process of developing their skills [1].

The objective of the article is to design a mobile game application for people with Down syndrome improving the development of cognitive and visual-motor skills through an interaction between the person and the application.

The rest of the document was organized as follows: Section II will define the Scrum methodology, the Troncoso method, and tools to be used, Section III will show the case study, Section IV will show the discussions, Section V will show the results, and finally Section VI will show the conclusion.

II. Methodology

In this article, we will implement the Troncoso method, the Scrum methodology, and some technological tools to develop a mobile application prototype for the learning of people with Down syndrome through interactive games.

A. Troncoso Method

Teaching people with Down syndrome to read is a very difficult activity that requires special pedagogical techniques. This syndrome causes disorders in the cognitive mechanism of attention, alertness, memory, correlation, analysis, and abstract thinking. Also, the visual-motor cognitive ability of these people is not fully developed, since a child with Down syndrome can not distinguish the hemispheres of the body [1].

Table I shows the percentage of people with Down syndrome in Spain according to the relation of activities since there is a high percentage that lacks illiteracy from 15 years old and only a low percentage that is in high school, making it known that there is no support for people to develop educational aspects.

TABLE I. POPULATION OF SPAIN WITH DOWN SYNDROME AGED 16 AND OVER ACCORDING TO AGE AND RELATIONSHIP WITH ACTIVITY (PERCENTAGES)

	Of 15 to 29	Of 39 to 44	Of 45 to 59	Whole population
He does not know to read nor to writes	36,8	60,9	79,9	55,39
Incomplete primary studies	41,9	24,8	13,7	29,34
Primary or equivalent studies	18,1	12,4	3,5	13,4
1st stage secondary education	3,2	1,9	2,9	2,47
Total	100,0	100,0	100,0	100,0

Table II shows the number of people in Peru who have some kind of deficiency that prevents them from developing intellectually, linguistically, visually, or morally, etc.

TABLE II. POPULATION WITH DOWN SYNDROME REGISTERED IN THE NATIONAL REGISTRY OF PERSONS WITH DISABILITIES, BY TYPE OF IMPAIRMENT

Type of deficiency	Quantity
Intellectuals	8 388
Other mental	6 494
Of language	7 798
Of audition	775
Of vision	1734
Visceral	1852
Musculoskeletal	3695
Others	1556

Because of this, it takes into account all the characteristics that are concluded with learning, as it is slow and it is necessary to follow a step-by-step process. Currently, there are educational methods designed for people with Down syndrome, one of them is the Troncoso method.

The Troncoso Method is a reading and writing teaching system designed specifically to work with people with Down syndrome. Also, it aims to ensure that students acquire a reading and writing ability good enough to have an independent life, trying to manage functionally and practically with their social and cultural environment [8].

1) Stages of reading instruction: It is done in the opposite way to the usual. Instead of teaching the child to recognize the fundamental units of words, it begins by helping him/her recognize some simple words through the use of pictures. The objective of the first stage is to make the child capable of recognizing written words based on their shape, without the need to know the individual letters. In the second stage, the objective is to help recognize the syllables by which words are formed so that they can be identified in other written texts [9].

2) Stages of teaching writing: It is divided into three phases, ranging from the most general and simple skills to the most complicated. In the first stage, you have to become familiar with the use of pencil and paper by drawing basic lines and shapes. In the second phase, the student will apply the skills acquired during the previous phase and will have to learn to write letters. Later, he or she must be able to put them together in syllables, words, and sentences. Finally, the final stage of the teaching of writing has to do with acquiring greater speed and fluency in reproducing all types of written texts [9].

3) Stages of teaching visual-motor coordination: The following stages of teaching do not belong to the Troncoso method, since that method only teaches reading and writing, but will be a perfect complement to the development of motor skills of people with Down syndrome. The teaching of visual-motor coordination consists of three different stages, but before the first stage, the same user or with the help of a person will place the mobile device in front, for example, on a table or shelf and must be standing at a distance of five or six feet.

The first stage is a presentation of concepts of the hemispheres, clearly showing the child which of each side is the left and which is the right [1]. After the presentation, the front camera will be activated and the user will be able to look at himself as in a mirror, at that moment an assistant will be shown that will order him to move his arms or legs from the bottom to the top, then he will be asked to move a specific arm, either right or left, if the user succeeds the assistant will congratulate him, but if he fails it will motivate him and will give him another order so he can keep on trying and can achieve the exercise.

B. Scrum Methodology

Scrum is a software development model that helps establish successful and organized step-by-step projects. Agility is what most characterizes the methodology, so it was chosen to implement in the project.

1) Processes: The methodology consists of four processes, and they are the following:

a) Sprint Planning: In this meeting, we will define the functionality in the planned increment and how it is being developed in the increment and output defining the objective that Sprint will have [10].

b) Sprint development work: This process of the methodology will make use of the Sprint Backlog tool, and each team member inspects the work that the rest is doing [11], also, no changes can be made so as not to affect the objective of the Sprint in process, the quality is maintained and if a Sprint is too large can increase the complexity, but also increase the risk.

c) Sprint Review: The review process is done at the end of each Sprint lasts approximately 4 hours, so that the increase can be inspected in detail. The interested parties and the Scrum team are the collaborators of the revision, besides, it is directed by the Product Owner who is in charge of watching over the fulfillment of the established. The result of this process is a Product Backlog with some possible changes in its estimate [10].

d) Retrospective of the Sprint: It is an opportunity for the team to inspect themselves for errors they may have made, and to create a plan for improvements in the next Sprint, with a time frame of approximately 4 hours [10]. And after this process, the cycle starts again after the Sprint.

2) Tools: The methodology has two important tools, and they are:

a) Product Backlog: It is a list of tasks that encompasses the entire project, managed by the Product Owner defining the order and priority. The list has a dynamic character and can change at the end of each Sprint [10].

b) Sprint Backlog: It is a subset of the Product Backlog tasks that the development team chooses for the Sprint, along with the plan to develop it [10].

Fig. 1 represents the order of the processes, how they are developed, and the use of tools.

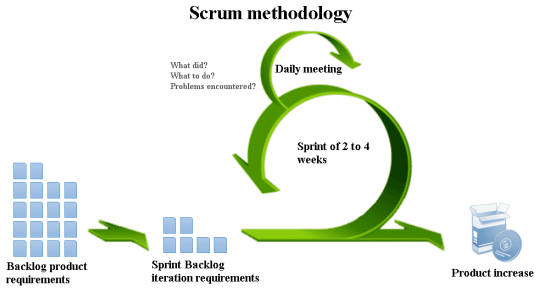


Fig. 1. Scrum Methodology Processes

C. Prototype Tools

For the design of the mobile application, we will use Balsamiq Wireframes, which is a graphic design tool for user interfaces, either for websites and desktop or mobile web applications [12]. The advantage of this tool is its interactivity when making the designs and the accessibility of the software.

D. Technological Tools

The following tools will be used for the development of the mobile application prototype.

1) Android Studio: The environment that will be used to develop the mobile application will be Android Studio, which is the official integrated development environment (IDE) for Android application development, based on IntelliJ IDEA. It is one of the best code editors and has many tools for Intellij [13]. Also, it is characterized, by being a fast emulator, with many functions, tools, with a wide variety of API (application programming interface) that you can use for any development project of a mobile application.

2) Kotlin: To develop the application, we will use the Kotlin language, which is a fairly modern programming language that simplifies the work, increases productivity, and leaves the developer satisfied [14]. It has a more secure, more concise, and optimal code concerning asynchronous programming.

3) SQLite: Regarding the database, the best option would be SQLite, which is a C language library that has implemented a small, fast, autonomous, highly reliable SQL database engine with almost all the functions [15]. It is the most widely used database engine in the world, is integrated into absolutely all cell phones and a large part of computers, and is included in almost all applications that people use.

III. Case Study

In this section of the research, the development of the mobile application prototype will be explained in detail, along with the methodologies mentioned, because it will perform the processes according to the requirements to have the prototype. Next, the planning of the Sprint of each of the modules to be exercised and the estimation of the time spent in each of the development stages will be carried out.

A. Sprint Planning

In this section will be made the User History that are descriptions of the functionalities that will have the software and will be the result of the collaboration between client and team and will be improved during the life of the project, besides, they are of fast forms to administer the requirements without elaborating great documents[16], next, the following format will be placed for the user history As, I want and for; that will be developed in the case of study.

- As a User, I want to enter the application through an account generated to access the application.
- As a User, I want to register to add my data to have an account and enter the application.
- As a User, I want to make the entrance evaluation to know at which level I am.
- As a User, I want to choose the learning method according to my preference and the results of the entrance test to acquire knowledge according to my level.
- As a User, I want to visualize the reports of my test scores to have control of my learning.

1) Time Estimation: In this phase, the duration of each of the Sprints is, observed, as shown in Table III. One of the best-known techniques used to estimate in Scrum is Planning Poker, which uses an arbitrary size measure to indicate the complexity in the size of a user story and whose value only makes sense to the development team in question [17]. The estimation of the project resulted in three months because the user stories were estimated between one or two weeks.

TABLE III. PRODUCT BACKLOG

Interface	Duration
Mobile application for learning development	3 months
Sprint1: Login Interface	1 week
Sprint1:Registration Interface	2 weeks
Sprint2: Input Evaluation Interface	2 weeks
sprint3: Written Learning Interface	2 weeks
Sprint3:Reading Learning Interface	2 weeks
Sprint3:Coordination Learning Interface	2 weeks
Sprint4: Reports interface	2 weeks

2) Product Scope: It can be understood as the characteristics that the product has as a result of a project. The functionalities that it has originated from a series of requirements given by the client or the executing organization, which indicate how the product is wanted. Therefore, if you want to know if the scope of the product was met, it is verified and evaluated that all requirements raised that are included within the resulting product, that is, it is as requested[18].

Fig. 2 shows the estimate of the time it will take for the equipment to have the points of the user history so that the estimates of the scope of the equipment are noted in the least detail.

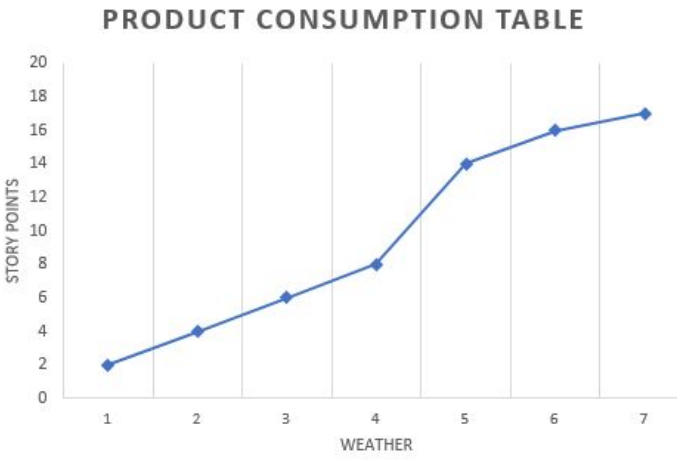


Fig. 2. Product Scope

B. Sprint Development Work

The Sprints developed in this process are:

1) Sprint 1 (Login and Registration): In this Sprint, the prototypes of the user login and registration interfaces were developed. The user will be able to enter his account with an email and a password, as shown in the login interface in Fig. 3. If you do not have an account you can register, as shown in the login interface in Fig. 3, you will have to enter your data such as your full name, age, gender, date of birth, email, and password.

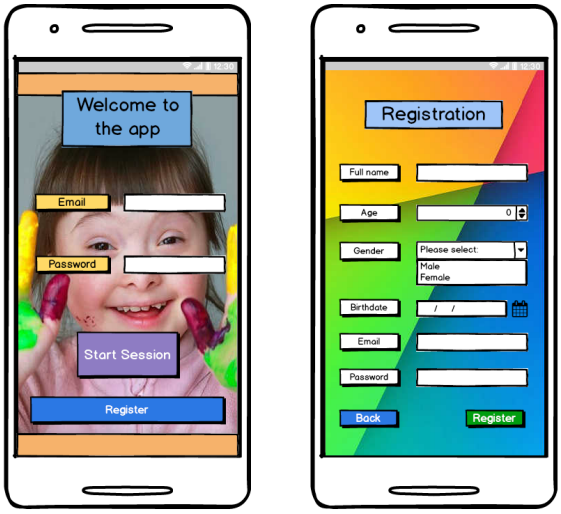


Fig. 3. Login and Registration

2) Sprint 2 (Input evaluation): In this second Sprint, the prototype of the input evaluation interface was developed to determine the user's level of knowledge. As shown in Fig. 4, the interface has three sections, one for writing, one for reading, and one for coordinating movements.

Fig. 4. Input Evaluation

3) Sprint 3 (Learning to write, read, and coordinate movements): In this Sprint was developed the prototype of the learning interfaces for the user, as shown in Fig.5 is the writing interface, there the user will learn to write the words correctly, the exercise consists of completing a word, where you will have three alternatives, and only choose one that is correct, also at the top of the word is a picture of support.

The reading interface is shown in Fig. 5, there the user will learn to read texts, the exercise consists of reading small stories, while reading, the words read will be shaded. If you do not read it correctly, the word will not be shaded, you will

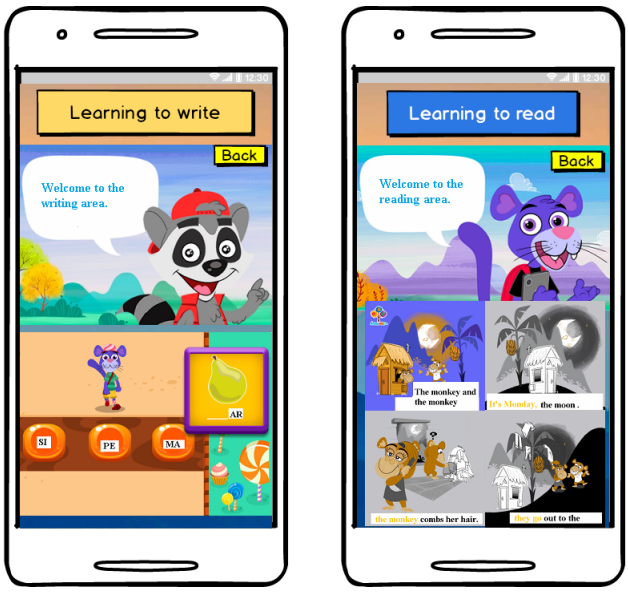


Fig. 5. Teaching Writing and Reading

have to read it again so that everything is shaded and thus complete the exercise.

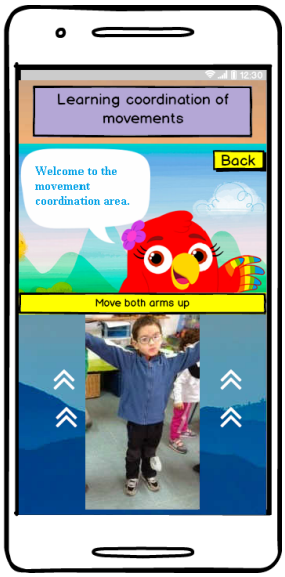


Fig. 6. Teaching Coordination

The interface shown in Fig. 6, has the objective that the user learns to recognize the right and left hemisphere, making movements of his arms or legs. As an example, we show a child who is looking at the front camera of the cell phone and is doing the action of raising both hands, because the application tells him to move to his right and left arm upwards. Another example is shown in Fig. 5 in the lower part where the learning of movement coordination is, the girl receives the indication to raise her right hand and leave her left hand down, but she also raises her right leg, but the application considers that she correctly fulfilled what was indicated.

4) Sprint 4 (Report on learning): In this last Sprint, the prototype of the interface for reporting the user's learning was developed, as shown in Fig. 7, the user will see the progress he has in the three learning, the first bar graph measures the number of correct answers in each learning and it is seen that it stands out in the reading. Also, a pie chart is generated showing a total of the correct answers per month, and it is seen that in November the user had more correct answers than in previous months.

Fig. 7. Report on Learning

C. Sprint Review

When the Sprint is finished, it is inspected by the team, with the review time of two to three hours. One of the team members will evaluate each task developed and decide if any changes are needed. Also, the team members will explain each process in detail and the solutions that were implemented.

D. Retrospective of the Sprint

The team evaluates each process and technique implemented during the development of the Sprint. For example, the use of the Kotlin programming language, the SQLite database manager, or the Balsamiq tool for the development of the prototypes. New ideas or methods can be suggested for the gradual improvement of each Sprint of the project.

IV. Discussion

In this section, the analysis will be carried out and compared with other researches, of which they present similarities or inequality concerning the development of the methodology chosen for the application, taking into account the phases involved in it and in its way of presenting the execution.

A. About Case Study

The prototype was made with Balsamiq because it is a simple and accessible tool so that the functionalities were reviewed and approved by the team. Compared to another

application called Hatle that was made with Android Studio, for children who can not read or write [19], the design was improved so that the user does not have difficulties with the application, increasing tests so you can see your progress.

B. About the Methodology

The Scrum methodology compared to the RUP methodology, has some controversy, has some essential characteristics of agile development processes such as iterative growth or is focused on the architecture, but in turn the rigidity of conventional methods [20].

Among other advantages that the Scrum methodology highlights, is the adaptation to a wide range of software development projects where development teams are small, with short deadlines, and high levels of quality [21].

Some limitations that were found in the research work that could have affected the results, was not choosing a good method that is very useful for people with Down syndrome can learn correctly in the application, and another is that there was not much application research done.

V. Results

Next, we will show the expected results regarding the development of the research within the case study and the methodology, taking into account the application to be developed and the implementation of the control system.

A. About Case Study

As a case study, the objective was to develop a learning tool for people with Down syndrome by improving the development of their cognitive and motor skills based on technologies, determining the effectiveness by comparing the progress of literacy, using measures of letter identification, reading, spelling and handwriting quality.

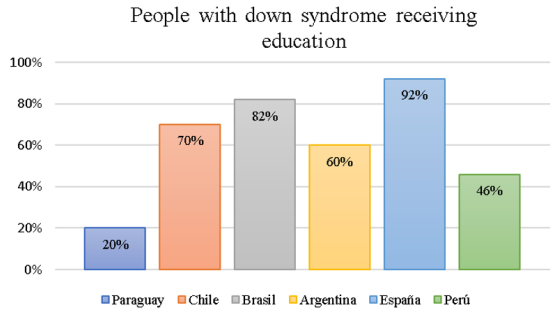


Fig. 8. People with Down Syndrome Receiving Education in Some Countries

Fig. 8 shows that countries like Peru and Paraguay have a low percentage of people with Down syndrome who receive an education. And of this percentage, only a small group of people receive it in specialized schools.

And current studies support the hypothesis that, through appropriate applications of the tablets, special students with disabilities or learning difficulties can enrich their literacy skills [19].

However, the use of this technology, while producing improvements in spatial reasoning and visual attention, does not produce the same effect in selective attention [22]. Furthermore, as we can see in Fig. 9, the data collected by the Inclusive Education Group of the Salesian Polytechnic University in the current research situation of the use of technology as an assistive resource in the education of students with disabilities shows that it is of great help to them.

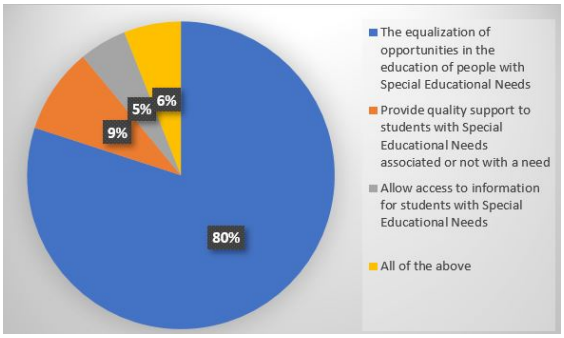


Fig. 9. Benefits of the use of ICT in the Education of Persons with Disabilities

B. About the Methodology

The implementation of the Scrum methodology, allowed to generate a comfortable work environment for the whole team because they worked collaboratively [23]. Besides, it has a structure that divides processes and tools easy to understand and develop.

VI. Conclusions

We developed the prototype of the application that will help people with Down syndrome to develop their cognitive and visual-motor coordination skills through interactive games with the Balsamiq tool that is well known for interface design. And with the implementation of the Scrum methodology and the Troncoso method allowed better teamwork. In this way, with the teachings received through the interactive games, people with Down syndrome will be able to integrate into society, be more independent in their daily activities and thus have a better quality of life. As future work, we want to develop the application implementing artificial intelligence and adding more interactive games in the different learning so that the user does not get bored with the same games and have more fun and optimal learning.

References

[1] T. Rodrigues, N. Valencia, D. Santos, A. Frizera, and T. Bastos, "Development of game-based system for improvement of the left-right recognition ability in children with down syndrome," pp. 627–634, 2019.
[2] P. V. Torres-Carrión, C. S. González-González, P. A. Toledo-Delgado, V. Muñoz-Cruz, R. Gil-Iranzo, N. Reyes-Alonso, and S. Hernández-Morales, "Improving cognitive visual-motor abilities in individuals with down syndrome," *Sensors*, vol. 19, no. 18, p. 3984, 2019.
[3] G. de Graaf, F. Buckley, and B. G. Skotko, "Birth and population prevalence for down syndrome in european countries," 2018.
[4] G. Castro and R. Jhonatan, "Síndrome de down en el perú," *Multimedia Tools and Applications*, p. 4, 2020.
[5] W. F. W. Ahmad, H. N. B. I. Muddin, and A. Shafie, "Number skills mobile application for down syndrome children," pp. 1–6, 2014.

[6] N. S. Abdul Aziz, W. F. W. Ahmad, and N. J. b. Zulkifli, "User experience on numerical application between children with down syndrome and autism," pp. 26–31, 2015.

[7] K. Caro, I. A. Encinas-Monroy, V. L. Amado-Sanchez, O. I. Islas-Cruz, E. A. Ahumada-Solorza, and L. A. Castro, "Using a gesture-based videogame to support eye-hand coordination and pre-literacy skills of children with down syndrome," *Multimedia Tools and Applications*, pp. 1–28, 2020.

[8] N. M. Rubio, "Método troncoso: qué es y cómo se aplica en niños y niñas," urlhttps://psicologiaymente.com/desarrollo/metodo-troncoso, 2016.

[9] R. P. Alejandro, "Troncoso method: what is it for, materials, stages," urlhttps://www.lifeder.com/metodo-troncoso/, 2020.

[10] J. C. A. Becerra and C. E. D. Vanegas, "Propuesta de un método para desarrollar sistemas de información geográfica a partir de la metodología de desarrollo ágil-scrum." *Cuaderno Activa*, vol. 10, pp. 29–41, 2018.

[11] M. S. Méndez Ramos, J. Preciado Garcia, D. P. Castrillón Arboleda *et al.*, "Modelo de transición de metodología rup a scrum en proyectos de desarrollo de software."

[12] Balsamiq, "What tools does balsamiq offer to help remove bad software from the world?" urlhttps://balsamiq.com/company/, 2020.

[13] Developer.Android, "Introduction to android studio," urlhttps://developer.android.com/studio/intro, 2020.

[14] Developer, "How to develop android apps with kotlin?" urlhttps://developer.android.com/kotlin?hl=es, 2019.

[15] SQLite, "What is sqlite?" urlhttps://www.sqlite.org/index.html, 2020.

[16] K. V. Suaza, J. J. T. García, and C. M. Z. Jaramillo, "Mejora de historias de usuario y casos de prueba de metodologías ágiles con base en tdd." *Cuaderno activa*, vol. 7, pp. 41–53, 2015.

[17] J. G. F. Mendoza, F. G. Vera, C. C. O. González, and A. S. Contreras, "Propuesta metodológica para la estimación de proyectos gestionados mediante scrum, con enfoque a la pequeña industria del software (methodological proposal for estimating projects managed in scrum, focusing on small business software industries)," *Pistas Educativas*, vol. 41, no. 134, 2019.

[18] R. B. Miguel, "Product scope vs. project scope," urlhttp://www.liderdeproyecto.com/articulos/, 2019.

[19] V. G. Felix, L. J. Mena, R. Ostos, and G. E. Maestre, "A pilot study of the use of emerging computer technologies to improve the effectiveness of reading and writing therapies in children with down syndrome," *British Journal of Educational Technology*, vol. 48, no. 2, pp. 611–624, 2017.

[20] D. A. V. Paredes, L. C. C. Martínez, R. M. L. Bermúdez, and S. R. P. Mendoza, "Análisis de la metodología rup en el desarrollo de software académico mediante la herramienta django," *RECIMUNDO*, vol. 3, no. 2, pp. 964–979, 2019.

[21] A. Kuz, M. Falco, and R. S. Giandini, "Comprendiendo la aplicabilidad de scrum en el aula: herramientas y ejemplos," *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*, no. 21, pp. e07–e07, 2018.

[22] A. Adrián, "La tecnología, gran aliada de los alumnos con down," urlhttps://www.magisnet.com/2019/04/la-tecnologia-gran-aliada-de-los-alumnos-con-down/, 2019.

[23] N. Tymkiw, J. M. Bournissen, and M. C. Tumino, "Scrum como metodología de enseñanza y aprendizaje de la programación," in *XXII Workshop de Investigadores en Ciencias de la Computación (WICC 2020, El Calafate, Santa Cruz)*, 2020.

Anti-Molestation: An IoT based Device for Women's Self-Security System to Avoid Unlawful Activities

Md. Imtiaz Hanif[1] , Shakil Ahmed[2] , Wahiduzzaman Akanda[3] , Shohag Barman[4]
Department of Computer Science
American International University-Bangladesh, Dhaka-1229, Bangladesh

***Abstract*—Now-a-days, the public, mostly women and children are facing much harassment from the societies. The unlawful activities against ladies and children have been increasing significantly, and regularly we find out about eve-teasing, sexual assault cases, and attempt to molest or even killing after rape in public places or open areas. Also, many cases had gone unwarranted due to short pieces of evidence. In Bangladesh, the current statistics of sexual assaults and various unlawful activities are proliferating. To acknowledge these problems, in this paper, we have designed an IoT-based (Internet of Things) embedded device that is able to communicate with the law enforcement agency by dialing "999" (An Emergency Telephone Number in Bangladesh) on demand. The device contains Arduino Pro-Mini Microcontroller with a GSM (Global System for Mobile communication) module and can send SMS (short message service) with the victim's present area to the law enforcement agency and relatives via GPRS (General Packet Radio Services). The proposed device's form factor is too tiny to carry out easily at anywhere and anytime. The device features the "Plug & Play" functionalities, which means one button to operate the entire device. Also, the device is cost-effective so that people of every level can afford it at a reasonable price.**

***Keywords*—Anti-rape; IoT device; smart-safety device; women safety; wearable device; GSM/GPRS**

I. Introduction

Women reserve the option to be liberated from savagery, badgering, and segregation. Eliminating the obstructions of a hazardous domain can assist women with satisfying their potential as individuals and as supporters of work, communities, and economies. However, the World Health Organization (WHO) evaluates, about 35% of ladies worldwide have confronted either physical or possibly sexual private assistant mercilessness or non-associated sexual brutality throughout their life. Furthermore, more than 15 million young ladies matured 15-19 years have encountered rape [1]. Brutality against ladies and kids does not separate by race, religion, culture, class, or a nation. Recently, in Bangladesh, these types of unlawful activities are booming rapidly. Every morning, when we take the newspaper in hand, we find out various rape cases, killing, and molestation against women. Also, unlawful activities like kidnapping, hijacking, and robbery, etc. are happening every day. According to the statistical report of "Ain o Salish Kendra" (ASK), in 2018 to Aug. 2020 in Bangladesh, more than 3000 women were raped, and 180 women were killed after rape [2]. Among those cases, only the child rape cases were more than 600. Working ladies, female understudies, and children especially confronted this unfortunate incident.

At this moment, we can undoubtedly get to the necessary information in real-time, from (nearly) whichever area we are at. Due to rapid advancements in embedded systems, the IoT-based devices are getting popular day by day because of its flexible interactivity. IoT energizes the correspondence between gadgets. It also makes robotics easy and minimizes humanoid sweats that can save much time. Nevertheless, some security and privacy issues also are there in the IoT system [3]. So in this paper, we have designed an IoT-based device, namely, "Anti-Molestation," to ensure women and child safety in daily life. The device can send SMS with the victim's current location to her friends and family. Also, the device can make the call to "999" (Emergency Number of Bangladesh) as on-demand. We have developed the device in such a way that every level of women or people can afford it at a reasonable price. Moreover, we designed the prototype model as too small, and also it can be formed in a smaller device to use it in a locket, hand-bag, and pocket or as a bracelet, etc. So the contributions of our device are:

- It can send SMS with the current location.
- It can make a call by pressing a single button.
- The device is cost and power-efficient.
- The device overcome burst-transmission phenomenon.

The paper is formulated as six sections. In Section 2, the previous work on women's safety has been discussed. In Section 3, we have discussed our system model and the overall framework. In Section 4, the implementation details of our device has been shown. In Section 5, the overall methodology and results are discussed. Lastly, we have concluded the paper by discussing the overall summaries and the scope of future work.

II. Related Work

By studying literature, we have come to know that many works had done to ensure women's safety by proposing various IoT and Application-based devices [4]. A.Z.M. Tahmidul et al. [5] proposed an application based wearable device. The primary function of this device is to send SMS and the victim's current area to the closest police headquarters and family members. The application interface is designed in such a way that the map indicates a safe location to survive from a criminal attack. This device is not user friendly for rural women. Many of the girls from rural areas are not familiar with mobile applications, or they may even not have a smartphone. However, the form factor of this device is too large to carry it easily. To avoid rape, another paper [6] also proposed a smart mobile application "BONITAA" which also warped with

various features such as SMS and location sending via GSM, health supports, medical supports, counseling, self-defense tips for the rape victims. To acknowledge the problem of rural women, they integrated the "Bangla" language in their application and tried to make it user friendly. However, the problem is that the women who are not familiar with using mobile applications may not relish the facilities of those applications.

The author in [7] proposed a wearable device to ensure women's safety to avoid sexual assault. The method also introduced a mobile application. They designed their device using GSM, GPS (Global Positioning System), and the Wi-Fi (Wireless Fidelity) module integrated with a Microcontroller. The device is also able to make the call and send the location to the pre-recorded numbers or the nearest police stations to avoid unlawful activities. The main problem is that the device needs the always-on internet for web server access. They spent a lot on building such a device with a mobile application. It can not be affordable for every woman in our society, and also the application-based interface is not flexible for all end users. In the paper [8], the author also proposed a women safety device. The device also can send SMS and location to the pre-saved phone numbers. To avoid sexual manipulations, the authors integrated three push-buttons, GSM, GPS, RFID (Radio-Frequency Identification), vibrator, buzzer, and display with a Microcontroller. The main problem is that their prototype model is too large to carry. The device also contains three switches, which are very difficult to navigate through those switches in a panic situation.

U. Rai et al. [9] developed a safety device using Raspberry-Pi with a GPS module and a push button. When a girl presses the push button, the device sends its longitude and latitude via a GPS module. They also developed their location finder application to identify the victim's location. The prototype is simple but large in size. Raspberry-Pi needs the always-on internet to capture the location's coordinates. N. R. Sogi et al. [10] also proposed a Raspberry-Pi based IoT device. They called their device as "SMARISA" smart ring for women's safety. The device also can send the location, and additionally, it contains a camera module. When a woman presses the button, the device is activated, and the camera module captures the incident and upload images in the local-host server and generate sound with the buzzer. They also developed a mobile application to communicate with the victim's trusted one. The problem of this device is that no GSM module was not considered, and costly to develop such devices. Moreover, "Raspberry-Pi", a mini-computer, is also power-hungry. Various sensors, such as pulse rate, motion, and temperature sensor, etc. are used in [13], [14], [18]. Those devices might not work correctly because of integration of various sensors in one module. The sensors may generate wrong readings in any situation that can activate the device. The form factors of those devices are also too large to carry. Another device [11] developed to help the victims from the teaser and molester. The device also sends the SMS and current location to the family members of the victim. Sensor-based devices are not efficient to use because of its fault result generation.

T. Sen et al. [12] developed a women's safety device with Raspberry-Pi. The device contains a monumental architecture with its camera module nerve simulator. The device can send the victim's location via GPS and GSM module. The authors also developed an android application and local server to maximize women's security. Nevertheless, the problem is that the device is too big and can not be carried out easily. V. Sharma et al. [15] also developed a smart shoe to ensure women's safety. The authors used Raspberry-Pi and Arduino-Uno Microcontrollers to implement their system. First of all, a shoe base video capturing devices are not efficient to capture correct frames. Besides, the use of Arduino Microcontroller rather than Raspberry-Pi is inadequate in this case. Also, the shock generation of 400KV can kill a human within some minutes. Another work [16] was proposed for supporting the women in danger. The device is also designed to send SMS and location to relatives of the victim. The device is too huge with its AAA size batteries and the large LCD screen. M. R. Ruman et al. [17] also developed a safety device for women's safety. The device is also able to send the location of the victim and can be rescued early. The device contains a shock generator also. Integration of various extra facilities, the prototype is too large to carry.

To acknowledge these problems, we have also built our safety device to support the women and child. The device is so tiny and can be carried out efficiently in daily life. Moreover, our device is very much cost efficient that people of all levels can afford it at a reasonable price. The device also has the feature of "Plug & Play". So, one operational button to activate the device for all functionalities.

III. Framework and System Design

In this section, we have discussed the overall framework of our "Anti-Molestation" safety device. The framework contains two types of system design.

A. Working Framework

At first, we turn on the device, and when the system is started, it will initialize the SIM800L module. Then the GPRS and GSM module will ready to read the data from the user when we press the button once it obtains the Geo-Coordinate of the current location via GPRS. Then the system sends the HTTP (Hypertext Transfer Protocol) POST request using the "AT" attention command along with the Geo-Coordinate location to the associated application server. Then the application server sends the Geo-Coordinate location to the pre-stored phonebook of the victim's relatives and "999" via SMS gateway. When we long-press the panic button, the system is ready to call the pre-saved emergency phone number, or "999" on demand. Here, Fig. 1 illustrates the overall working framework of our proposed "Anti-Molestation" safety system.

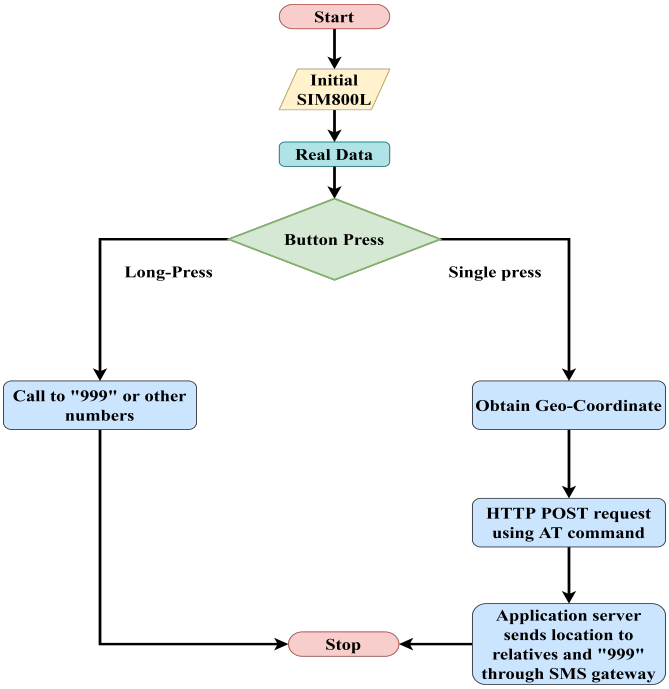


Fig. 1. The Working Framework for our Anti-Molestation Safety Device. Here, "999" is the Bangladeshi Emergency Telephone Number, and the "AT" Denotes the Attention Command to Send Location via Geo-Coordinate.

B. Block Diagram and Components

The square diagram of the framework in Fig. 2 illustrates overall hardware required for developing the gadget.

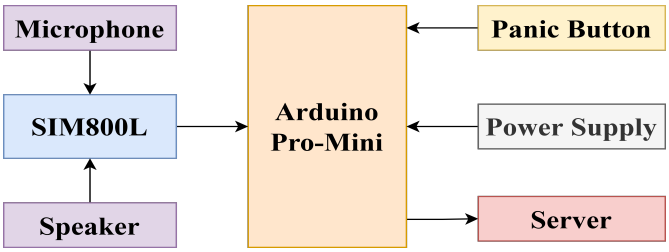
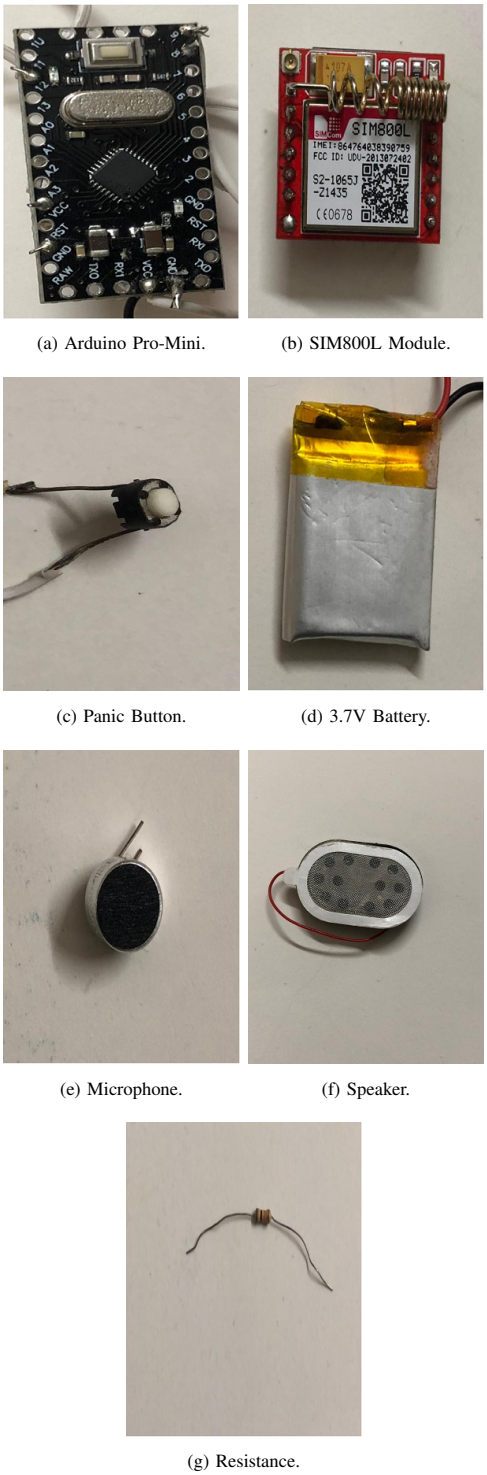


Fig. 2. The Block Diagram of our Device. The Block Diagram shows all the Hardware we have used to Develop our Safety Device.

The main heart of this device is the Arduino Pro-Mini Micro-controller to handle the entire mechanisms. A local-host server is proposed for keeping the truck of every SMS and deliver it to the adjacent police stations. A power supply is needed to power up the device. We have used Li-Poly (Lithium-Polymer) battery as a power supply. The block diagram also contains a panic button to operate the entire device. A SIM800L module has been used for taking advantage of the GPRS and GSM functionalities. We have used a condenser microphone and a speaker to communicate with the pre-saved emergency numbers. All those components are integrated with the Arduino Pro-Mini Microcontroller.



(a) Arduino Pro-Mini. (b) SIM800L Module.

(c) Panic Button. (d) 3.7V Battery.

(e) Microphone. (f) Speaker.

(g) Resistance.

Fig. 3. The figure shows the Necessary Components that we have used for Developing "Anti-Molestation" Safety Device.

1) Microcontroller: A Microcontroller (Fig. 3a) is a minimized incorporated circuit proposed to oversee a specific activity in a rather system. AA normal microcontroller fuses a processor, memory, and I/O (Input-Output) peripherals on a single chip. Here, we have used an Arduino Pro-Mini Microcontroller, which is too small and can handle all the

things we need.

2) SIM800L Module: We have used SIM800L GSM/GPRS module (Fig. 3b) for calling and sending SMS functionalities. The module is attached to the Microcontroller. This module has a small form factor, and a Subscriber Identify Module (SIM) is attached to it.

3) Panic Button: We have used a panic-button (Fig. 3c) to perform "Plug & Play" features. The panic-button regulates two functionalities like single press and long-press.

4) Power Supply: We have used a 3.7V Lithium-Polymer (Li-Poly) rechargeable battery (Fig. 3d) to power the device. We used this because of its small dimensions. Also, a charging port is included to charge the device.

5) Condenser Microphone and Speaker: We have used a condenser microphone (Fig. 3e) and a speaker (Fig. 3f) to transmit the voice through the safety device so that the law enforcement agency can hear the sounds around the prey.

6) Resistance and Some wires: We used a 10KΩ resistance (Fig. 3g) for voltage regularization. Moreover, some wires were used for making connections to the entire peripherals.

IV. Implementation Details

In this section, we have discussed the step by step procedure of the overall hardware implementation of our safety device. Here, Fig. 4 illustrates the circuit diagram of our safety system.

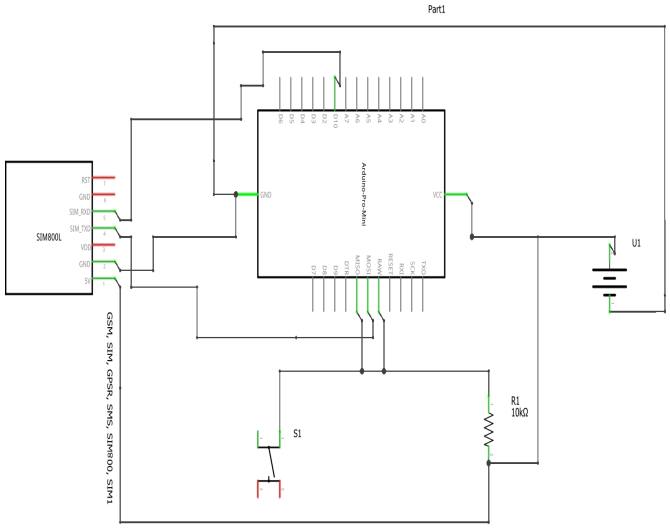


Fig. 4. The Circuit Diagram of the Device. Here, S1, R1, and U1 denotes the Switch, Resistance and Battery, Individually.

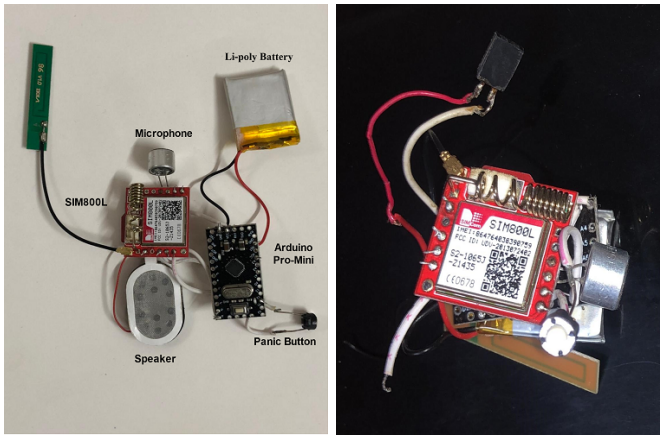
First of all, we need the Arduino Pro-Mini Microcontroller to connect all peripherals because it is the heart and only one thing that can operate all other hardware. Here, in Fig. 4, we can see that SIM800L has several pins and the SIM_RXD pin connects with Arduino's digital Input-Output (I/O) 10 pin for serial communication. The SIM_RXD is the receiver pin. The SIM_TXD is the transmitter pin and connects with the Microcontroller's Master Out Slave In (MOSI) pin for sending data to the peripherals. We have used a rechargeable 3.7V Li-poly battery to power up the whole device.

TABLE I. The cost distribution of used components. We calculate the cost both in US Dollar and Bangladeshi Taka.

Product Name	Price (USD)	Price (BDT)
Arduino Pro-Mini	1.56	2
SIM800L	1.57	134
Panic Button	0.02	2
Li-poly Battery	1.09	93
Condenser Mic.	0.09	8
Speaker	0.02	2
10KΩ Resistor	0.05	5
Total (approx.)	4.40 USD	377 BDT

The positive wire of the battery is connected with the Microcontroller's VCC pin, and the negative wire is connected to the GND pin. The Microcontroller has the Master In Slave Out (MISO), and RAW pins are serially connected with the switch S1 to on-off the device. A 10KΩ resistor is also used for the voltage regulator. The SIM800L module has the MIC pin to connect the condenser microphone and the SPK pin to connect the speaker. To build this device, we have spent some money to buy the necessary components. The total cost of developing the "Anti-Molestation" safety device is approximately 400 BDT or 4.42 USD. Here, Table 1 shows the cost distribution of our necessary components.

The pictorial view of Fig. 5 illustrates the overall view of the Anti-Molestation safety device. Fig. 5a shows the dissection view of the entire device and Fig. 5b shows the developed view of the device.



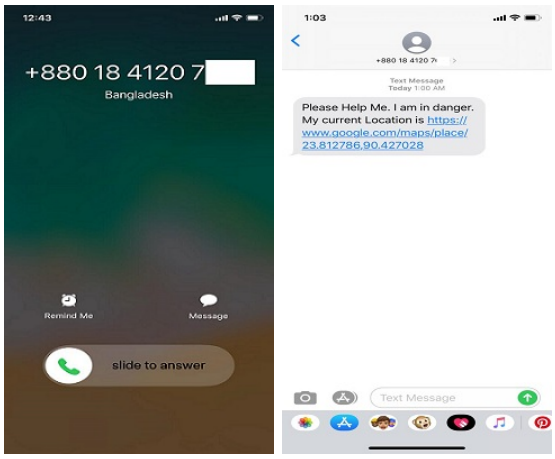
(a) Dissection View. (b) Developed View.

Fig. 5. The Developed Form Factor of the Anti-Molestation Device.

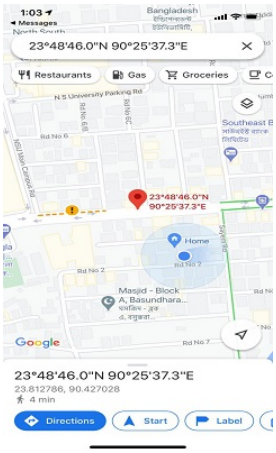
V. Result

In this section, we have discussed the methodology and result of our Anti-Molestation safety device. When a prey

presses the panic button, the SIM800L module call to the "999" law enforcement agency number through the cell phone tower, and by GPRS technology, the device sends SMS with his/her current location and update it in the application server. The authority notifies the nearby police station to rescue the prey. Fig. 5b illustrates the form factor of our device which is too small in size to carry anywhere easily. The device can send location continuously. If the law enforcement agency or police try to make a call to the device number, the device automatically discard the call and again send SMS with victim's current location. Here, Fig. 6a, 6b, and 6c shows the calling mechanism, location sending mechanism, and location traced in map of our safety device accordingly. Moreover, our device is so much power and cost efficient. The device runs a long with a single hour charge. Our safety device has shown a significant result among previous devices with the small form factor.



(a) Making Call. (b) Sending Location.



(c) Location in Map.

Fig. 6. The Qualitative Result of our Safety Device which can make a Call, send SMS with Location.

VI. Conclusion

In this paper, we have proposed and illustrated our device, namely, "Anti-Molestation", an IoT-based safety device. This safety device aims to help women and children from being harassed in any situation. The device can directly inform law enforcement agencies to take legal actions against the culprits. Also, the device can send the victim's current position to the nearest police station to rescue the victim.

The device is so much user-friendly, and people of all levels can use it without any hassle at a low cost. Even though we have built a sophisticated system, we still have some limitations finding the victim's actual location. Location can be distorted up to 100 meters radius from the victim's position where he/she presses the button. We are trying to overcome this issue as soon as possible. Moreover, we will also give our product an aesthetic look so that we can take it to the mass production level quickly.

References

[1] "Violence against women." https://www.who.int/en/news-room/fact-sheets/detail/violence-against-women (accessed Sep. 12, 2020) [online].

[2] "Rape Archives - Ain o Salish Kendra(ASK)." http://www.askbd.org/ask/category/hr-monitoring/violence-against-women-statistics/rape/ (accessed Sep. 12, 2020) [online].

[3] P. Brous, M. Janssen, and P. Herder, "The dual effects of the Internet of Things (IoT): A systematic review of the benefits and risks of IoT adoption by organizations," International Journal of Information Management, vol. 51. Elsevier Ltd, p. 101952, Apr. 01, 2020, DOI: 10.1016/j.ijinfomgt.2019.05.008.

[4] R. Ramachandiran, L. Dhanya, and M. Shalini, "A survey on women safety device using IoT," 2019 IEEE Int. Conf. Syst. Comput. Autom. Networking, ICSCAN, 2019, pp. 1–6, 2019, DOI: 10.1109/ICSCAN.2019.8878817.

[5] A. Z. M. Tahmidul Kabir, A. M. Mizan, and T. Tasneem, "Safety Solution for Women Using Smart Band and CWS App," pp. 566–569, 2020, DOI: 10.1109/ecti-con49241.2020.9158134.

[6] S. R. Mahmud, S. N. Tumpa, A. B. Islam, C. N. Ferdous, N. Paul, and T. T. Anannya, "BONITAA: A smart approach to support the female rape victims," 5th IEEE Reg. 10 Humanit. Technol. Conf. 2017, R10-HTC, 2017, vol. 2018-January, pp. 730–733, 2018, DOI: 10.1109/R10-HTC.2017.8289061.

[7] M. N. Islam et al., "SAFeBanD: A wearable device for the safety of women in Bangladesh," ACM Int. Conf. Proceeding Ser., pp. 76–83, 2018, DOI: 10.1145/3282353.3282363.

[8] S. Priyanka, Shivashankar, K. P. Roshini, S. P. Reddy, and K. Rakesh, "Design and implementation of SALVUS women safety device," 2018 3rd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. RTEICT 2018 - Proc., pp. 2438–2442, 2018, DOI: 10.1109/RTEICT42901.2018.9012442.

[9] U. Rai, K. Miglani, A. Saha, B. Sahoo, and M. Vergin Raja Sarobin, "ReachOut Smart Safety Device," 2018 6th Ed. Int. Conf. Wirel. Networks Embed. Syst. WECON 2018 - Proc., pp. 131–134, 2018, DOI: 10.1109/WECON.2018.8782071.

[10] N. R. Sogi, P. Chatterjee, U. Nethra, and V. Suma, "SMARISA: A Raspberry Pi Based Smart Ring for Women Safety Using IoT," Proc. Int. Conf. Inven. Res. Comput. Appl. ICIRCA, 2018, no. Icirca, pp. 451–454, 2018, DOI: 10.1109/ICIRCA.2018.8597424.

[11] S. K. Punjabi, S. Chaure, U. Ravale, and D. Reddy, "Smart Intelligent System for Women and Child Security," 2018 IEEE 9th Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEMCON, 2018, no. Apr 9600, pp. 451–454, 2019, DOI: 10.1109/IEMCON.2018.8614929.

[12] T. Sen, A. Dutta, S. Singh, and V. N. Kumar, "ProTecht - Implementation of an IoT based 3 -Way Women Safety Device," Proc. 3rd Int. Conf. Electron. Commun. Aerosp. Technol. ICECA, 2019, pp. 1377–1384, 2019, DOI: 10.1109/ICECA.2019.8821913.

[13] K. Thamaraiselvi, S. Rinesh, L. Ramaparvathy, and V. Karthick, "Internet of Things (IOT) based smart band to ensure the security for women," Proc. 2nd Int. Conf. Smart Syst. Inven. Technol. ICSSIT 2019, no. Icssit, pp. 1093–1096, 2019, DOI: 10.1109/ICSSIT46314.2019.8987928.

[14] M. R. Tejonidhi, Aishwarya, K. Chaithra, M. K. Dayana, and H. Nagamma, "IoT Based Smart Security Gadget for Women's Safety," 1st IEEE Int. Conf. Adv. Inf. Technol. ICAIT 2019 - Proc., pp. 348–352, 2019, DOI: 10.1109/ICAIT47043.2019.8987242.

[15] V. Sharma, Y. Tomar, and D. Vydeki, "Smart Shoe for Women Safety," 2019 IEEE 10th Int. Conf. Aware. Sci. Technol. iCAST 2019 - Proc., pp. 1–4, 2019, DOI: 10.1109/ICAwST.2019.8923204.

[16] N. Islam, M. R. Hossain, M. Anisuzzaman, A. J. M. Obaidullah, and S. S. Islam, "Design and Implementation of Women Auspice System by Utilizing GPS and GSM," 2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE, 2019, pp. 1–6, 2019, DOI: 10.1109/ECACE.2019.8679202.

[17] M. R. Ruman, J. K. Badhon and S. Saha, "Safety Assistant And Harassment Prevention For Women," 2019 5th International Conference on Advances in Electrical Engineering (ICAEE), Dhaka, Bangladesh, 2019, pp. 346-350, DOI: 10.1109/ICAEE48663.2019.8975648.

[18] V. Hyndavi, N. S. Nikhita, and S. Rakesh, "Smart Wearable Device for Women Safety Using IoT," no. Icces, pp. 459–463, 2020, DOI: 10.1109/icces48766.2020.9138047.

Non-Linear Control Strategies for Attitude Maneuvers in a CubeSat with Three Reaction Wheels

Brayan Espinoza García[1], Ayrton Martin Yanyachi[2], Pablo Raúl Yanyachi[3]
Electronic Engineering Professional School, Universidad Nacional de San Agustin, Arequipa, Peru[1]
Pedro Paulet Astronomical and Aerospace Institute, Universidad Nacional de San Agustin, Arequipa, Peru[3]
Vrije Universiteit Brussel, Brussels, Belgium[2]

***Abstract*—Development of nanosatellites with CubeSat standard allow students and professionals to get involved into the aerospace technology. In nanosatellites, attitude plays an important role since they can be affected by various disturbances such as gravity gradient and solar radiation. These disturbances generate a torque in the system that must be corrected in order to maintain the CubeSat behavior. In this article, the kinematic and dynamic equations applied to a CubeSat with three reaction wheels are presented. In order to provide a solution to the attitude maneuvering problem, three robust control laws developed by Boskovic, Dando, and Chen are presented and evaluated. Furthermore, these laws are compared with a feedback control law developed by Schaub and modified to use Quaternions. The simulated system was subjected to disturbances caused by a Gravity Gradient Torque and misalignments in the reaction wheels. The effectiveness of each law is determined using the Average of Square of the Commanded Control Torque (ASCCT), the Error Euler Angle Integration (EULERINT), the settlement time, the estimated computational cost (O), and the steady-state error (e_{ss}).**

Keywords*—Attitude control; attitude maneuvers; adaptive control; feedback control; CubeSat; Quaternions; reaction wheels; comparison

I. Introduction

The CubeSat standard, developed in 1999, is intended to reduce development time and costs, as well as increase accessibility to space for students and teachers [1]. As the control of attitude maneuvers is a key factor in the performance of a CubeSat, the aim of this research is to compare and choose the most adequate control law to perform rest-to-rest maneuvers under constant and non-linear disturbances.

Sidi [2], Slotine [3], and Márquez [4] defined the attitude problem as a system of two non-linear matrix equations. Among the attitude system controllers, the one presented by Sidi in [2] is observed. Sidi linearized the attitude equations for small angles and applied PID controllers using Euler Angles and Quaternions. However, this controller was not effective under external disturbances. Control algorithms have also been developed for non-linear systems. These include a feedback control developed by Schaub in [5] where Modified Rodrigues Parameters (MRP) are used to describe the attitude of a satellite, and a Variable Speed Control Moment Gyro (VSCMG) is used as an actuator. Although Schaub's control algorithm had a quick response to constant errors of the inertia tensor, it did not guarantee a quick response to external disturbances. Adaptive control laws have also been developed following the guidelines described by Slotine in [3]. Also, Dando [6] proposed an adaptive control law based on sliding surfaces that calculated the error between the real and the calculated inertia tensor instead of estimating the tensor of inertia itself as proposed by Ahmed in [7]. Nevertheless, this method was only valid for a constant error. Scarritt in [8] estimated a gain applied to the modeled inertia tensor and a rotation associated with the misalignment of the reaction wheel obtaining a robust control algorithm with a high computational cost (O).

All the controllers mentioned above neglected the input control constraints. However, Boskovic [9] focused on evaluating stability under saturation inputs obtaining a robust control algorithm based on the variable structure control. Boskovic's controller was independent of the inertia tensor and has a lower computational cost (O) than the average robust algorithms but its disturbance rejection property cannot be observed. Similarly, Chen [10] developed a robust controller that considered input constraints based on the fast non-singular terminal sliding mode surface (FNTSMS). This controller needed an inertia a priori information but was capable to reject non-linear disturbances and to keep the tracking error around zero. However, only non-asymptotic stability was guaranteed in a finite time.

This article is divided into the following sections: Section II presents the dynamic and kinematic equations for a rigid body. In Section III, the equations that describe the attitude for a CubeSat with three reaction wheels are introduced. Sections IV and V present the equations necessary to simulate the gravity gradient torque and the internal disturbances torques produced by reaction wheels misalignment. Section VI presents a bibliographic review of adaptive control laws developed by Dando, Boskovik, and Chen. Also, a modification of the control law developed by Schaub to use three reaction wheels instead of VSCMG and Quaternions instead of MRP is presented. In Section VII, simulations subject to the aforementioned perturbations and to a miscellaneous torque were performed. In order to measure the effectiveness of the control algorithm, a comparison was made applying five criteria. The first criterion of comparison is the steady-state error (e_{ss}), the second is the Error Euler Angle Integration (EULERINT) used by Sidi in [2], the third is the Average of Square of the Commanded Control Torque (ASCCT) presented by Xiao in [12], the fourth

is the settlement time, and the fifth is the estimated computational cost (O). Finally, the results, conclusions, and future work are shown in Sections VIII, IX, and X, respectively.

II. Attitude Equations

Two equations are required to describe the attitude of a rigid body. First, the kinematic equation that describes the motion of the satellite itself without taking into account the action of forces. Second, the dynamic equation that takes into account the torques acting on the body.

A. Kinematics Equation

The kinematics can be described using Euler Angles, Rodrigues Parameters, Modified Rodrigues Parameters, Quaternions, and other parameters derived from these. In this article, Quaternions are used for the description of attitude [13] [14].

The kinematic equation for rotations between 0^o to 360^o is defined in (1):

$$\dot{q} = \frac{1}{2}\Xi(q)\omega \tag{1}$$

where '$\Xi(q)$' is defined in (2).

$$\Xi(q) \equiv \begin{bmatrix} q_4 & -q_3 & q_2 \\ q_3 & q_4 & -q_1 \\ -q_2 & q_1 & q_4 \\ -q_1 & -q_2 & -q_3 \end{bmatrix} \tag{2}$$

'q' is a Quaternion that describes the rotation of the inertial frame to the body frame and has the following form:

$$q = \begin{bmatrix} q_1 & q_2 & q_3 & q_4 \end{bmatrix}^T \tag{3}$$

A Quaternion 'q' has a scalar part denoted by 'q_4' and a vector part denoted by '$q_{13} = [q_1, q_2, q_3]^T$'. Also, '$q$' must have a unit norm to describe a pure rotation, as seen in (4).

$$||q|| = q_1^2 + q_2^2 + q_3^2 + q_4^2 = 1 \tag{4}$$

B. Dynamics Equation

The dynamics of the CubeSat, neglecting the effect of the reaction wheels and modeled as a rigid body, is described by (5) [17].

$$\dot{\omega} = \tilde{J}_B^{-1}\left[L - \omega \times \tilde{J}_B\omega\right] \tag{5}$$

Where:

- ω: Angular velocity of the satellite relative to the inertial frame.
- $\tilde{J}_B$: Body inertia tensor.
- L: External torque applied to the center of mass expressed in the body frame.

All parameters in (5) are expressed in the body frame according to [2], [4] and [16].

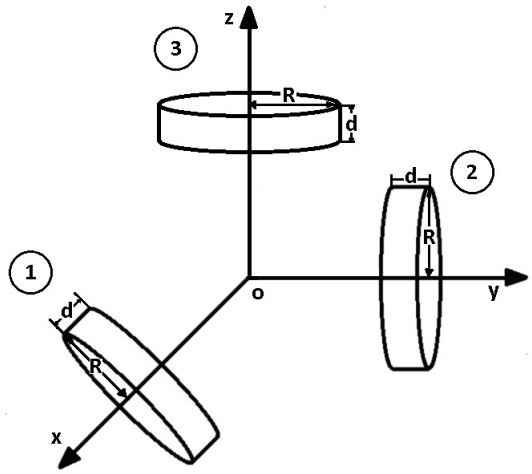


Fig. 1. Distribution of the reaction wheels described in the body frame which origin coincides with the center of mass and geometric. The loads are considered to be fully balanced around the 'o' origin. Adapted from [2].

III. Three reaction wheels CubeSat Model

To consider the effects that reaction wheels have on the system, a CubeSat is simulated with three reaction wheels with the distribution shown in Fig. 1.

'$\tilde{J}_B$' is defined as the inertia tensor without contributions from the reaction wheels and the general inertia tensor 'J_B' is defined in (6) according to [4].

$$J_B \equiv \tilde{J}_B + \sum_{l=1}^{n} J_l^{\perp}(I_3 - W_l W_l)^T \qquad l = 1, 2, 3 \tag{6}$$

Where:

- W_l: Rotation axes.
- $J_l^{\perp}$: Reaction wheel inertia perpendicular to the axis of rotation.
- $J_l^{\parallel}$: Reaction wheel inertia parallel to the axis of rotation.
- I_3: 3x3 Diagonal identity matrix.

Decomposing (6) and taking into account the distribution of the reaction wheels in Fig. 1, the general inertia tensor is defined by (7).

$$J_B = \begin{bmatrix} \tilde{J}_{xx} + 2J^{\perp} & -\tilde{J}_{xy} & -\tilde{J}_{xz} \\ -\tilde{J}_{yx} & \tilde{J}_{yy} + 2J^{\perp} & -\tilde{J}_{yz} \\ -\tilde{J}_{zx} & -\tilde{J}_{zy} & \tilde{J}_{zz} + 2J^{\perp} \end{bmatrix} \tag{7}$$

In the case of reaction wheels with radius 'R_{rw}' , thickness 'd_{rw}' and mass 'm_{rw}', the parallel and perpendicular inertias are defined by (8) and (9), respectively.

$$J^{\parallel} = \frac{1}{2}m_{rw}R_{rw}^2 \tag{8}$$

$$J^{\perp} = \frac{1}{4}m_{rw}R_{rw}^2 + \frac{1}{12}m_{rw}d_{rw}^2 \tag{9}$$

The dynamics of the satellite is defined in (10) according to [2] and [4].

$$\dot{\omega} = J_B^{-1}\left[L - L_{rw} - \omega \times (J_B\omega + H_B^w)\right] \tag{10}$$

Where:

- L: External torque applied to the center of mass expressed in the body frame.
- L_{rw}: Torque delivered by reaction wheels.
- H_B^w: Angular momentum delivered by the reaction wheels.

Equation (6) excludes the inertia of the reaction wheels parallel to the rotation axis since this is considered as a body that rotates freely when the reaction wheel is turned off. However, the parallel inertias are taken into account when the general angular momentum of the system is evaluating (11).

The angular momentum delivered by the reaction wheels is defined by (11):

$$H_B^w = \sum_{l=1}^{n} J_l^{\parallel}(W_l\omega + \Omega_l)W_l \qquad l = 1, 2, 3 \tag{11}$$

where Ω_l represents the angular speed of the reaction wheels.

Particularly, (11) is expressed as follows:

$$H_B^w = J_{rw}\left[\omega + \Omega\right] \tag{12}$$

where J_{rw} is defined by (13).

$$J_{rw} = \begin{bmatrix} J^{\parallel} & 0 & 0 \\ 0 & J^{\parallel} & 0 \\ 0 & 0 & J^{\parallel} \end{bmatrix} \tag{13}$$

The torque produced by the reaction wheels can be calculated with (14).

$$L_{rw} = J_{rw}(\dot{\omega} + \dot{\Omega}) \tag{14}$$

With (12) and (13), the dynamics described in (10) can be rewritten as (15).

$$\dot{\omega} = J_B^{-1}\left[L - L_{rw} - \omega \times (J_B\omega + J_{rw}\left[\omega + \Omega\right])\right] \tag{15}$$

Finally, the attitude for a CubeSat with three reaction wheels is fully modeled with (1) and (15), the acceleration of the reaction wheels can also be calculated with (16).

$$\dot{\Omega} = J_{rw}^{-1}L_{rw} - \dot{\omega} \tag{16}$$

IV. External disturbances

A. Gravity Gradient Torque

The external torques that affect a CubeSat are various. However, for low orbit satellites, the major disturbance that a CubeSat is exposed to is the gravity gradient torque 'L_{gg}' which is defined in [13] and [19] as (17).

$$L = L_{gg} = 3\omega_o^2 c_3 \times J_B c_3 \tag{17}$$

And 'ω_o'and 'c_3' is defined in (18) and (19), respectively.

$$\omega_o^2 = \frac{\mu}{r_c^3} \tag{18}$$

$$c_3 = C(q)\begin{bmatrix}0\\0\\1\end{bmatrix} = \begin{bmatrix}2(q_1q_3 - q_2q_4)\\2(q_2q_3 + q_1q_4)\\1 - 2(q_1^2 - q_2^2)\end{bmatrix} \tag{19}$$

Where:

- μ: Earth's gravitational coefficient ($\mu = 3.986 \times 10^{14}$).
- r_c^3: Distance from the center of earth.
- $C(q)$:Attitude Matrix.
- c_3: Nadir-pointing unit vector.

V. Reaction wheels disturbances

Among the most common disturbances, the friction presented in the motors of the reaction wheels [11] and the misalignments [12] are considered.

A. Reaction Wheel Misalignments

The reaction wheel configuration presented in Fig. 1 is an ideal configuration. In practice, the torque produced by the misalignments is modeled as shown in Fig. 2.

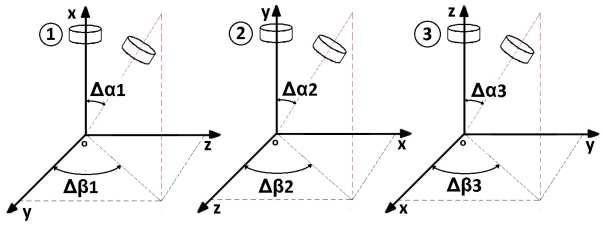


Fig. 2. Misalignments in the reaction wheels. '$\Delta\alpha_1$', '$\Delta\alpha_2$', '$\Delta\alpha_3$' are small angles and '$\Delta\beta_1$', '$\Delta\beta_2$', '$\Delta\beta_3$' are in the range of '$[-pi, pi]$'. Adapted from [12].

The equation (20) describes the total torque produced by the reaction wheels when the misalignment angles are small.

$$L_{rw}^* = L_{rw} + \Delta D L_{rw} \tag{20}$$

Where:

$$\Delta D = \begin{bmatrix} 0 & \Delta\alpha_2 sin(\Delta\beta_2) & \Delta\alpha_3 cos(\Delta\beta_3) \\ \Delta\alpha_1 cos(\Delta\beta_1) & 0 & \Delta\alpha_3 sin(\Delta\beta_3) \\ \Delta\alpha_1 sin(\Delta\beta_1) & \Delta\alpha_2 cos(\Delta\beta_2) & 0 \end{bmatrix}$$

The angles '$\Delta\alpha_1$', '$\Delta\alpha_2$', '$\Delta\alpha_3$', '$\Delta\beta_1$', '$\Delta\beta_2$', '$\Delta\beta_3$' are defined in Fig. 2.

VI. Control Laws

In this article, three control laws for attitude maneuvers are compared. In order to maintain consistency throughout the article, the nomenclatures of each author were changed to those defined in [13] by Markley & Crassidis as shown in Appendix A (Table VI). The 'u' control signal becomes the torque delivered by reaction wheels 'L_{rw}' from (10).

Before starting to describe the control laws, first define the Quaternion error 'δq' between the CubeSat's attitude 'q'and the desired attitude 'q_d' according to Scarritt [8].

$$\delta q = \begin{bmatrix}\delta q_{13}\\ \delta q_4\end{bmatrix} \tag{21}$$

Where:

$$\delta q_{13} = \Xi^T(q_d)q \tag{22}$$
$$\delta q_4 = (q_d)^T q \tag{23}$$

And the angular velocity error is defined as '$\delta\omega = \omega - \omega_d$'. Where '$\omega$' is the angular velocity of the CubeSat relative to the inertial frame expressed in the body frame and 'ω_d' is the desired angular velocity expressed in the frame of the body.

In most cases, only the desired angular velocity expressed in the inertial frame 'ω_{dR}' is available. Then, the rotation matrix '$C(\delta q)$' is needed to express it in the body frame [6].

$$\delta\omega = \omega - C(\delta q)\omega_{dR} \tag{24}$$

The skew-symmetric matrix is also defined in (25) [17].

$$[\omega\times] = \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix} \tag{25}$$

A. Quaternion Feedback Controller

The tracking problem is defined by the error variables '$\delta\omega$' and '$\delta\beta$' defined in (26).

$$\delta\beta = q_{13} - q_{d13} \tag{26}$$

The control law is defined in (27).

$$u = P\delta\omega + K\delta\beta - [\omega\times](J_B\omega - J_{rw}(\omega+\Omega)) - J_B(\dot{\omega}_d - [\omega\times]\omega_d) \tag{27}$$

Where 'K' and 'P' are positive definite matrices. Lyapunov's candidate function is defined by Schaub in [5] as (28).

$$V = \frac{1}{2}\delta\omega^T J_B\delta\omega + K\delta\beta^T\delta\beta \tag{28}$$

The dynamics of the closed-loop model is defined by (29).

$$J_B\frac{d^B}{dt}\delta\omega + K\beta + P\delta\omega = 0 \tag{29}$$

Where $\frac{d^B}{dt}$ is the derivative with respect to the body frame. By deriving (28) and replacing it with (29), (30) is described as follows:

$$\dot{V} = \delta\omega^T[u - [\omega\times]J_B\omega - [\omega\times]J_{rw}(\omega+\Omega) - J_B\dot{\omega_d} + J[\omega\times]\omega_d + K\delta\beta] \tag{30}$$

1) Stability Proof: By replacing (27) in (30), (31) is obtained..

$$\dot{V} = -\delta\omega^T[P]\delta\omega \tag{31}$$

It can be seen that (31) is a positive semi-defined function. Thus, the system is stable. However, since (28) is dependent on two variables, asymptotically stability can not be ensured.

Nevertheless, using the Mukherjee and Chen theorem [18] we can show that the closed-loop system is asymptotically stable. However, it is important to mention that the value of the calculated inertia tensor will never be equal to the real inertia tensor. Hence, the equation (27) is rewritten as (32).

$$u = P\delta\omega + K\delta\beta - [\omega\times](J_B^*\omega - J_{rw}^*(\omega+\Omega)) - J_B^*(\dot{\omega}_d - [\omega\times]\omega_d) \tag{32}$$

Where 'J_B^*' and 'J_{rw}^*' are the calculated inertia tensor and the matrix of parallel inertias calculated for the reaction wheels, respectively.

B. Boskovic Robust Controller

Boskovic's work [9] is based in the variable structure approach and his control technique does not require previous knowledge of the inertia tensor. In addition, Boskovic designed an adaptive gain that allows to compensate the disturbances torques and to ensure that attitude and angular velocity errors will tend to zero.

The control law proposed by Boskovic is given by (33).

$$-u_i(t) = -u_{max}\frac{s_i(t)}{|s_i(t)| + k^2(t)\delta_k} \qquad i = 1,2,3 \tag{33}$$

Where 'δ_k' is a positive constant, 'K(t)' is the adaptive gain and 'u_{max}' is the torque limit for all control torques. The Boskovic Sliding Vector '$s(t)$' is defined by (34).

$$s(t) = \delta\omega(t) + k^2\delta q_{13}(t) \tag{34}$$

Boskovic defined, in [9], the adjustment law for the time-varying control gain as (35).

$$\dot{k}(t) = \frac{\gamma k(t)}{1 + 4\gamma[1-\delta q_4(t)]}\left\{u_{max}\sum_{i=1}^{3}\left[\frac{\delta\omega_i\delta q_i(t)}{|s_i(t)| + k^2(t)\delta_k} - \frac{|\delta\omega_i(t)|(1+\delta_k)}{|\delta\omega_i(t)| + k^2(t)(1+\delta_k)}\right] - \delta\omega^T\delta q_{13} - k^2\delta q_{13}^T\delta q_{13}\right\} \tag{35}$$

where 'γ' is a positive scalar and is called the convergence rate.

C. Dando Adaptive Controller

The adaptive control law proposed by Dando [6] is intended to estimate the error of the inertia tensor instead of the inertia tensor itself.

The closed-loop error dynamics are first defined in (36).

$$J_B\dot{s(t)} = -[\omega(t)\times]J_B\omega(t) - J_B\alpha_r(t) + u(t) \tag{36}$$

The Dando Sliding Vector is defined in (37).

$$s(t) = \omega(t) - \omega_r(t) \tag{37}$$

And the other parameters are defined as:

$$\alpha_r(t) = C(\delta q)\dot{\omega}_{dR} - [\delta\omega\times]C(\delta q)\dot{\omega}_{dR} - \lambda\delta\dot{q}_{13} \tag{38}$$
$$\omega_r = C(\delta q)\omega_{dR} - \lambda sgn[\delta q_4]\delta q_{13} \tag{39}$$

In this control, a priori knowledge of the inertia tensor with a certain level of uncertainty is assumed as defined in (40).

$$J_B = J_B^* + \tilde{J}_E \tag{40}$$

Where 'J_B' is the real satellite inertia tensor and 'J_B^*' is the calculated inertia tensor and '$\tilde{J}_E$' is the inertia uncertainty.

The control law is given by (41) and (42).

$$-u = -K_D s(t) + J_B^* \alpha_r(t) + [\omega(t)\times] J_B^* \omega(t) + \tilde{u}(t) \quad (41)$$

$$\tilde{u}(t) = -\tilde{J}_E \alpha_r(t) - [\omega_r(t)\times]\tilde{J}_E \omega(t) \quad (42)$$

Equation (42) is a torque related to the uncertainty inertia tensor. In order to develop an adaptive control, Dando parametrized '$\tilde{u}(t)$' in terms of '$\tilde{\theta}$' defined by (43).

$$\tilde{\theta}(t) = [\tilde{J}_{E_{xx}}, \tilde{J}_{E_{yy}}, \tilde{J}_{E_{zz}}, \tilde{J}_{E_{yz}}, \tilde{J}_{E_{xz}}, \tilde{J}_{E_{xy}}]^T \quad (43)$$

Dando in [9], and Ahmed in [7] introduced the 'L_{op}' operator defined by (44).

$$L_{op}(a) = \begin{bmatrix} a_1 & 0 & 0 & 0 & a_3 & a_2 \\ 0 & a_2 & 0 & a_3 & 0 & a_1 \\ 0 & 0 & a_3 & a_2 & a_1 & 0 \end{bmatrix} \quad (44)$$

Applying (44) and (43) to parameterize (42) the following equation is obtained:

$$\tilde{u}(t) = \Phi^T \tilde{\theta}(t) \quad (45)$$

Where 'Φ' is defined as (46).

$$\Phi(\omega, \omega_r, \alpha_r) = -(L_{op}(\alpha_r) + [\omega_r\times] L_{op}(\omega))^T \quad (46)$$

And the adaptive law to estimate '$\dot{\tilde{\theta}}$' is given by (47).

$$\dot{\tilde{\theta}}(t) = -\Gamma \Phi s(t) \quad (47)$$

D. Chen Robust Controller

The robust control law proposed by Chen [10] is based on the Fast Non-singular Terminal Sliding Mode Surface (FNTSMS) method and adaptive control methods to compensate the inertia tensor uncertainties under constraints in the reaction wheels.

The Chen Sliding Vector is defined in (48).

$$S = [S_1, S_2, S_3]^T = \delta\omega + \alpha_1 \delta q_{13} + \alpha_2 \beta(\delta q_{13}) \quad (48)$$

In (48), 'α_1', 'α_2' are positive constants and the 'β' operator is defined in (49) and (50).

$$\beta(\delta q_i) = \begin{cases} sig^{\gamma}(\delta q_i) & |\delta q_i| > \eta \\ r_1(\delta q_i) + r_2 sgn(\delta q_i)\delta q_i^2 & |\delta q_i| \le \eta \end{cases} \quad (49)$$

$$sig^{\gamma}(\delta q_i) = sgn(\delta q_i)|\delta q_i|^{\gamma} \qquad i = 1, 2, 3 \quad (50)$$

Where '$r_1 = (2-\gamma)\eta^{(\gamma-1)}$', '$r_2 = (\gamma-1)\eta^{(\gamma-2)}$', '$\gamma > 0$', '$\eta < 1$' and '$sgn(.)$' is the sign function.

The system to compensate the actuator saturation is defined in (51).

$$\dot{\zeta} = \begin{cases} 0 & ||\zeta||_2 \le \zeta_0 \\ -k_1\zeta - k_2 sig^{\gamma_1}(\zeta) - \frac{||S^T \Delta u||_1 + 0.5\Delta u^T \Delta u}{||\zeta||_2^2}\zeta + \Delta u & ||\zeta||_2 > \zeta_0 \end{cases} \quad (51)$$

Where '$\Delta u = sat(u) - u$', 'k_1', 'k_2' and 'ζ_0' are positive constants and '$sig^{\gamma_1}(\zeta) = [sig^{\gamma_1}(\zeta_1), sig^{\gamma_1}(\zeta_2), sig^{\gamma_1}(\zeta_3)]^T$'.

The Chen Control Law is given by (52).

$$-u = -F - \alpha_1 J_B^* \dot{\delta q_{13}} - \alpha_2 J_B^* \dot{\beta}(\delta q_{13}) - k_3\zeta - \frac{1}{2}S + u_r + u_n + u_a \quad (52)$$

Where 'u_a' and 'F' are defined in (53) and (56) respectively.

$$u_a = -\hat{u}\frac{S}{||S||_2 + \epsilon}, \epsilon = \frac{k_0}{1+\hat{u}} \quad (53)$$

$$\hat{u} = \hat{c}_0 + \hat{c}_1||\delta\omega||_2 + \hat{c}_2||\delta\omega||_2^2 \quad (54)$$

$$\dot{\hat{c}}_n = p_n(||S||_2||\delta\omega||_2^n - \chi_n \hat{c}_n), n = 1, 2, 3 \quad (55)$$

$$F = -[\omega\times] J_B^* \omega + J_B^*([\delta\omega\times]C(\delta q)\omega_{dR} - C(\delta q)\dot{\omega}_{dR}) \quad (56)$$

Equations for 'u_r' and 'u_n' are defined in (57) and (58) respectively.

$$u_r = -\tau_1 S - \tau_2 sig^{\rho}(S) \quad (57)$$

$$u_n = -k_4 sig^{\gamma_1}(S) \quad (58)$$

Where '$\rho > 0$', '$\gamma_1 < 1$', 'τ_1', 'τ_2', 'k_3', 'k_4' are positive constants, and '$k_1 - \frac{1}{3}k_3^2 - \frac{1}{2} > 0$'.

Chen defines the saturation function as (60).

$$sat(u) = [sat(u_1), sat(u_2), sat(u_3)]^T \quad (59)$$

$$sat(u_i) = \begin{cases} (U_{max} - a) + atanh(\frac{u_i - U_{max} + a}{a}), & u_i \ge U_{max} - a \\ u_i, & a - U_{max} < u_i < U_{max} - a \\ (a - U_{max}) + atanh(\frac{u_i + U_{max} - a}{a}), & u_i \le a - U_{max} \end{cases} \quad (60)$$

Where 'U_{max}' is the maximum control torque delivered by the reaction wheels, 'a' is positive constant, and 'i=1,2,3'.

VII. Numerical Simulations

Numerical simulations were carried out to perform regulation maneuvers for long angles. The block diagram shown in Fig. 3 describes how simulations were performed and the parameters taken from [15]. Parameters, initial conditions, and desired attitude can be seen in Tables I, II, and III, respectively.

The second-order Simpson's rule, shown in Appendix B, was used as the numerical integration method and the fourth-order Runge-Kutta algorithm was used to approximate the solution in differential equations. Simulation time was 200 seconds with a total of 100,000 iterations with a step of 0.2 milliseconds.

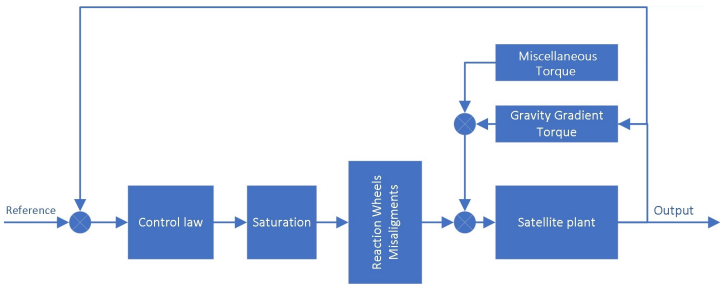


Fig. 3. Block diagram for a CubeSat subject to gravity gradient torque, and misalignment. The control law block may have adaptation algorithms depending on the simulated control law.

TABLE I. PARAMETERS OF SATELLITE AND REACTION WHEELS

Satellite Parameters	
Parameter	**Values**
CubeSat Inertia Tensor	$\tilde{J}_B = \begin{bmatrix} 8.46 & 1.1 & 1.5 \\ 1.1 & 8.46 & 1.6 \\ 1.5 & 1.6 & 8.46 \end{bmatrix} \times 10^{-3}$ (Nm)
Desired trajectory	q_d=$[0.2393, 0.1893, 0.0381, 0.9515]^T$ ω_d=$[0, 0, 0]^T$ (deg/s)
Initial conditions	$q = [0, 0, 0, 1]^T$ $\omega = [0, 0, 30]^T$ (deg/s) $\omega_{rw} = [0, 0, 30]^T$ (deg/s)
Reaction wheels parameters	
Parameter	**Values**
Radius	$R_{rw} = 4.3$ cm
Mass	$m_{rw} = 25$ g
Width	$d_{rw} = 1.5$ cm
Maximum torque	1.343×10^{-2} (Nm)

TABLE II. PARAMETERS OF DISTURBANCES

Gravity Gradient Torque	
Parameter	**Values**
r_c	6471×10^3 m
Reaction wheel misalignments	
Parameter	**Values**
$\Delta\alpha_1$, $\Delta\alpha_2$, $\Delta\alpha_3$	3, -4, 5 deg
$\Delta\beta_1$, $\Delta\beta_2$, $\Delta\beta_3$	10, -50, 70 deg

TABLE III. CONTROL PARAMETERS

Quaternion Feedback Controller	
Parameter	**Values**
Calculated Inertia Tensor [a]:	$J_B^* = I_3$ (Nm)
K[a], P[a] :	I_3, I_3
Calculated Reaction wheels parameters	$R_{rw} = 4cm$ $d_{rw} = 1cm$ $m_{rw} = 20g$
Boskovic Robust Controller	
Parameter	**Values**
δ_k, γ	0.01, 0.001
U_{max}	1.343×10^{-2} Nm
Initial condition	$K_0 = 1$
Dando Adaptive Controller	
Parameter	**Values**
Calculated Inertia Tensor [a]	$J_B^* = I_3$ (Nm)
λ, γ, Kd	1, 0.001, 1
Initial condition	$\theta_0 = [1, 1, 1, 0, 0, 0]^T$
Chen Robust Controller	
Parameter	**Values**
Calculated Inertia Tensor [a]	$J_B^* = I_3$ (Nm)
α_1, α_2, η	1, 0.5, 0.0001
γ, γ_1, ρ	0.6, 0.7, 0.7
k_1, k_2, k_3, k_4	2, 1, 0.3, 1
τ_1, τ_2, k_0	10, 10, 0.0005
p_0, p_1, p_2 χ_1, χ_2, χ_3	0.1, 0.1, 0.1 0.001, 0.001, 0.001
ζ_0, a	0.0001, 0.5

[a] I_3: Diagonal identity matrix.

A. Attitude Regulation Maneuvers and Torque Magnitude Constraints

The rest-to-rest attitude maneuver for a non-spinning CubeSat is simulated with all control laws subject to saturation with a maximum torque of 13.45×10^{-3} Nm emulating the physical limitations of the reaction wheels. Euler angles, angular rate error, and control torque are shown in Appendix C (Fig. 9, 10, 11, and 12), respectively.

B. Disturbances Torque Rejection

In order to evaluate the Disturbance Torques Rejection, a miscellaneous disturbance torque was added as follows:

$$T_{mis} = 0.7[sin(t), 2cos(2t), 3sin(3t)]^T \times 10^{-3} Nm \quad (61)$$

The results obtained were shown in Euler Angles since the attitude Quaternion does not have any physical sense. The conversion of Quaternions to Euler Angles was carried out using the asymmetric XYZ sequence. The results obtained are shown in Figures 4, 5, 6, and 7.

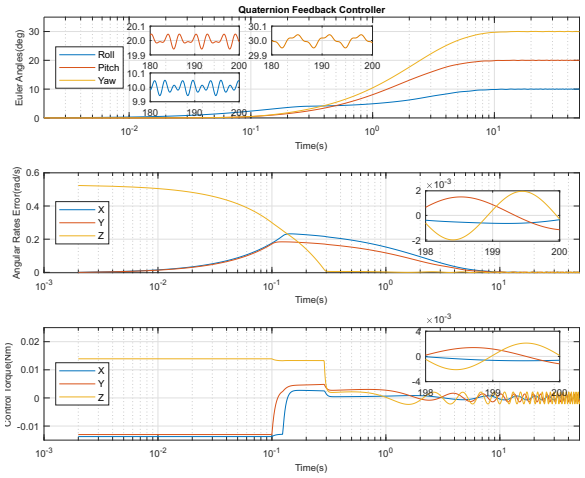


Fig. 4. Euler Angles and Angular Rate Errors produced by the miscellaneous disturbance torque in the Controller defined in (32) with a settling time (ts) of 6.81, $\pm$ 0.04 degrees in the Euler Angles oscillations, $\pm$ 1.9 $\times 10^{-3}$ rad/s oscillations in angular rates, and $\pm 2.1 \times 10^{-3}$ Nm in Control Torque.

C. Evaluation of Performance Between Laws of Control

To compare the efficiency of the different control laws, several articles have been published [4][5][6]. Precision, computational cost, and stability are used as evaluation criteria in [20]. In [21], a performance index that considers both the thruster activity and the attitude tracking performance is used.

In this paper, five criteria of comparison have been used: EULERINT, ASCCT, settlement time at 5% (t_s), average computational cost (O), and steady state error (e_{ss}).

- **Error Euler Angle Integration (EULERINT).** Sidi [2] defines it as the integral of the error angle about the Euler axis of rotation. This is a quality indicator since it shows the accumulated angle error that the CubeSat travels to reach the desired position. The adapted formula to calculate it with Quartenions is described in (62). This parameter is similar to the attitude tracking performance metric defined in [21].

$$EULERINT = \int_0^T 2cos^{-1}(\delta q_4) \quad (62)$$

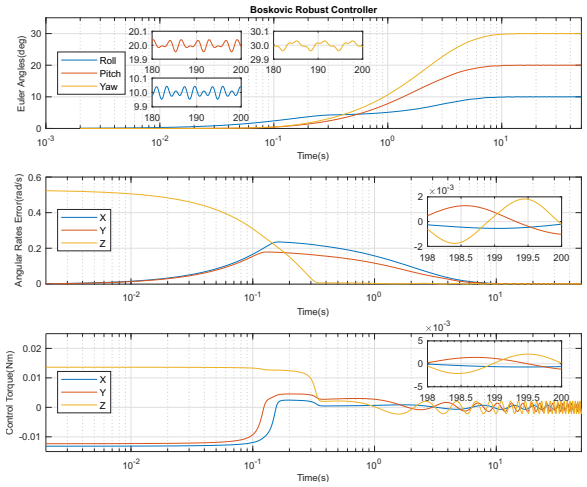


Fig. 5. Euler Angles and Angular Rate Errors produced by the miscellaneous disturbance torque in the Controller defined in (33) with a settling time (ts) of 6.44, $\pm$ 0.04 degrees in the Euler Angles oscillations, $\pm$ 1.8 $\times 10^{-3}$ rad/s oscillations in angular rates, and $\pm 2.1 \times 10^{-3}$ Nm in Control Torque.

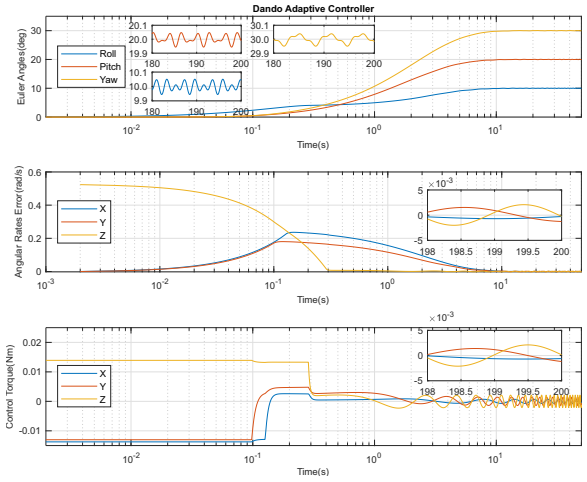


Fig. 6. Euler Angles and Angular Rate Errors produced by the miscellaneous disturbance torque in the Controller defined in (41) and (45) with a settling time (ts) of 6.48, $\pm$ 0.05 degrees in the Euler Angles oscillations, $\pm$ 2.0 $\times 10^{-3}$ rad/s oscillations in angular rates, and $\pm 2.1 \times 10^{-3}$ Nm in Control Torque.

- **Average of Square of the Commanded Control Torque (ASCCT).** It is defined by [12] as a measure of magnitude equivalent to the effective average torque exerted on the three satellite axes. This parameter is similar to the thruster activity performance metric defined in [21].

$$ASCCT = \frac{1}{T}\int_0^T ||u(t)||^2 dt \tag{63}$$

- **Computational Cost (O).** The average time that the algorithm takes to calculate the new control command 'u'.

The conducted simulations in Section VII-A were used to compare the control laws (27), (33), (41), and (52). The

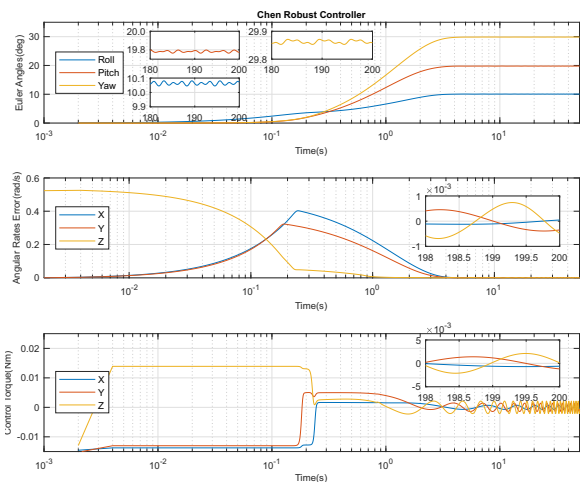


Fig. 7. Euler Angles and Angular Rate Errors produced by the miscellaneous disturbance torque in the Controller defined in (52) with a settling time (ts) of 2.75, $\pm$ 0.01 degrees in the Euler Angles oscillations, 0.2 Euler Angles offset, $\pm$ 0.7 $\times 10^{-3}$ rad/s oscillations in angular rates, and $\pm 2.1 \times 10^{-3}$ Nm in Control Torque.

calculated ASCCT and EULERINT are shown in Fig. 8.

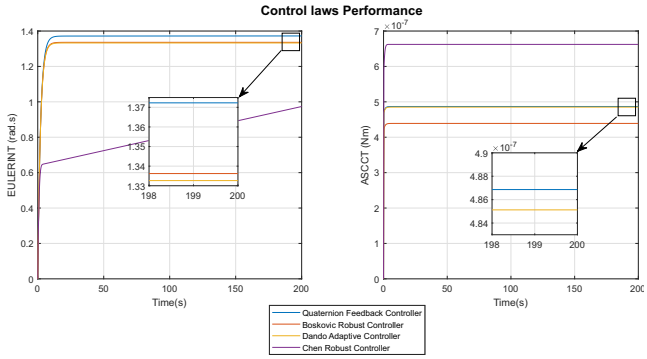


Fig. 8. EULERINT and ASCCT performance for control laws (27), (33), (41), and (52) without effects of miscellaneous disturbance torque.

VIII. Results

This section presents the results obtained in the numerical simulations developed in Section VII. The results obtained from Sections VII-A and VII-B are shown in Table IV and Table V, respectively.

TABLE IV. Simulation Results without Effects of Miscellaneous Torques

Controller	ts (5%)	EULERINT	ASCCT	O	e_{ss}
Feedback	6.83s	1.372 rad.s	0.486 μN.m	0.027ms	$-0.16\ 10^{-3}$
Boskovic	6.32s	1.336 rad.s	0.430 μN.m	0.16ms	$0.71\ 10^{-6}$
Dando	6.33s	1.333 rad.s	0.485 μN.m	0.20ms	$0.96\ 10^{-6}$
Chen	2.75s	Growing	0.640 μN.m	0.31ms	-0.21

The torque per gravity gradient for a balanced CubeSat of one unit is minimal. A magnitude of $[-3.14, 6.44, -5.02]^T \times$

TABLE V. SIMULATION RESULTS WITH EFFECTS OF MISCELLANEOUS TORQUES

Controller	ts (5%)	Amplitude in Euler Angles oscillations	OFFSET
Feedback	6.81s	0.04 deg	0 deg
Boskovic	6.44s	0.04 deg	0 deg
Dando	6.48s	0.05 deg	0 deg
Chen	2.75s	0.01 deg	0.2 deg

10^{-9} Nm for the Gravity Gradient Torque is obtained as seen in Appendix D (Fig. 13).

IX. CONCLUSION

In this article the simulation and comparison of four control laws in a CubeSat environment were carried out to find the most efficient control law that will be applied in the authors' future CubeSat projects.

The Quaternion Feedback Controller explained in Section VI-A is a basic algorithm that needs previous knowledge of the inertia tensor. Although, the miscellaneous torque rejection in this algorithm is not robust, it has an acceptable behavior rejecting constant inertia tensor uncertainties as can be seen in Fig. 4 and 9. However, the steady-state error is higher than the other robust algorithms and stability cannot be ensured in the case of control input constraints.

The Boskovic Robust Controller explained in Section VI-B is a robust algorithm that was developed considering the control input constraints so the authors ensure its global stability. To avoid the problem of chattering in the simulations, gain values for the adaptive parameter 'k' were chosen in order to not achieve convergence before the attitude and angular rate errors. As seen in Fig. 5, the controller does not reject the miscellaneous torque completely but the estimated computational cost is lower than the other robust algorithms as seen in Table IV. Moreover, the ASCCT parameter is the lowest but the EULERINT is only lower than the quaternion feedback algorithm.

The Dando Adaptive Controller law explained in Section VI-C is a robust algorithm that need a priori knowledge of the inertia tensor. Global stability cannot be ensured because this algorithm is not evaluated with control input constraints. Even though the saturation time to perform the maneuver is minimal, the behavior of the control law can be acceptable as seen in Fig. 6 and 11. The disturbance rejection is similar to Boskovic Controller but its steady state behavior is better than the other control laws.

The Chen Robust Controller law explained in Section VI-D is a robust controller that takes into account the control input constraints and needs a priori knowledge of the inertia matrix. Even though the asymptotic convergence in a finite time cannot be ensured as seen in Fig. 7 and 8, the disturbance rejection is better than the other controllers. In the simulations, the minimum steady-state is 0.02 degrees and the ASCCT torque needed to produce the rest-to-rest maneuver is the highest.

In order to choose the best control algorithm the priority was given to the steady-state error because it allows to perform precise attitude maneuvers. As a second evaluation criterion, the EULERINT parameter was chosen since it allows to perform a maneuver with the least possible error. Computational cost (O) was not consider a major factor as complex algorithms are not a challenge in modern micro-controllers. The settling time (ts) was not relevant for the chosen application.

Following these guidelines and according to the results obtained, it can be seen that the best performing control law for rest-to-rest maneuvers is the Boskovic Control law. This controller is capable of maintaining a steady-state error of 0.71×10^{-6} degrees while rejecting disturbances caused by misalignment, gravity gradient, and miscellaneous torques.

X. FUTURE WORK

Further research will implement and compare the studied control laws in an air bearing testing platform as proposed in [15]. Also, it is interesting to evaluate the performance of these control laws under noise effects and uncertainties in the overall system. Moreover, based on [20], further work will be focused on the measurement of the current consumed by each reaction wheel and the overall electric potential in each maneuver.

ACKNOWLEDGMENT

The authors are grateful for the support provided by the Center for Acquisition and Processing of Satellite Data (CAPDS) located in the Pedro Paulet Astronomical and Aerospace Institute (IAAPP) of the Universidad Nacional de San Agustin of Arequipa.

REFERENCES

[1] J. Carnahan, "CubeSat Design Specification," CubeSat Program, Cal Poly SLO, 1999.

[2] M. J. Sidi, "Spacecraft Dynamics and Control," Cambridge University Press, 1997.

[3] J. E. Slotine et al., "Applied nonlinear control," Englewood Cliffs, NJ: Prentice hall, 1991.

[4] H. J. Marquez, "Nonlinear control systems. John Wiley & Sons," 2003.

[5] Schaub, Hanspeter et al., "Feedback control law for variable speed control moment gyros," Journal of the Astronautical Sciences, 1998, vol. 46, no 3, p. 307-328.

[6] A. J. Dando, "Robust Adaptive Control of Rigid Spacecraft Attitude Maneuvers," PhD Thesis, Queensland University of Technology, 2008.

[7] J. Ahmed, V. T. Coppola, and D. S. Bernstein, "Adaptive asymptotic tracking of spacecraft attitude motion with inertia matrix identification," Journal of Guidance, Control, and Dynamics, 1998, vol. 21, no 5, p. 684-691.

[8] S. Scarritt, "Nonlinear model reference adaptive control for satellite attitude tracking," in AIAA Guidance, Navigation and Control Conference and Exhibit. 2008. p. 7165.

[9] J. D. Boskovic, S. Li, and R. K. Mehra, "Robust tracking control design for spacecraft under control input saturation," Journal of Guidance, Control, and Dynamics, 2004, vol. 27, no 4, p. 627-633.

[10] H. Chen, S. Song, and Z. Zhu, "Robust finite-time attitude tracking control of rigid spacecraft under actuator saturation," International Journal of Control, Automation and Systems, 2018, vol. 16, no 1, p. 1-15.

[11] V. Carrara and H. K. Kuga, "Estimating friction parameters in reaction wheels for attitude control," Mathematical Problems in Engineering, 2013, vol. 2013.

[12] B. Xiao et al., "Attitude tracking control of rigid spacecraft with actuator misalignment and fault," in IEEE Transactions on Control Systems Technology, 2013, vol. 21, no 6, p. 2360-2366.

[13] F. L. Markley and J. L. Crassidis, Fundamentals of spacecraft attitude determination and control. New York: Springer, 2014.

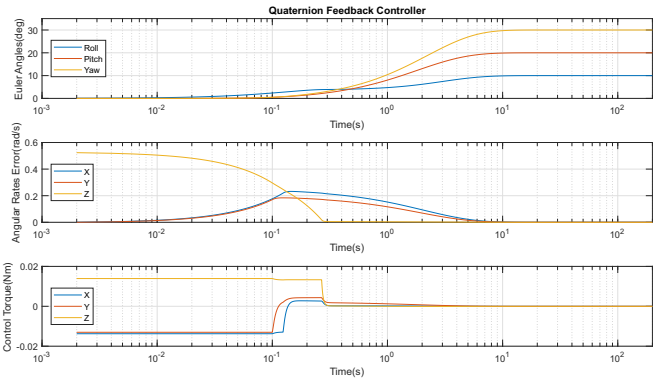


Fig. 9. Euler Angles, Angular Rate Errors and Control Torque produced by the Quaternion Feedback Controller defined in (27) with a settling time (ts) of 6.83 s and Average Computational Cost (O) of 0.027 ms. The Control Torque saturation time was 267.5 ms.

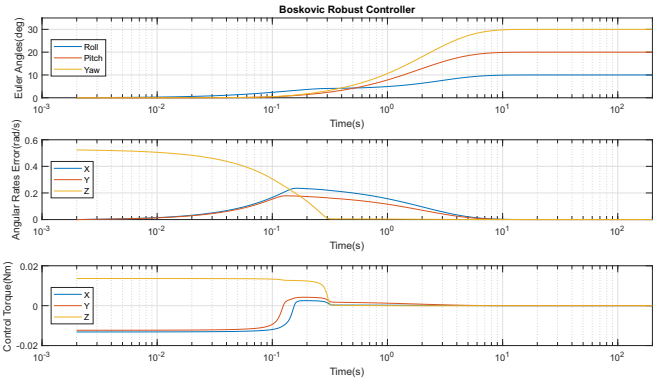


Fig. 10. Euler Angles, Angular Rate Errors and Control Torque produced by the Boskovic Robust Controller defined in (33) with a settling time (ts) of 6.32 s and average computational cost (O) of 0.159 ms. The Control Torque saturation time was 100.5 ms.

[14] P. R. Yanyachi, "Modelagem e Control de Attitude de Satélites Artificiais com Apendices Flexiveis," Doctoral Thesis, Sao Paulo University, Brasil, 2005.

[15] H. Mamani, "Design and implementation of a testing platform for the picosatelite CubeSat attitude control and determination system," Undergraduate Thesis, National University of San Agustin, Perú, 2019.

[16] T. T. Arif, "A Decentralized Adaptive Control of Flexible Satellite," 2007 IEEE Aerospace Conference, Big Sky, MT, 2007, pp. 1-7.

[17] P. S. Pereira da Silva, F. S. Freitas and P.R. Y. Aco-Cardenas, "Flat systems, flat information and an application to attitude control," 1999 European Control Conference (ECC), Karlsruhe, 1999, pp. 3202-3207.

[18] R. Mukherjee and D. Chen, "Asymptotic stability theorem for autonomous systems," Journal of Guidance, Control, and Dynamics, 1993, vol. 16, no 5, p. 961-963.

[19] I. Kök, "Comparison and analysis of attitude control systems of a satellite using reaction wheel actuators," Dissertation, 2012.

[20] A. Bello et al., "Experimental comparison of attitude controllers for nanosatellites," in 8th European Conference for Aeronautics and Space Sciences, 2019.

[21] N. Coulter and H. Moncayo, "Comparison of Optimal and Bioinspired Adaptive Control Laws for Spacecraft Sloshing Dynamics," Journal of Spacecraft and Rockets, 2020, vol. 57, no 1, p. 12-32.

APPENDIX A

Nomenclature used in this paper taken from Markley's book [13] .

TABLE VI. NOMENCLATURE

Name	Symbol
Attitude Quaternion	$q = [q_1, q_2, q_3, q_4]$
Quaternion vectorial part	q_{13}
Quaternion scalar part	q_4
Satellite angular velocity	ω
Reaction wheels angular velocity	Ω
Inertia tensor without reaction wheels	J_B
Total inertia tensor	J_B
Perpendicular reaction wheel inertia	$J^{\perp}$
Parallel reaction wheel inertia	$J^{\parallel}$
Perpendicular reaction wheel inertia matrix	J_{rw}
External torques	L
Reaction wheels torques	L_{rw}
Gravity gradient torque	T_{gg}
Reaction wheel radio	R_{rw}
Reaction wheel thickness	d_{rw}
Reaction wheel mass	m_{rw}
Error Quaternion	δq
Angular velocity error	$\delta\omega$
Earth's gravitational coefficient	μ
Distance from the center of Earth	r_c

APPENDIX B

Numerical integration method used in Section VII: The second-order Simpson's rule.

$$\int_{t_1}^{t_2} f(t)dt \approx \frac{dt}{6}\left[f(t_1) + 4f\left(\frac{t_1+t_2}{2}\right) + f(t_2)\right] \quad (64)$$

APPENDIX C

The rest-to-rest attitude maneuver for a non-spinning CubeSat simulated in Section VII-A.

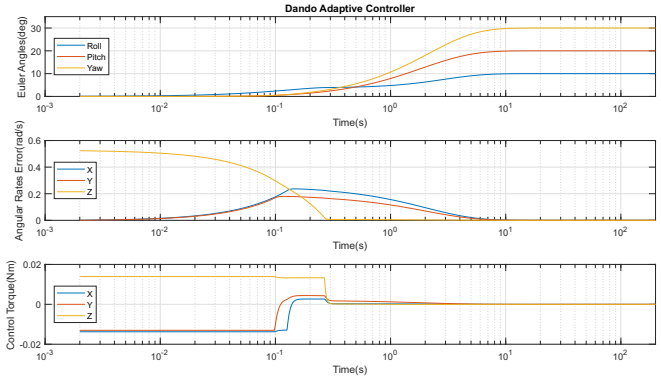


Fig. 11. Euler Angles, Angular Rate Errors and Control Torque produced by the Dando Adaptive Controller defined in (41) and (45) with a settling time (ts) of 6.33 s and average computational cost (O) of 0.200 ms. The Control Torque saturation time was 267 ms.

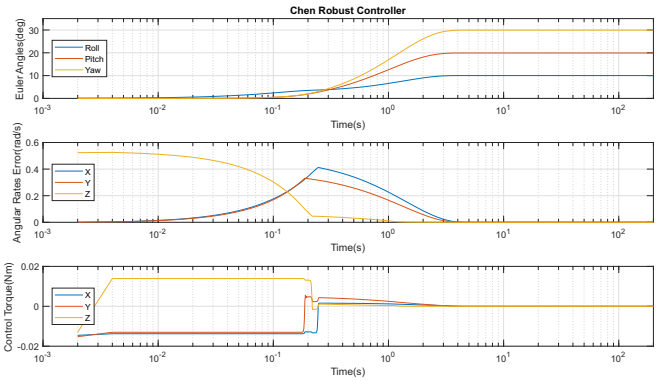


Fig. 12. Euler Angles, Angular Rate Errors, and Control Torque produced by the Chen Robust Controller defined in (52) with a settling time (ts) of 2.75 s and an average computational cost (O) of 0.309 ms. The Control Torque saturation time was 235 ms.

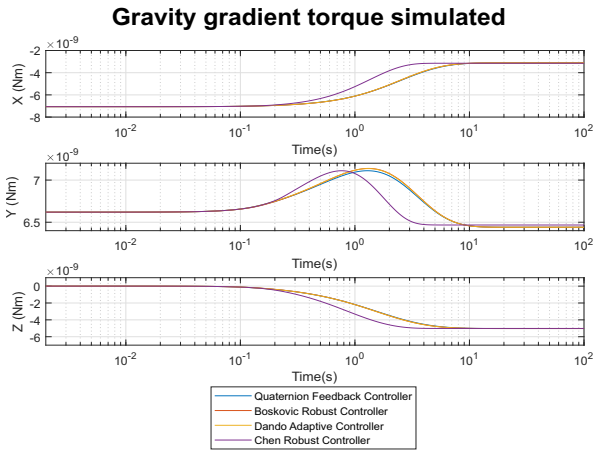


Fig. 13. Gravity gradient torque disturbance for a 100 Km low orbit CubeSat.

APPENDIX D

Gravity Gradient Torque described by (17) in Section IV.

Comparison of the CatBoost Classifier with other Machine Learning Methods

Abdullahi A. Ibrahim[1], Raheem L. Ridwan[2], Muhammed M. Muhammed[3], Rabiat O. Abdulaziz[4], Ganiyu A. Saheed[5]
Department of Mathematical Sciences, Baze University, Abuja, Nigeria[1,3]
African Institute for Mathematical Sciences, Accra, Ghana[2]
Department of Energy Engineering, PAUWES, University of Tlemcen, Algeria[4]
Institute of Mathematics, University of Silesia, Katowice, Poland[5]

***Abstract*—Machine learning and data-driven techniques have become very famous and significant in several areas in recent times. In this paper, we discuss the performances of some machine learning methods with the case of the catBoost classifier algorithm on both loan approval and staff promotion. We compared the algorithm's performance with other classifiers. After some feature engineering on both data, the CatBoost algorithm outperforms other classifiers implemented in this paper. In analysis one, features such as loan amount, loan type, applicant income, and loan purpose are major factors to predict mortgage loan approvals. In the second analysis, features such as division, foreign schooled, geopolitical zones, qualification, and working years had a high impact on staff promotion. Hence, based on the performance of the CatBoost in both analyses, we recommend this algorithm for better prediction of loan approvals and staff promotion.**

***Keywords*—Machine learning algorithms; data science; CatBoost; loan approvals; staff promotion**

I. Introduction

Machine learning and data-driven techniques have become very significant and famous in several areas. Some of the machine learning algorithms used in practice include; support vector machine, logistic regression, CatBoost, random forest, decision tree, AdaBoost, extreme gradient boosting, gradient boosting, naive Bayes, K-nearest neighbor, and many more. In supervised machine learning, classifiers have been widely used in areas such as fraud detection, spam email, loan prediction, and so on. In this work, we shall look into the applications of some machine learning methods in areas of loan prediction and staff promotion.

The issuance of loans is one of the many profit sources of financial institutions. However, the problems of default by applicants have been of major concern to credit providing institutions [1]. Studies conducted in the past were mostly empirical and as such the problems of default have not been definitively dealt with. The furtherance of time to the $21st$ century was accompanied by bulks of archived data collected from years of loan applications. Statistical techniques have been developed to study past data to develop models that can predict the possibility of defaults by loan applicants; thus, providing a score of creditworthiness. The availability of voluminous data called Big data necessitated the introduction of machine learning tools that can be used to discriminate loan applicants based on creditworthiness. This study considered some of these machine learning techniques to classify loan applicants based on available data to assess the probability of default and also recommend the technique that yields the best performance.

Since the advent of machine learning, several pieces of research has been conducted to discriminate against a loan applicants. In Goyal and Kaur [2], the authors developed an ensemble model by aggregating together Support Vector Machine (SVM), Random Forest (RF), and Tree Model for Genetic Algorithm (TMGA). The ensembled model was compared with each of these models individually and eight other machine learning techniques namely Linear Model (LM), Neural Network (NN), Decision Trees (DT), Bagged CART, Model Trees, Extreme Learning Machine (ELM), Multivariate Adaptive Regression Spline (MARS) and Bayesian Generalized Linear Model (BGLM) and was concluded from the analysis that the ensembled algorithm provided an optimum result. Alomari and Fingerman [3] tried to discriminate against loan applicants by comparing six machine learning techniques. The study compared DT, RF, K-Nearest Neighbour (KNN), OneR (1R), Naïve Bayes (NB), and Artificial Neural Networks (ANN) in which Random Forest gave the best performance with an accuracy of 71.75%. In Ibrahim and Rabiat [1], four classifiers were used to prediction in titanic analysis and XGBoost achieved the highest accuracy. Also, Ulaga *et al.* [4] conducted exploratory research where the suitability of RF was tested in classifying loan applicants and accuracy of 81.1% was achieved. In related research by Li [5], RF, BLR, and SVM were used to predict loan approvals and RF outperformed the other techniques with an accuracy of 88.63%. Xia *et al.* [6] predicted approvals for a peer-to-peer lending system by comparing Logistic Regression (LR), Random Tree (RT), Bayesian Neural Network (BNN), RF, Gradient Boosted Decision Trees (GBDT), XGBoost, and CatBoost and the results indicated that CatBoost gave the best performance over the other classifiers. The review of past literature showed tremendous developments in the applications of machine learning classifiers and how ensembled classifiers outperform single classifiers. However, only a few pieces of research considered CatBoost classifier in loan prediction approvals; hence, this research seeks to compare eight machine learning methods namely Binary Logistic Regression, Random Forest, Ada Boost, Decision Trees, Neural Network, Gradient Boost, Extreme Gradient Boosting, and CatBoost algorithms in the prediction of loan approvals.

The application of machine learning in employee promotion is another area we shall look into. Employees/staff play a

significant role in the development of an enterprise. Employee promotion in an enterprise is a major concern to both the employer and employee. In human resource management, staff promotion is very vital for organizations to attract, employ, retain, and effectively utilize their employee's talents [7]. Promotion of staff in an organization is based on some factors among which are age [8], gender [9], education [10], previous experience [11] and communication strategy or pattern [12]. In Long *et. al* [7], the authors applied some machine learning algorithms on Chinese data to predict employee promotion. It was discovered that, among all the available features in their dataset, the number of the different positions occupied, the highest departmental level attained and the number of working years affect staff promotion. In Sarkar *et. al* [13], joint data clustering, and decision trees were used to evaluate staff promotion. Saranya *et. al* [14] researched why the best and performing employees quit prematurely and predicted performing and valuable employees likely to quit prematurely. The proposed algorithm was recommended to the human resource department to determine valuable employees likely to quit prematurely. Previous works showed tremendous developments in the applications of machine learning but only few researchers have considered the CatBoost classifier in staff promotion. This research seeks to compare four machine learning methods namely Random Forest, Gradient Boost, Extreme Gradient Boosting, and CatBoost algorithms in the prediction of staff promotion.

Some of these literature only discussed the applications without emphases on the mathematics behind this algorithm. This paper will differ from others by highlighting the mathematics of the algorithm, the process of data cleaning, applying the supervised learning algorithms and evaluating these algorithms. This paper aim to develop a predictive machine learning model from supervised machine learning in areas of loan prediction and staff promotion. To achieve this aim, we shall set some objectives which will also be our contribution:

- Perform data science process such as exploratory analysis, perform data cleaning, balancing, and transformation
- Develop a predictive model from machine learning methods
- Apply some model evaluation metrics to determine the performance of the implemented models.

The rest of the paper is structured as follows: some machine learning algorithms are given in Section II while designs and nomenclatures are presented in Section III. Section IV presents the analytical results and Section V concludes the paper.

II. Materials and Algorithms

The following algorithms; Binary logistic regression, Random forest, Adaptive Boosting, Decision trees, Neural networks, gradient boost, XGBoost and Catboost, shall be discussed in this section.

A. Binary Logistic Regression

Consider a dataset with response variable (Y) classified into two categories, Y = 'Loan approved','not approved' or $Y = \{'promoted','notpromoted'\}$. Logistic regression models the probability of Y belongs to a specific category. With approach (1) below to predict this probability:

$$p(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{1}$$

The conditions $p(X) < 0$ and $p(X) > 0$ can be predicted for values of X, except for range of X is limited. To keep away from this, $p(X)$ must be modelled with the help of a logistic function that generates between 0 and 1 values as output. The function is defined as in (2)

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}} \tag{2}$$

The *'maximum likelihood'* method is used to fit (2). The unknown coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ in (2) should be approximated based on the data available for training the model. The intuition of likelihood function can be expressed mathematically as in (3):

$$\ell(\beta_0, \ldots, \beta_n) = \prod_{i:y_i=1} p(x_i) \prod_{i':y'_i=0} (1 - p(x'_i)) \tag{3}$$

The estimates $\beta_0, \ldots, \beta_n$ are selected to maximize this function [1]. More explanation can be obtained in [15].

Basic Assumptions of Binary Logistic Regression

(i) The response variable must be binary.

(ii) The relationship between the response feature and the independent features does not assume a linear relationship.

(iii) Large sample size is usually required.

(iv) There must be little or no multicollinearity.

(v) The categories must be mutually exclusive and exhaustive.

B. Random Forest

Random forest (RF) algorithm is a well-known tree-based ensemble learning method and the bagging-type ensemble [16]. RF differs from other standard trees, each node is split using the best among a subset of predictors randomly chosen at that node [17]. This additional layer of randomness is what makes RF more robust against over-fitting [18]. To improve the bagged trees in RF, a small tweak that de-correlates the trees are made. As in bagging, we build several decision trees on bootstrapped training sets. But when building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p-predictors [19]. The RF approach for both classification and regression is presented in Algorithm 1.

Algorithm 1 Random Forest Algorithm

(i) Draw m_{tree} bootstrap samples from the initial data.

(ii) Initialize an *unpruned* tree, for every bootstrap sample, with the modification given as follow: instead of choosing the best-split among all predictors at each node, sample randomly n_{try} of the predictors and select the best-split from among those features. Bagging can be seen as a special case of random forests which can be obtained when $n_{try} = k$, number of predictors.

(iii) new data is predicted by aggregating the predictions of the m_{tree} trees.

C. Adaptive Boosting

Adaptive boosting (AdaBoost) algorithm is another machine learning method used to improve the accuracy of other algorithms. It is a boosted algorithm generated by training weaker rules to develop a boosted algorithm. In Adaboost, training sets $(x_1, y_1), \ldots, (x_m, y_m)$ is the input, where each x_i belongs to some *instance space X*, and each *feature* y_i is in some label set Y (in this case assuming that $Y = \{-1, +1\}$. This method calls repeatedly a given weak or base learning algorithm in a given series of rounds $t = 1, \ldots, T$. One of the significant and vital ideas of the algorithm is to keep a distribution or set of weights over the training set. The weight of this distribution on training samples *i* on round *t* is represented by $D_t(i)$.

At initial, all weights are set equally, but on each round, the weights of misclassified samples are increased so that the weak learner is forced to focus on the hard samples in the training set. The weak learner's job is to find a weak hypothesis $h_t : X \to \{-1, +1\}$ appropriate for the distribution D_t [20]. A metric used to measure the goodness of a weak hypothesis is its error. The algorithm procedure is presented in algorithm 2.

Algorithm 2 Adaboost Algorithm

Given: $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_i \epsilon X$, $y_i \epsilon Y = \{-1, +1\}$

Initialize: $D_1(i) = \frac{1}{m}$ for $i = 1, \ldots, m$. For $t = 1, \ldots, T$:

- Train weak learner using distribution D_t.
- Get weak hypothesis $h_t : X \to \{-1, +1\}$ with error
- $\epsilon_t = Pr_{i \sim D_t}[h_t(x_i) \neq y_i]$
- Choose $\alpha_t = \frac{1}{2} \ln(\frac{1 - \epsilon_t}{\epsilon_t})$
- Update: $D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} exp(-\alpha_t) & \text{if } h_t(x_i) = y_i \\ exp(\alpha_t) & \text{if } h_t(x_i) \neq y_i \end{cases} = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$

where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).

Output the final hypothesis:

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$$

D. Decision Trees

Decision trees are one of the supervised learning algorithms that can be applied to both classification and regression problems [21]. We shall briefly consider regression and classification tree problems. There are two steps (as explained in [21]) for building a regression tree:

(i) Divide the set of feasible values $X_1, \ldots, X_n$ for into $I-$distinct and non-overlapping regions, $R_1, R_2, \ldots, R_i$.

(ii) For each sample that falls into R_i, the same prediction is made, which is the average of the dependent feature for the training sets in R_i.

In order to construct regions $R_1, \ldots, R_i$, we elaborate on step (i) above. In theory, $R_1, R_2, \ldots, R_i$ could take any shape or dimension.

However, for simplicity, we may split the predictor space into high-dimensional boxes and for easy interpretation of the predictive model. The aim is to obtain boxes $R_1, \ldots, R_i$ that minimizes the Residual Sum of Squares (RSS) as given in the mathematical expression in (4)

$$\sum_{i=1}^{I} \sum_{j \in R_i} (y_j - \hat{y}_{R_i})^2 \quad (4)$$

Where $(\hat{y}_{R_i})$ is the mean response of the training sets in the ith box.

The classification tree on the other hand predicts a qualitative response variable. In a classification tree, we predict that every observation belongs to the 'most frequently occurring' class of training sets in the region to which it belongs since

we intend to allocate sample in a given region to the 'most frequently occurring' class of training sets in that region, the classification error rate is the part of the training sets in that region that do not belong to the most frequent class, as given in (5).

$$E = 1 - \max_{l}(\hat{p}_{ml}) \tag{5}$$

where $\hat{p}_{ml}$ denotes the ratio of training samples in the mth region from lth class. However, it turns out that classification error is not sensitive enough for tree-growing. The *Gini index* which is defined mathematically in (6)

$$G = \sum_{l=1}^{L} \hat{p}_{ml}(1 - \hat{p}_{ml}) \tag{6}$$

A measure of total variance over the L classes. Further details can be found in [21].

E. Neural Network

An Artificial Neural Network (ANN) is an imitation of the interconnections made up in the human brain. The inputs in ANN represent the dendrites in the human brain which receives electrochemical signals from other neurons into the cell body. Every input carries a signal which is obtained by the product of its weight and the input to a hidden layer in the neuron powered by an activation function usually a sigmoid function, other activation functions like tangent hyperbolic function, linear function, step function, ramp function, and Gaussian function can also be used [22]. The last layer is the output layer which represents the axon extending to the synapse that connects two different neurons. A typical ANN architecture has inputs, output, and a bias. The ANN architecture differs majorly by layers. The most common and simple architecture is a Perceptron which has two inputs, a hidden layer, and a single output. The neural networks are mostly backpropagated to be used for classification and prediction. The back and forth movement in a neural network between the input and output layers is referred to as an epoch. A neural network undergoes several epochs until a tolerable error is achieved and thus the training of an artificial neural network is achieved. ANN architecture is shown in Fig. 1.

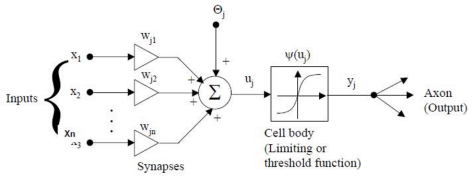


Fig. 1. Architecture of an Artificial Neural Network [23]

where Θ = external threshold, offset or bias w_{ji} = synaptic weights x_i = inputs y_i = output as in (7)

$$y_i = \psi(\sum_{i=1}^{n} w_{ji}x_i + \Theta_i) \tag{7}$$

F. Gradient Boost

Gradient boost is a boosted algorithm used for regression and classification. It is derived from the combination of Gradient Descent and Boosting. It involves fitting an ensemble model in a forward stage-wise manner. The first attempt to generalize an adaptive boosting algorithm to gradient boosting that can handle a variety of loss functions was done by [24], [25]. The steps for gradient boosting algorithm is outlined in algorithm 3.

Algorithm 3 Gradient Boost Algorithm

Inputs:

- Input data $(x, y)_{i=1}^{N}$
- number of iterations M
- choice of the loss-function $\psi(y, f)$
- choice of the base-learner model $h(x, \theta)$

Algorithm:

- intialize $\hat{f}_0$ with a constant
- compute the negative gradient $g_t(x)$
- fit a new base-learner function $h(x, \theta_t)$
- find the best gradient descent step-size ρ_t : $\rho_t = \arg\min_{\rho} \sum_{i=1}^{N} \psi[y_i, \hat{f}_{t-1}(x_i) + \rho h(x_i, \theta_t)]$
- update the function estimate: $\hat{f} \leftarrow \hat{f}_{t-1} + \rho_t h(x, \theta_t)$
- end for

G. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is one of the boosted tree algorithms [16], which follows the principle of gradient boosting [24]. When compared with other gradient boosting algorithms, XGBoost makes use of a more regularized model formalization in other to control over-fitting of data, which gives it better performance [16]. In other to achieve this, we need to learn functions h_i, with each containing structure of tree and leaf scores [26]. As explained in [27], Given a data with m-samples and n-features, $\mathcal{D} = \{(X_j, y_j)\}(|D| = m, X_j \in \mathbb{R}^n, y_j \in \mathbb{R})$ a tree ensemble model makes use of L additive functions to predict the output as presented in (8).

$$\hat{y}_j = \phi(X_j) = \sum_{l=1}^{L} h_l(X_J), \qquad h_l \in \mathcal{H} \tag{8}$$

where $\mathcal{H} = \{h(X) = w_q(X)\}(q : \mathbb{R}^n \rightarrow U, w \in \mathbb{R}^U)$ is the space of regression trees. q denotes the structure of each tree that maps a sample to its corresponding leaf index. U denotes number of leaves in the tree. Each h_l corresponds to independent structure of tree q and leaf weights w.

To learn the set of functions used in the model, the regularized objective is minimized (9) as follows:

$$\mathcal{L}(\phi) = \sum_{j} l(\hat{y}_j, y_j) + \sum_{l} \Omega(h_l),\ \Omega(h) = \gamma U + \frac{1}{2}\lambda||w||^2 \tag{9}$$

where l is differentiable convex loss function which measures difference between the target y_j and predicted $\hat{y}_j$. Ω

penalizes the complexity of the model to avoid over-fitting. The model is trained in an additive way. A score to measure the quality of a given tree structure q is derived as given in (10):

$$\hat{\mathcal{L}}^{(u)}(q) = -\frac{1}{2}\sum_{j=1}^{U}\frac{(\sum_{i=I_j} f_i)^2}{\sum_{i=I_j} g_i + \lambda} + \gamma U \tag{10}$$

where $f_i = \partial_{\hat{y}^{(u-1)}} l(y_i, \hat{y}^{(u-1)})$ and $g_i = \partial^2_{\hat{y}^{(u-1)}} l(y_i, \hat{y}^{(u-1)})$ are the gradient and second order gradient statistics, respectively. Further explanation can be obtained in [27].

H. CatBoost

Another machine learning algorithm that is efficient in predicting categorical feature is the CatBoost classifier. CatBoost is an implementation of gradient boosting, which makes use of binary decision trees as base predictors [28]. Suppose we observe a data with samples $D = \{(X_j, y_j)\}_{j=1,\ldots,m}$, where $X_j = (x_j^1, x_j^2, \ldots, x_j^n)$ is a vector of n features and response feature $y_j \in \mathbb{R}$, which can be binary (i.e yes or no) or encoded as numerical feature (0 or 1). Samples (X_j, y_j) are independently and identically distributed according to some unknown distribution $p(\boldsymbol{.}, \boldsymbol{.})$. The goal of the learning task is to train a function $H : \mathbb{R}^n \to \mathbb{R}$ which minimizes the expected loss given in (11)

$$\mathcal{L}(H) := \mathbb{E}L(y, H(X)) \tag{11}$$

where $L(\boldsymbol{.}, \boldsymbol{.})$ is a smooth loss function and (X, y) is a testing data sampled from the training data D.

The procedure for gradient boosting [24] constructs iteratively a sequence of approximations $H^t : \mathbb{R}^m \to \mathbb{R}, t = 0, 1, \ldots$ in a greedy fashion. From the previous approximation H^{t-1}, H^t is obtained in an additive process, such that $H^t = H^{t-1} + \alpha g^t$, with a step size α and function $g^t : \mathbb{R}^n \to \mathbb{R}$, which is a base predictor, is selected from a set of functions G in order to reduce or minimize the expected loss defined in (12):

$$\begin{aligned} g^t &= \arg\min_{g\in G}\mathcal{L}(H^{t-1} + g) \\ &= \arg\min_{g\in G}\mathbb{E}L(y, H^{t-1}(X) + g(X)) \end{aligned} \tag{12}$$

Often, the minimization problem is approached by the Newton method using a second-order approximation of $\mathcal{L}(H^{t-1} + g^t)$ at H^{t-1} or by taking a (negative) gradient step. Either of these functions is gradient descent [29], [30]. Further explanation of CatBoost algorithm can be obtained in [28].

III. Design and Nomenclatures

Some evaluation metrics such as confusion matrix, the area under the curve (AUC), accuracy, error rate, true positive rate, true negative rate, false-positive rate, and false-negative rate shall be discussed.

A. Confusion Matrix

A confusion matrix contains information about actual and predicted classifications from a classifier. The performance of such a classifier is commonly evaluated using the data in the matrix. Table I shows the confusion matrix for classifier [1], [31].

TABLE I. Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	True Negative (TN)	False Negative (FN)
	Positive	False Positive (FP)	True Positive (TP)

True Positive: The classifier predicted a true event and the event is actually true.

True Negative: The classifier predicted that an event is not true and the event is actually not true.

False Positive: The classifier predicted that an event is true but the event is actually not true.

False Negative: The classifier predicted that an event is not true but the event is actually true.

The confusion matrix can be interpreted as: the TN and TP are the correctly classified classes while FN and FP are the misclassified classes.

B. Model Evaluation Metrics

The Model training time, model accuracy, and memory utilized are some good metrics for comparing the performance of the classifiers. Also, the area under the Receiver Operating Characteristics Curve (ROC-AUC) is a performance metric for classification accuracy. The AUC is another metric which checks the performance of multiple-class classification accuracy [26]. Model accuracy is the proportion of the correct predictions (True positive and True negative) from the total predictions defined in (13).

$$\begin{aligned} \text{Accuracy} &= \frac{TN + TP}{TP + TN + FP + FN} \times 100\% \\ \text{Error Rate} &= \frac{FP + FN}{FP + FN + TP + TN} \times 100 \end{aligned} \tag{13}$$

The error rate is the proportion of all incorrect predictions divided by the total number of samples, given in (13).

True Positive Rate (TPR), also called the sensitivity or recall, is the proportion of correct positive predicted class from total positive class. The best sensitivity is 1.0 and the worst is 0.0. True Negative Rate (TNR), also called the specificity, is the proportion of correct negative predictions from the total number of negative classes. The best specificity is 1.0 and the worst is 0.0. The TPR and TNR are given in (14).

$$\begin{aligned} \text{True Positive Rate} &= \frac{TP}{FN + TP} \times 100 \\ \text{True Negative Rate} &= \frac{TN}{FP + TN} \times 100 \end{aligned} \tag{14}$$

Precision is the number of correctly predicted positive value out of the total number of positive class, as given in (15). False Positive Rate (FPR) is the number of incorrect positive prediction out of the total number of negatives as in (15).

$$\begin{aligned} \text{False Positive Rate} &= \frac{TP}{FNP+TP} \times 100 \\ \text{False Negative Rate} &= \frac{FP}{FP+TN} \times 100 \end{aligned} \tag{15}$$

C. Calibration Plots

Calibrated methods (classifiers) are probabilistic classifiers for which the outcome of the predicted probabilities of a particular classifier can be interpreted as a confidence interval. The metric is used to determine whether the predicted probability can be interpreted as a confidence interval.

D. System Specification

All classifiers were run on Jupyter notebook in python 3.7.4 on Linux 19.10 version. The codes were run on 8GB HP elite book, core $i5$.

IV. Results and Discussion

In this section, we shall perform two analyses to determine the performance of all the machine learning algorithms discussed previously. We begin by exploring the data to obtain the numerical statistics, identify missing values, outliers, and if the independent feature is balanced or not. After initial exploration we were able to identify missing values and outliers, the independent feature is balanced.

A. Analysis 1: Predicting Mortgage Approvals from Government Data

The analysis is based on US Government data concerning predicting mortgage approvals [32]. This is a binary classification problem. Our analysis was based on the $500,000$ observations with 23 features from the training data-set of mortgage approvals government data, each containing specific characteristics for a mortgage application which will either get approval ("1"); or not ("0"). We tested our model on a data-set with 150,000 samples.

1) Exploratory Analysis: Before developing a predictive model, we need to understand the data-set by exploratory analysis. In the exploratory analysis, we intend to find answers to some questions such as (i) which features have missing values, (ii) features with outliers, (iii) is the response feature balanced? (iv) the distribution of the data points and so on. We present some visualizations in Fig. 2 and 3 to answer these questions.

Fig. 2 shows the both classes give almost the same frequency, with $250,114$ for the *accepted* data points and $249,886$ for *not accepted* data points. The data distribution seem balanced and other analysis can be performed. Fig. 3 shows the distributions of the three classes of the loan purpose that we have. The Loan amount follows a normal distribution for both the *accepted* and *not accepted* in Fig. 4. Fig. 5 shows the two classes of loan type. The

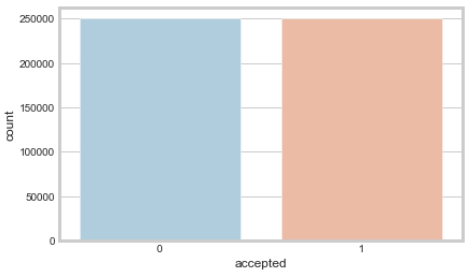


Fig. 2. Response Features

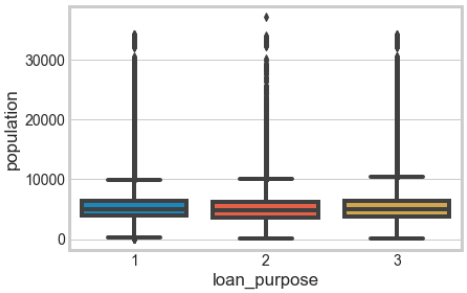


Fig. 3. Loan Purpose

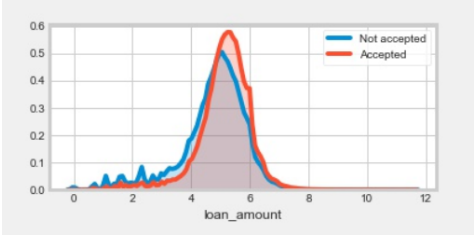


Fig. 4. Loan Amount

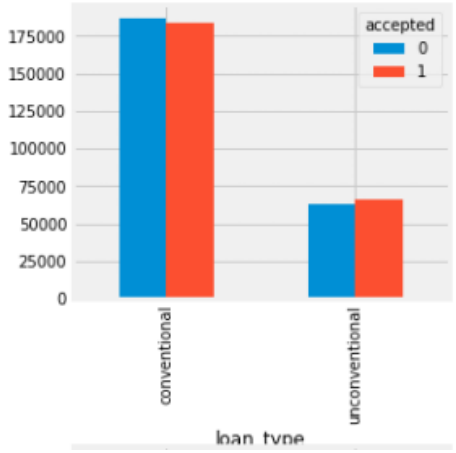


Fig. 5. Loan Type

accepted class of the conventional loan type has highest frequency with around $180,000$. In Fig. 6 state code 6 has the highest number ($\approx$ 890) of district lenders, with state code 50 having the least. Fig. 7 shows the fea-

tures "msa-md, applicant-income, number-of-owner-occupied-units, number-of-1-to-4-family-units, tract-to-msa-md-income-pct" having numbers of 76982, 39948, 22565, 22530, 22514, respectively. Generally, from our visualizations we can see that the major features that contribute to the prediction of mortgage loan are loan amount, loan type, applicant income and loan purpose.

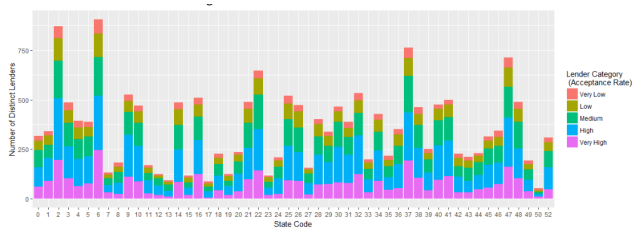


Fig. 6. District Lenders by State Code

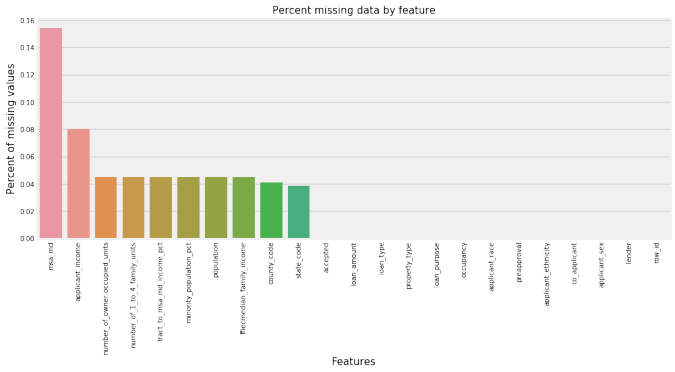


Fig. 7. Features with Missing Information

2) Data Pre-processing: To replace the missing values (NA's) for both numerical and categorical features. Starting with the categorical features, the NA's encountered were replaced with the mode of that feature. Also, categorical features were one-hot encoded, which means each of the distinct categories in a particular feature was converted to numerical fields. For numerical features, the NA's were replaced on a case by case basis. Features like "applicant-income, number of owner-occupied units" were replaced with the median as it handles the presence of outliers, unlike mean imputation. The test set used has 150, 000 samples for each of the models.

TABLE II. PERFORMANCE COMPARISON OF THE ALGORITHMS

Algorithm	Score	Avg. time (fit)	Avg. time (score)(s)	F1 score	AUC	Precision
Logistic Regression	0.62	73.806	0.032	0.62	0.67	0.61
Random Forests	0.69	19.045	1.271	0.64	0.71	0.68
Adaboost	0.67	28.632	1.664	0.63	0.72	0.66
XGBoost	0.69	28.322	1.342	0.65	0.75	0.68
Neural Networks	0.68	27.234	1.123	0.73	0.66	0.66
Gradient Boosting	0.69	30.233	3.432	0.66	0.75	0.68
CatBoost	**0.732**	46.657	6.725	0.75	0.78	0.83
Decision Trees	0.68	4.272	1.012	0.054	0.62	0.66

3) Results and Discussion: False positive rate is a method of committing a type I error in null hypothesis testing when conducting multiple comparisons. For the problem used in this paper, the false positive rate is an important metric as it would be a disaster if the system predicts a client would be given a loan but in reality, he was not. From Table II, the CatBoost algorithm achieved the highest accuracy. This means that the confusion metrics for CatBoost, the value of correctly classified (TP + FN) is higher than the other six algorithms implemented. And with the least number of miss-classified. Other metrics such as f1 score, AUC, and precision are also shown in Table II.

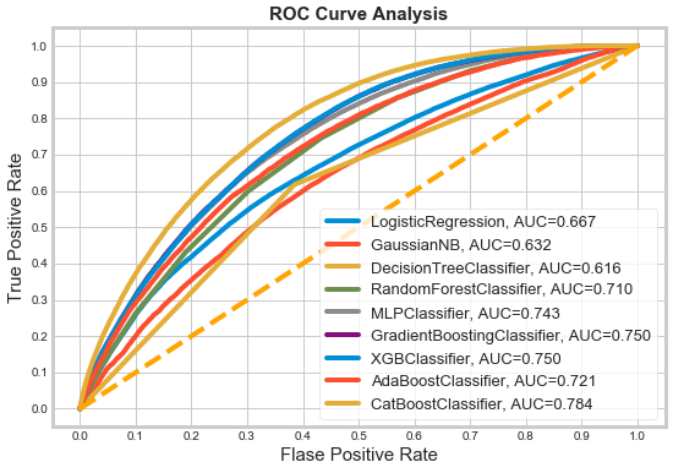


Fig. 8. ROC Curves for the Algorithms

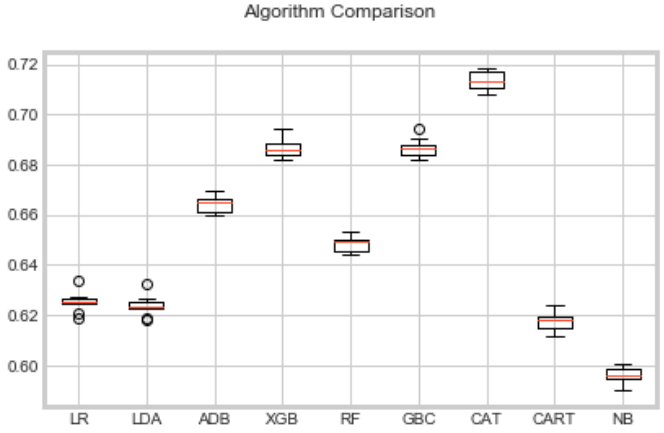


Fig. 9. Boxplots for the Algorithms

After training the models on the training set and predicted the probabilities on the test set, we then obtained the true positive rate, false positive rate, and AUC scores. From Fig. 8, CatBoost achieved highest AUC value of 0.78 which is closer to 1 than other classifiers. Also, Fig. 9 shows the comparison with other implemented algorithms. The names of the algorithms are written in the short form where LR denotes Logistic regression, ADB for AdaBoost, RF denotes random forest, GBC for gradient boosting, CART for decision trees, NB for naive Bayes, XGB denotes XGBoost, and CAT for CatBoost algorithms. Box 9 shows the spread of the accuracy scores across each cross-validation fold for each algorithm. Box 9 is generated based on the mean accuracy and standard

deviation accuracy of the algorithms. In Fig. 10, the calibration plots for all the implemented algorithms are plotted, CatBoost method produced well-calibrated predictions as it optimizes log-loss. Fig. 11 shows the model performance graph for the CatBoost classifier.

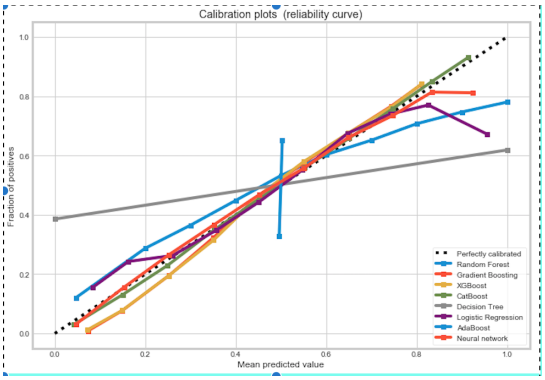


Fig. 10. Calibration Plots

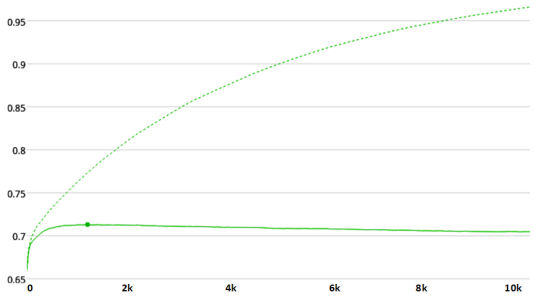


Fig. 11. Graph of the Model Training

From the plot 9, it would suggest that CatBoost is perhaps worthy of further study on this problem due to its performance. The result has been presented in Table II which contains the model accuracies, AUC, the average time to fit, and score. In summary, features such as *loan amount*, *loan type*, *applicant income* and *loan purpose* played a key role in predicting mortgage loan approvals. This mean, for an individual to get a mortgage loan, the amount of loan, what type of load, income of the loan applicant and purpose for wanting to secure loan are the key questions that needs to be addressed before a loan will be approved.

B. Analysis 2: Staff Promotion Algorithm

HR analytics using machine learning will revolutionize the way human resources departments now operate. This will lead to higher efficiency and better results overall. This analysis uses predictive analytics in identifying the employees most likely to get promoted or not using historical staff promotion datasets [32]. We trained the model on a dataset with $38,312$ samples and 19 features and tested it on a data-set with $16,496$ samples.

1) Exploratory Analysis: Before developing a predictive model, we need to understand the data points and have a pictorial view of what the data-set contains. In the exploratory analysis, we intend to find answers to some questions such as (i) which features have missing values, (ii) features with outliers, (iii) is the response feature balanced? (iv) the distribution of the data points and so on. We present some visualizations in Fig. 12, 13, 14, 15, 16, 17, and 18 to answer these questions.

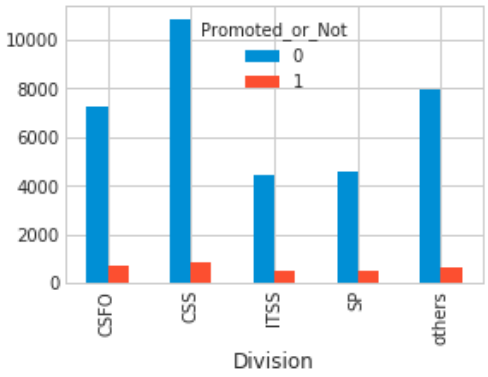


Fig. 12. Division

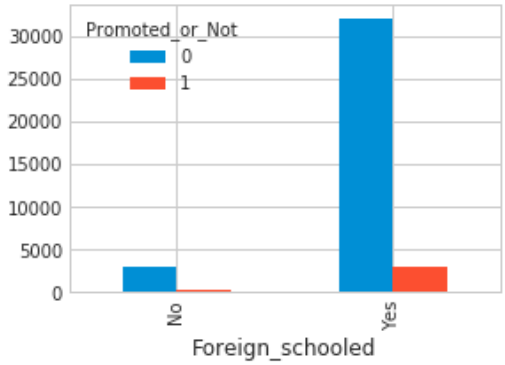


Fig. 13. Foreign School

Starting with Fig. 12, the CSS class have the highest number (over 1000) of staff that was promoted in the Division feature. In Fig. 13, more than 30000 staff who got promoted were foreign-schooled and $\approx$ 2500 studied locally. Fig. 14 shows the geographical zones of staffs, the South-West zone recorded the highest number of promoted staff while the North-East zone has the least number of promoted staff. Fig. 15 shows the frequency for the two classes in the response variable. It was observed that most of the staffs fall in the "not promoted" class, with a ratio of "promoted" to "not promoted" as $8:92\%$.

Furthermore, in Fig. 16, employees from Oyo state (in South-West) appears to have the highest number of working years (38 years) while employees from Zamfara state had 24 years of working experiences. This could further support Fig. 14 with staff from the South-West zone having the highest

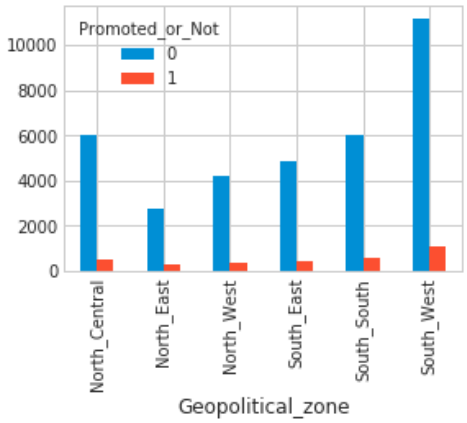


Fig. 14. Geographical Zones

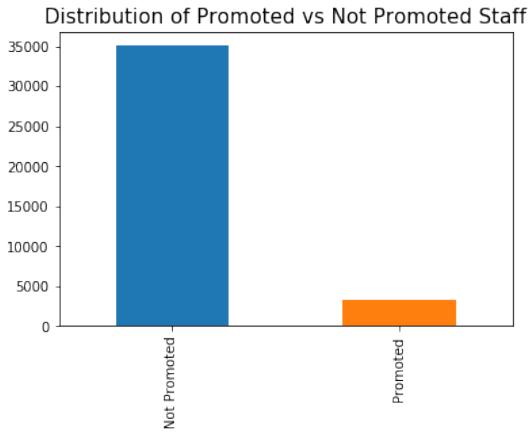


Fig. 15. Response Feature

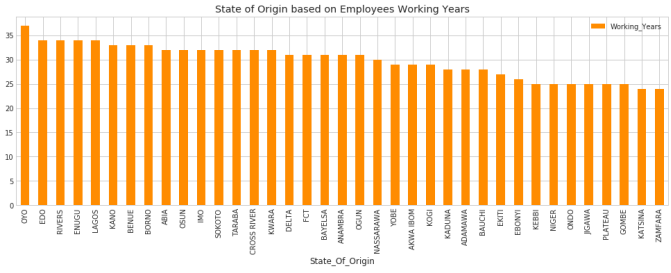


Fig. 16. State of Origin

number of promotion because of their working years. In Fig. 17, the graduate employees appear older than the Non-graduate employee. This could be due to the number of years spent studying before joining the workforce. The distribution of staff channel of recruitment is shown in Fig. 18. In summary, features such as division, foreign schooled, geopolitical zones, qualifications, and working years had a high impact on staff promotion.

2) Data-preprocessing: We replaced the missing values (NA's) for both numerical and categorical features. For the categorical features, the NA's encountered were replaced with

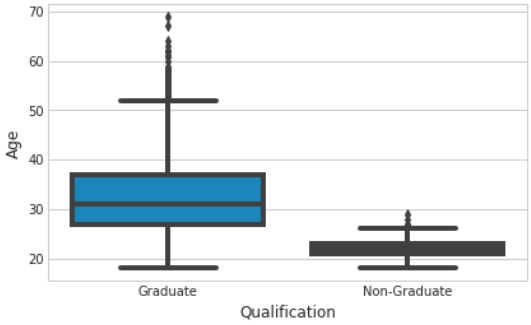


Fig. 17. Educational Status

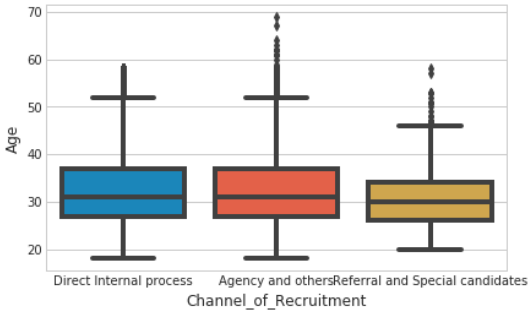


Fig. 18. Channel of Recruitment

the mode of that feature. For Numerical features, the NA's were replaced on a case by case basis. For the imbalanced response feature, it was balanced with a Resampling technique, to improve our prediction. After this step, we split the data and proceed to the prediction phase. The test set used has $16,496$ samples for each of the models.

3) Results and Discussion: Table III shows a summary of the evaluation metrics for implemented algorithms. CatBoost and XGBoost achieved the highest score with 94%. And when uploaded into Kaggle [33] online, we had a difference of 0.01. Other metrics are also shown in this table.

TABLE III. PERFORMANCE COMPARISON OF THE ALGORITHMS

Algorithm	Score (PC)	Score (Kaggle)	AUC	Precision	F1-Score
Random forest	0.93	0.88	0.71	0.70	0.94
XGBoost	0.94	0.93	0.82	0.93	0.92
Gradient Boost	0.90	0.84	0.82	0.93	0.95
CatBoost	**0.94**	0.93	0.82	0.91	0.95

In Fig. 19 the AUC value of the applied algorithms are plotted, the random forest classifier had the least value of 0.71 while other algorithms achieved 0.82. The distribution of the algorithms is shown in Fig. 20. Again, the random forest classifier achieved the least value while other algorithms achieved high values. The model performance graph for the CatBoost algorithm showing how the model was trained, the number of iterations, and the accuracy is shown in Fig. 21.

The proportion of the training set and the test error rate is plotted in Fig. 22. CatBoost and XGBoost had little error compared to other implemented algorithms. Fig. 23 shows that the CatBoost and XGBoost methods produced well-calibrated predictions.

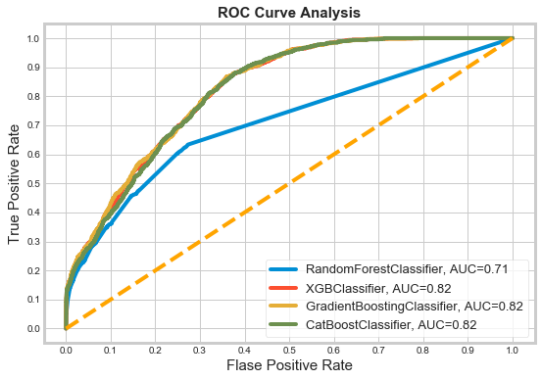


Fig. 19. ROC Curves for the Algorithms

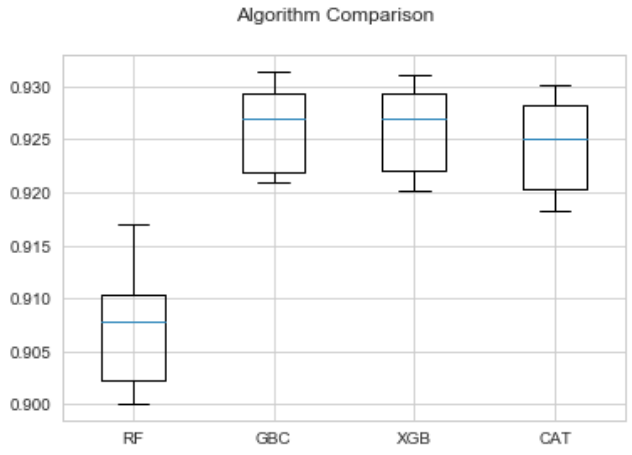


Fig. 20. Boxplots for the Algorithms

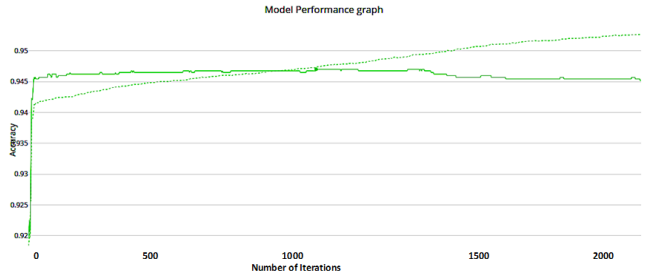


Fig. 21. Graph showing the Model Performance of the CatBoost Algorithm

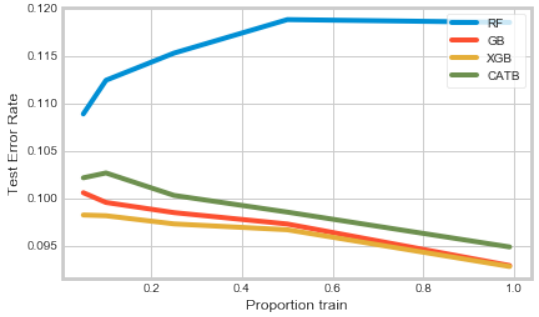


Fig. 22. Reliability Curves for the Test Error Rate

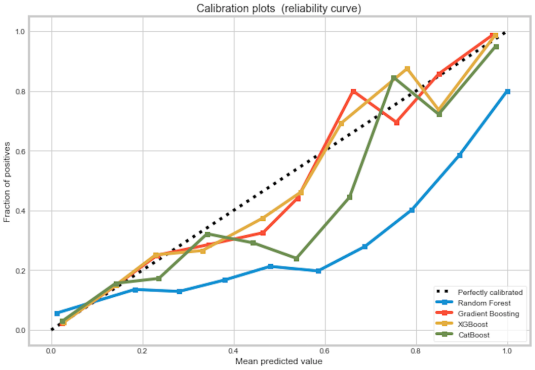


Fig. 23. Reliability Boxplots for the Algorithms

V. Conclusion

Having applied all the mentioned algorithms in our methodology, this paper aimed to compare some predictive machine learning algorithms from supervised learning with applications in areas of loan prediction and staff promotion. We performed two analyses: loan prediction and staff promotion. Each analysis started with exploratory analysis where we find insights from the data, then the data was cleaned, balanced, and transformed for prediction. The machine learning algorithms discussed in this paper were then implemented and some metrics were used to evaluate the implemented models' performance. CatBoost classifier did pretty well achieving the highest score (accuracy) in both applications (or analysis). Other evaluation metrics also support the performance of this algorithm. We thereby recommend the CatBoost classifier for the predictive model.

For the mortgage loan analysis, features such as *loan amount*, *loan type*, *applicant income* and *loan purpose* played a significant role in predicting mortgage loan approvals. And for the staff promotion analysis, features such as *division*, *foreign schooled*, *geopolitical zones*, *qualifications*, and *working years* had a significant impact towards staff promotion.

Future work might consider cross-validation. Cross-validation could also be used to compute the model's accuracy based on different combinations of training and test samples. Besides, some other classifiers with larger datasets may be applied.

Acknowledgment

The authors would like to appreciate Sowole Samuel for his suggestions.

References

[1] Abdullahi Adinoyi Ibrahim and Rabiat Ohunene Abdulaziz, *Analysis of Titanic Disaster using Machine Learning Algorithms*, Engineering Letters, vol. 28, no. 4, pp1161-1167, 2020.

[2] Anchal Goyal and Ranpreet Kaur, *Loan Prediction using Ensemble Technique.* International Journal of Advanced Research in Computer and Communication Engineering. Vol. 5(3), 2016.

[3] Zakaria Alomari and Dmitriy Fingerman, *Loan Default Prediction and Identification of Interesting Relations between Attributes of Peer-to-Peer Loan Applications.* New Zealand Journal of Computer-Human Interaction ZJCHI. Vol. 2(2), 2017.

[4] K. Ulaga Priya, S. Pushpa, K. Kalaivani and A. Sartiha, *Exploratory Analysis on Prediction of Loan Privilege for Customers using Random forest.* International Journal of Engineering and Technology. Vol. 7(2.21) Pp. 339-341, 2018. https://doi.org/10.14419/ijet.v7i2.21.12399

[5] Li Ying, *Research on Bank Credit Default Prediction Based on Data Mining Algorithm.* The International Journal of Social Sciences and Humanities Invention Vol. 5(6): 4820-4823, 2018.

[6] Y. Xia, L. He, Y. Li, N. Liu and Y. Ding, *Predicting Loan Default in Peer-to-Peer Lending using Narrative Data.* Journal of Forecasting. Pp. 1–21, 2019. https://doi.org/10.1002/for.2625

[7] Y. Long, J. Liu, M. Fang, T. Wang, & W. Jiang, *Prediction of Employee Promotion Based on Personal Basic Features and Post Features.* In Proceedings of the International Conference on Data Processing and Applications (pp. 5-10), 2018.

[8] C. S. Machado and M. Portela, *Age and Opportunities for Promotion.* IZA Discussion Paper No.7784, 2013.

[9] F. D. Blau and J. DeVaro, *New Evidence on Gender Differences in Promotion Rates: An Empirical Analysis of a Sample of New Hires.* Industrial Relations: A Journal of Economy and Society. 46, 3 (July. 2007), 511-550. DOI= https://doi.org/10.1111/j.1468-232x.2007.00479.x.

[10] S. Spilerman and T. Lunde, *Features of Educational Attainment and Job Promotion Prospects.* American Journal of Sociology. 97, 3 (Nov. 1991), 689-720, 1991. DOI= https://doi.org/10.1086/229817

[11] I. E. De Pater, A. E. Van Vianen, M. N. Bechtoldt and U. C. KLEHE, *Employees' Challenging Job Experiences and Supervisors' Evaluations of Promotability.* Personnel Psychology. 62, 2 (Summer 2009), 297-325. DOI= https://doi.org/10.1111/j.1744-6570.2009.01139.x

[12] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi and T. W. Malone, *Evidence for a Collective Intelligence Factor in the Performance of Human Groups.* Science. 330, 6004 (Oct. 2010), 686-688. DOI= https://doi.org/10.1126/science.1193147.

[13] A. Sarker, S. M. Shamim, M. S. Zama, & M. M. Rahman, *Employee's Performance Analysis and Prediction using K-Means Clustering & Decision Tree Algorithm.* Global Journal of Computer Science and Technology, 2018.

[14] S. Saranya and J. S. Devi, *Predicting Employee Attrition using Machine Learning Algorithms and Analyzing Reasons for Attrition.* 2018

[15] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An Introduction to Statistical Learning: with Applications in R.*. Springer Texts in Statistics 103, DOI 10.1007/978-1-4614-7138-7 2017.

[16] R. Punnoose and P. Ajit, *Prediction of Employee Turnover in Organizations using Machine Learning Algorithms.* International Journal of Advanced Research in Artificial Intelligence (IJARAI), Vol. 5, No. 9, 1-5, 2016.

[17] A. Liaw and M. Wiener, *Classification and Regression by Random forest,* R news, 2(3), 18-22, 2002.

[18] L. Breiman, *Random forests. Machine Learning,* 45(1), 5–32, 2001.

[19] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R. Springer textbook*, 2013.

[20] Yoav Freud and E. Robert Schapire, *A Short Introduction to Boosting.* Journal of Japanese Society for Artificial Intelligence. Vol 14(5):771-780, 1999.

[21] G. James et al., *An Introduction to Statistical Learning: with Applications in R,* Springer Texts in Statistics 103, DOI 10.1007/978-1-4614-7138-7 8

[22] Shiruru, Kuldeep, *An Introduction to Artificial Neural Network.* International Journal of Advance Research and Innovative Ideas in Education. 1. 27-30, 2016.

[23] Kumamoto University http://www.cs.kumamoto-u.ac.jp/epslab/ICinPS/Lecture-2.pdf

[24] J. H. Friedman, *Greedy Function Approximation: A Gradient Boosting Machine*, Annals of statistics, 1189-1232, 2001.

[25] N. Alexy and K. Alois, *Gradient Boosting Machines: A Tutorial. Frontiers in Neurorobotics.* Vol 7(21) pp 3, 2013.

[26] S. Lessmann and S. Voß, *A Reference Model for Customer-centric Data Mining with Support Vector Machines*, European Journal of Operational Research 199, 520–530, 2009.

[27] T. Chen and C. Guestrin, *XGBoost: Reliable Large-scale Tree Boosting System*, 2015.

[28] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush & A. Gulin, *CatBoost: Unbiased Boosting with Categorical Features.* In Advances in Neural Information Processing Systems (pp. 6638-6648), 2018.

[29] J. Friedman, T. Hastie and R. Tibshirani, *Additive Logistic Regression: A Statistical View of Boosting.* The Annals of Statistics, 28(2):337–407, 2000.

[30] L. Mason, J. Baxter, P. L. Bartlett and M. R. Frean, *Boosting Algorithms as Gradient Descent.* In Advances in Neural Information Processing Systems, pages 512–518, 2000.

[31] R. Kohavi and F. Provost, *Glossary of Terms. Machine Learning—Special Issue on Applications of Machine Learning and the Knowledge Discovery Process.* Machine Learning, 30, 271-274, 1998.

[32] Microsoft Capstone Project https://www.datasciencecapstone.org/ (Assessed in April 2019)

[33] Kaggle https://www.kaggle.com/c/intercampusai2019 (Assessed in August 2019)

Hindustani or Hindi vs. Urdu: A Computational Approach for the Exploration of Similarities Under Phonetic Aspects

Muhammad Suffian Nizami
Department of Computer Science
FAST National University
Chiniot-Faisalabad, Pakistan

Tafseer Ahmed
Center for Language Computing
Mohammad Ali Jinnah University
Karachi, Pakistan

Muhammad Yaseen Khan
Center for Language Computing
Mohammad Ali Jinnah University
Karachi, Pakistan

***Abstract*—The semantic coexistence is the reason to adopt the language spoken by other people. In such human habitats, different languages share words typically known as loan words which appears not only as of the principal medium of enriching language vocabulary but also for creating influence upon each other for building stronger relationships and forming multilingualism. In this context, the spoken words are usually common but their writing scripts vary or the language may have become a digraphia. In this paper, we presented the similarities and relatedness between Hindi and Urdu (that are mutually intelligible and major languages of Indian sub-continent). In general, the method modifies edit-distance; and works in the fashion that instead of using alphabets from the words it uses articulatory features from the International Phonetic Alphabets (IPA) to get the phonetic edit distance. This paper also shows the results for the languages consonant under the method which quantifies the evidence that the Urdu and Hindi languages are 67.8% similar on average despite the script differences.**

***Keywords*—Lexical Similarity; Urdu; Hindi; Edit Distance; Pho-netics; Natural Language Processing; Computational Linguistics**

I. Introduction

In the Indian sub-continent hundreds of different languages are spoken throughout the area it spans, most of which belong to the Dravidian and Aryan families. It is the accepted fact by linguists that the Aryan family of languages evolved from Sanskrit [1]. Historically, during the medieval period of India, Sanskrit was the language of rulers and of the people from the upper-class, this period also shows the witness for Prakrits and the other languages derived from the Sanskrit [2]. Followed by time, we see the rule of Persian language in Indian courts; and at the near end of Mughal era, Urdu eventually became the official language of the court [2], [3]. Many of the researchers argue that the languages Urdu and Hindi are same because they share the same grammar and a large number of words in their common vocabulary; while in the same context, many other researchers express their findings in refutal. The debate engages the development and origin of the Urdu. A common understanding behind the development of Urdu language shows that it is a creole language which came into being through the mixing of local Indian people and the foreign invaders from the different background and ethnicities [4]; and hence, often referred with the 'camp language'. In contrast, veteran Urdu lexicographer Parekh [5] rejected the theories describing Urdu as a creole language; and maintained that it is the 'Khariboli' of the central zone of India with the exception that if its vocabulary draws words from the Persian and Arabic it would become Urdu, and similarly, if it uses words from Sanskrit it would be Hindi. The Urdu and Hindi are the mutually intelligible languages; however, are the victim of language-split which resulted in the usage of modified Perso-Arabic script called 'Nastaliq' and Devanagari script, respectively, for writing. Amongst many other characteristics of these scripts the two which appear salient are: Nastaliq is a cursive scripted and supposed to be written in right-to-left direction; whereas, Hindi is block-letter and follows left-to-right direction. Figure 1 depicts the languages of the world and their respective sizes (in terms of size of leaves spread), where specifically Urdu and Hindi appear on the top-left, in the Indo-European→Indo-Iranian→Indo-Aryan branch. In the same context, the two languages in the world of today can be combined under the common term 'Hindustani' and also recognized as separate Persianised and Sanskritised registers of the Hindustani language.

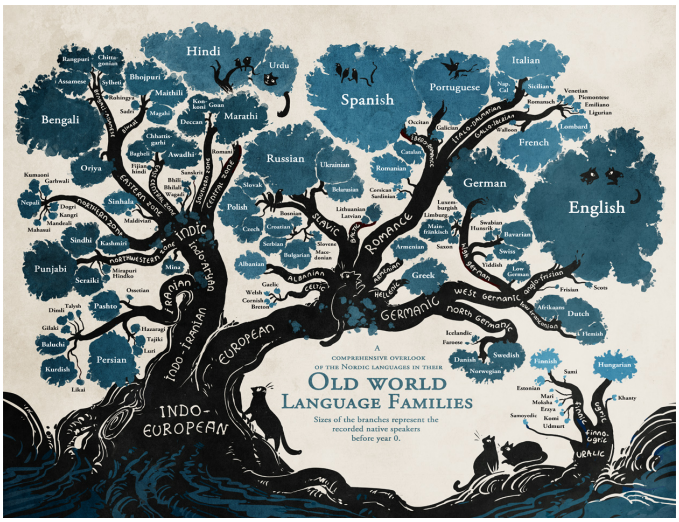


Figure 1: Major languages of the world presented in the format of family tree. Courtesy Minna Sundberg.

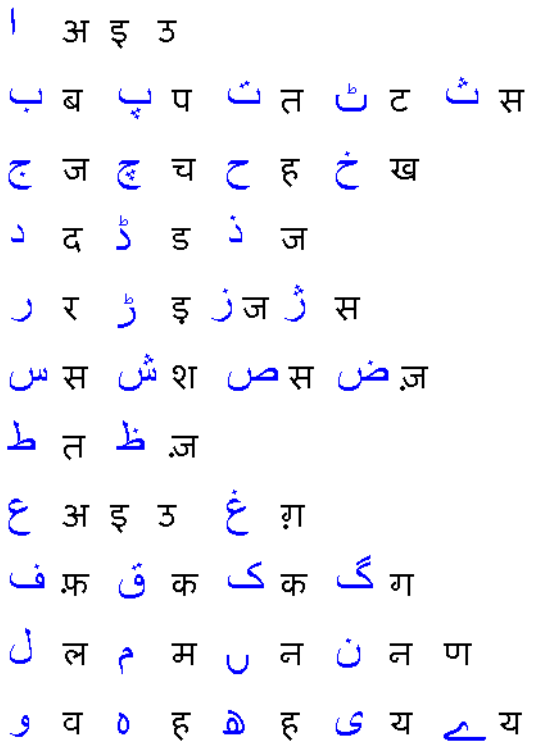


Figure 2: Alphabets of Urdu and Hindi languages in modified Perso-Arabic script (shown in blue colour) and Devanagari script (show in white colour) respectively.

Before we proceed further, another behaviour is noteworthy for which we see that since the division of India happened in 1947, a special focus has been made on official grounds for inducting Persian and Arabic words in Urdu and Sanskrit in Hindi by the right governments and print and electronic media associates of the two countries (i.e., Pakistan and India). Thus, where these languages share a vast vocabulary and morphological structure, their speakers attempt to distinguish them through the word borrowing or with *loan words* from the source languages as mentioned earlier. Hence, it is observable that it may be very difficult for the youth of contemporary time to comprehend the news bulletin announced officially in the pure Urdu and Hindi. However, movies and other channels of entertainment can be accounted for as the principal medium of vocabulary enhancement.

With more than 329.1 million native speakers all around the world, [2] and being a victim of digraphia; the main challenges and point of research investigation—w.r.t the similarities between Hindi and Urdu languages—taken into the consideration for this paper are discussed in the subsequent paragraphs.

The difference in writing script leads the matter not only towards the inabilities of reading but also to the pronunciation. We see that the pronunciation of certain Perso-Arabic alphabets are improper w.r.t the core Hindi speakers such that they are not able to differentiate 'ज', 'ज़', 'ज़', 'ज़', and 'ज़'; as they tend to pronounce 'ज' for all of them. It is seen very often to associate a diacritic symbol, namely *bindi*, in the transformation of 'ज'→'ज़' for differentiating 'ض' and 'ظ' from the rest of aforementioned Urdu alphabets. Similarly, for Urdu alphabets 'گ' and 'غ' they use a single Hindi alphabet'ग' and add bindi in it ('ग़') to substitute it for 'غ'; in a very similar fashion, it uses 'क' and 'क़' for 'ک' and 'ق' respectively. In a contrasting manner, for the two Urdu alphabets such as 'ا' and 'ع' Hindi corresponds with the three alphabets 'अ', इ', and उ'; similarly, for 'ح' and 'ھ', and 'ہ' Hindi has only one alphabet 'ह'; and lastly, for 'ث', 'س', and 'ص' Hindi has got only one alphabet i.e. 'स'; so with the Urdu alphabets 'ے' and 'ی' it has got a single alphabet i.e. 'य'. Collectively all of the alphabet marking are mapped in the figure 2. Thus, with these many-to-many mappings among the alphabets of two scripts, we can easily anticipate the production of very severe semantic mistakes. For example, take the Urdu word 'ذلیل' [ðaˈliːl] (*humiliated*) for which the Hindi may have the chance to pronounce as 'जलील' [dʒaˈliːl] (*exalted*, *magnificent*). Similarly, for the multi-words, the Urdu language has to give an additional space hence a single word would consist of multiple tokens; for example, 'ان پڑھ' [ənˌpəɽʱ] (*illiterate*) which is 'ان' + 'پڑھ', however; Hindi language has no compulsion of giving a white-space in between tokens, so for the given Urdu example 'ان پڑھ' , it will render 'अनपढ़'(pronounced as per same IPA and meant into the same thing).

Thus, in order to find the similarity between the two languages, we are required to transform every cognate as per a similar scheme. For such scheme, Romanized transliteration is the popular way but it undergoes with the same issue i.e. many-to-many alphabet mapping; for example 'ث', 'س', and 'ص' will have only substitute in the Latin script i.e. 'S' *et cetera* [6]. The alternate approach, as used in this paper, is taking the IPA into account for the transformation. In addition to it, this paper presents a modified version of conventional edit distance, namely 'Phonetic Edit Distance' (PED), where the articulatory features of the IPA are employed. This will also help us to see the relatedness of the same word, spoken/pronounced by the people of core Urdu and Hindi backgrounds, at a slight/negligible distance; instead of getting a hard distance through standard edit distance metric (yielded on romanized transliteration). Likewise, Nizami *et al.* [7], we considered to find the similarities w.r.t the Parts-of-Speech (PoS); such that it would be more interesting to find the right cognate of *book* as a noun in the list of nouns of the other language, rather the make a generalized look-up on all possible words.

The rest of the paper organization is as The literature review about the lexical similarity of languages and earlier techniques is in Section II, the methodology is described in Section III, the detailed results are shared in Section IV, followed by the conclusion and future works in Section V, and bibliographical references in the end.

II. Literature Review

In this section, the literature review, the existing techniques, and approaches for lexical similarity concerning script and sound are described.

The two most frequent method for resolving the problem of this kind are string matching algorithms and employment of the Soundex algorithm. The edit distance algorithm has different variants for different types of tasks like string alignment and

spells correction in language processing [8]. The problem is that it takes the characters or letters as distinct units, in such cases if the characters are completely similar then no operation needed. It depends on the user that it may use different weights for the operations. There is another algorithm known as Soundex which works on sound-based matching instead of letter or spell matching [9].

Jinugu [10] presented which is the variant of the Tarhio-Ukkonen algorithm [11] for maximizing the matching of a string by finding the longest patterns in the string and ignores the mismatches of characters. This algorithm works in multiple shifts on the variant lengths of strings for matching purpose, the shift distance and number of characters involved in matching also matters for its performance.

The work [12] shows usage of the Soundex algorithm for retrieving noun words from the database consisting of vowels and consonants for the Hindi language. Likewise, the Soundex algorithm provides classes for letters as their agreement classes which are six in number, where the vowels are eliminated and only consonants are changed into their relevant phonetic class [13]. Other similar work is by [14] and [15] which is phonetic matching using rule-based algorithms and utilizing encoding scheme for homophone words matching scripted in different languages.

The Soundex algorithm considers many letters due to the articulation of sound. The IPA chart is present at the website[1]. The IPA's are ordered according to the manner and place of articulation. Few letters have same articulation i.e. *plosive*, *bilabial* [16]. Similarly few letters are *voiced* and *unvoiced* consonant [17], *aspirated* and *pharyngeal* etc. The set of features can represent IPA symbols. In Germanic languages, according to [18] and [19], there are some voiced-stops that can change into voiceless stops and vice versa.

There are different studies found on the lexical similarity like [20] worked on the dialectal differences among the pair of texts using cosine similarity, Hamming distance, and Levenshtein distance and [21] worked on cognates identification among different languages based on inter-related vocabulary. It shows that the lexical similarity can be computed by using the phonetic level features of words rather than orthographic features.

Another work is done for the similarity of words on a limited PoS by using synsets in WordNet, to extend this lexical similarity on phrase level and sentence level computed with the help of word-level similarity [22]. Similarly, the lexical similarity computed for the source code by using string level matching [23]. Some other researchers find multiple dimensions for lexical similarity like knowledge-based, string-based, and corpus-based [24]. An experiment conducted by [25] on cross-language similarity for the cognates of Dutch and English, on similar grounds [26] lexical similarity was computed by using phonetics based cognates with high frequencies.

As the researchers in [25] and [26] showed that in cross-language cognates and loan words similarity matters phonologically. Similarly, the historic background and origin of these languages are analyzed like [27] did for Urdu and French words. Another work in support of phonological level similarity of languages was done on English word structures as a network of language where links were made between words phonologically [28].

III. Methodology

In this section, the main components are described as: The detailed discussion about the languages chosen for the experiment, the source of the dataset, and the specific PoS word lists for lexical similarity is done in section III-A. Discussion about the standard edit distance and the proposed phonetic modifications based on articulatory features is described in section III-B. The proposed method modified phonetic edit distance is explained in section III-C. The detailed discussion for the computation of lexical similarity for the chosen languages is given in section III-D.

The basic task of lexical similarity calculated on similar words ratio or count in between two languages. But, there are few reservations like:

1) How it will be inferred that two words are similar?
2) To which extent comparison should be done and which criteria should follow to choose the words?

The answer to the raised questions is that there should be a dynamic method to decide whether two words are cognates or not, the edit distance should be employed as a measure of similarity, the comparison should be on the equal words or some acceptable count of words and lastly, the criteria is also regarding the origin and source of words picked for the lexical comparison. To explain the first part of the answer we need to propose a modified edit distance method for computing lexical similarity which is explained in the coming part as phonetic edit distance. The next parts of the answers are related to word lists and their specific feature or aspect for selecting comparison. For this, we have chosen different parts-of-speech (PoS). This decision is made due to the importance of PoS as these are the rich and main content of any language, also some previous similar work of lexical similarity was done by using PoS tag set [7]. For the complete pictorial view of the proposed phonetic edit distance method, the system diagram is given in figure 3.

A. Dataset

We used the Universal Dependency[2] (UD) corpora for extracting PoS word lists. UD has some standard data about all languages in a standard format. Another reason to choose the only PoS for similarity is that in the textual corpora each language can be divided into PoS tag set. In this experiment, we have chosen Urdu and Hindi with majorly two scripts Devanagari and Perso-Arabic. Also, the conversion system for these scripts to IPA is developed. The part-of-speech (PoS) used for the similarity purpose are verbs, proper nouns, nouns, particles, auxiliary, pronoun, coordinating and subordinating conjunctions, and adposition. The length of each PoS tag word list is shown in Figure 4 comparing the size of both Hindi and Urdu languages.

[1] https://www.internationalphoneticassociation.org/content/full-ipa-chart

[2] https://universaldependencies.org/

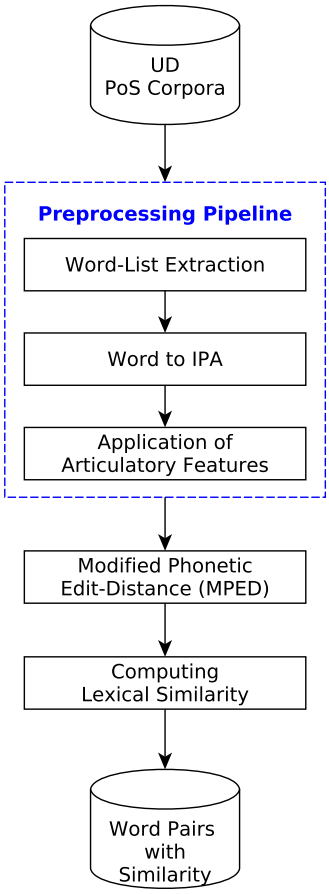


Figure 3: Phonetic Edit Distance based Lexical Similarity System Diagram.

B. Understanding the Phonetic Edit Distance

The standard edit distance [29] takes two words or strings and returns the distance between them. In this process, the internal mechanism of the edit distance method is based on the insertion, deletion, and update operations which compute the cost for two strings as a distance. Each operation is given a unit cost which aggregates during the comparison of two strings. This is simple string matching which doesn't provide any information about the sound-related features like phonetic articulatory features. In our proposed method the edit distance is modified based on these articulatory features and called here as the future of edit distance as phonetic edit distance.

The proposed Phonetic Edit Distance (PED) works the same as standard edit distance but its internal mechanism is based on phonetic features, it is explained in section III-C. It takes IPA encoded two words and then returns the phonetic distance between them. If the sound of both words is the same then the phonetic distance will be zero. But, if the words are not similar then the insertion and deletion operation computes cost as well as the phonetic cost of the words also aggregates to total in case of mismatch of sound.

If we take the standard edit distance, the distance of two IPA strings /bæd/ and /pæd/ is Φ(/bæd/, /pæd/) = 1, these IPA strings represent English words *bad* and *pad* respectively;

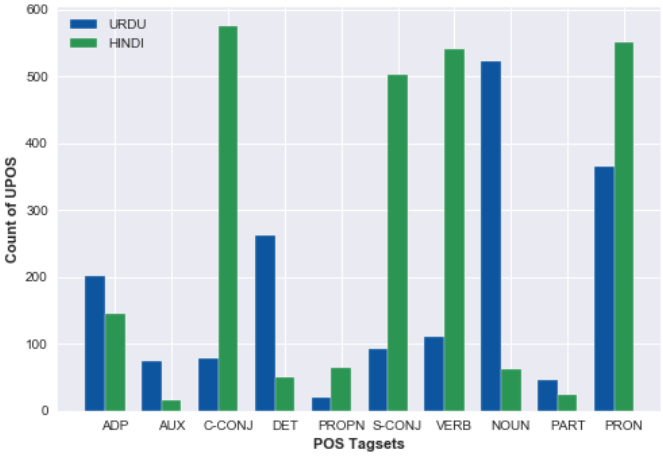


Figure 4: word-list size PoS-wise for Hindi and Urdu.

and Φ is the edit distance. There is only one replace operation of 'b' with 'p' to make the similar string but, with our proposed PED method, by using the same pair of IPA encoded strings is 0.2, in this case, the cost of replacement operation of 'b' with 'p' is 0.2 as the sound of both letters make less difference and take place near in the articulatory feature-based IPA chart, and thus phonetic similarity of the 'bad' and 'pad' is much lesser than the standard ED.

The lexical similarity in true sense is by calculating for both types of words. But vowels in different languages contain add-on features like in Urdu vowels contain short vowels as their composing part [30]. In our paper, we are skipping the vowels as future work due to their complex nature of features for phonetics. Thus, for *consonants*, we have proposed the following features (and their respective values) *voiced* (binary), *airflow* (discrete), *place* (continuous), *aspirated* (binary), *pharyngeal* (binary) and *manner* (discrete).

The features are picked from their positions, the *place* is the articulation of place and inside the human mouth, these places are present. we have assigned the value as per their feature positions. Like lips (bilabial) position have 0.05, teeth position have 0.15, and the throat (glottal) position has 0.95. The other two important features are *type* and *label*, the *label* is the IPA of sound and *type* ensures that sound is a vowel or a consonant.

C. Algorithm for Phonetic Edit Distance

As the sound based articulatory features presented in section III-B and represented with their corresponding values. In this work, we didn't compare the vowels with consonants but only consonants with consonants.

For the comparison of consonants, as shown in algorithm: 1, we have given $^2/_3$ value to the *place* and *manner* features; and $^1/_3$ value is assigned to the remaining features. The *voiced* feature has $^1/_5$ value; and the other features have the remaining weight/value shown as $1 - {}^2/_3 - {}^1/_5$. At present, in the proposed system, the remaining features are *airflow* and *aspirated*. Although, we can increase the features without decreasing the weight of major features.

Further, *manner* and *place* features are more substantial. So, we have considered the distance only when *manner* and *place* represented as (δ_{m+p}) is equal or lesser than the threshold which is $^1/_2$. If the joint distance is above the threshold level then we don't add the distance of other features and return only this distance. Also, rule-based distance for feature *manner* is done by using dictionary look-up when the key of $\langle a_{\text{manner}}, b_{\text{manner}}\rangle$ is given. Finally, a_D,b_D, shows the Manhattan Distance for the remaining all features as presented in-line 10.

Algorithm 1 Phonetic Difference for Consonant (PDC)

1: **procedure** PDC(a, b) ▷ a and b are strings, where $a \neq b$.
2: $\delta_{\text{manner}} \leftarrow \Theta[\langle a_{\text{manner}}, b_{\text{manner}}\rangle]$ ▷ Calculating Manhattan Distance between a and b using open and back features.
3: $\delta_{\text{placed}} \leftarrow \Theta(\langle a_{\text{placed}}, b_{\text{placed}}\rangle)$
4: **if** $\delta_{\text{manner+placed}}$ >threshold **then**
5: **return** $\delta_{\text{manner+placed}}$
6: **else**
7: $\delta_{\text{manner+placed}} \leftarrow \delta_{\text{manner+placed}} \cdot \frac{2}{3}$
8: $\delta_{\text{voiced}} \leftarrow \Theta(\langle a_{\text{voiced}}, b_{\text{voiced}}\rangle) \cdot \frac{1}{5}$
9: $\delta_{\text{remain\_features}} \leftarrow \Theta(\langle a_D, b_D\rangle) \cdot \beta$
10: **return** $\delta_{\text{manner+placed}} + \delta_{\text{voiced}} + \delta_{\text{remain\_features}}$

Algorithm 2 Modified Phonetic Edit Distance (MPED)

1: **procedure** MPED($\mathbf{X}, \mathbf{Y}$) ▷ X and Y are the IPA encoded strings
2: $x \leftarrow$ length of $\mathbf{X}$.
3: $y \leftarrow$ be the length of $\mathbf{Y}$.
4: **if** $\min(x, y) = 0$ **then**
5: **return** $\max(x, y)$
6: **if** $\mathbf{X}[x-1] = \mathbf{Y}[y-1]$ **then**
7: cost= 0
8: **else**
9: ins_cost $\leftarrow$ ED($\mathbf{X}[0 : x-1], \mathbf{Y}$) + 1
10: del_cost $\leftarrow$ ED($\mathbf{X}, \mathbf{Y}[0 : y-1]$) + 1
11: rep_cost $\leftarrow$ ED($\mathbf{X}[0 : x-1], \mathbf{Y}[0 : y-1]$) + PDC($\mathbf{X}[x-1], \mathbf{Y}[y-1]$)
12: **return** $\min$(ins_cost, del_cost, rep_cost)

In Algorithm 1, PDC is the phonetic difference of consonants, Θ is computing unit for two features, MPED is the modified phonetic edit distance and ED is the edit distance. The pseudo-code of overall lexical similarity for experimental languages is described in Algorithm 3, in which on finding similarity the result is in the range of (0,1) if the sound is the same then 0 otherwise 1. This way all PoS words from the Urdu language compared with Hindi language using this modified phonetic edit distance to find the lexical similarity and the ratio of loan words or cognates between the languages.

A brief example. If we take the example of Urdu word 'صاحب' and its standard IPA [sɑː.ɦɪb][3] (*sir* or *mister*) we may have many similar words/cognates in Hindi which are pronounced as gv[s̪äː.ɦɪb] , [sɑː.ɦəb] , [s̪äː.ɦəb] , [sɑːɦb] , [s̪äːɦb] , [sɑːb] , [s̪äːb][4]. To sense, the difference between

[3]https://en.wiktionary.org/wiki/صاحب

[4]https://en.wiktionary.org/wiki/साहिब#Hindi

Table I: Mapping of articulatory features for the Urdu word 'صاحب' [sɑː.ɦɪb].

	Meta Features	IPA letters				
	Label →	s	ɑː	ɦ	ɪ	b
	Type →	c	v	c	v	c
Phonetic Articulatory Features	Method	0	-	0	-	0
	Place	.45	-	.95	-	0.15
	Manner	fr	-	nsfr	-	pl
	Voice	0	-	0	-	0
	Aspirated	0	-	0	-	0
	Open	NA	-	NA	-	NA
	Back	NA	-	NA	-	NA
	Rounded	NA	-	NA	-	NA

Table II: Mapping of articulatory features for the Hindi word 'साहिब '[s̪äː.ɦəb] .

	Meta Features	IPA letters				
	Label →	s̪	äː	ɦ	ə	b
	Type →	c	v	c	v	c
Phonetic Articulatory Features	Method	0	-	0	-	0
	Place	.35	-	.95	-	0.15
	Manner	fr	-	nsfr	-	pl
	Voice	0	-	0	-	0
	Aspirated	0	-	0	-	0
	Open	NA	-	NA	-	NA
	Back	NA	-	NA	-	NA
	Rounded	NA	-	NA	-	NA

Table III: Comparison of standard edit distance and proposed method.

Source & Target Words	Standard Edit Distance	Proposed Phonetic Edit Distance
[sɑː.ɦɪb] vs. [s̪äː.ɦɪb]	2	.1
[sɑː.ɦɪb] vs. [sɑː.ɦəb]	1	0
[sɑː.ɦɪb] vs. [s̪äː.ɦəb]	3	.1
[sɑː.ɦɪb] vs. [sɑːɦb]	1	0
[sɑː.ɦɪb] vs. [s̪äːɦb]	4	.1
[sɑː.ɦɪb] vs. [sɑːb]	3	.1
[sɑː.ɦɪb] vs. [s̪äːb]	5	1.05

the IPA letter /sɑ/ and /s̪/ we can substitute/suppose Urdu alphabets ص and س respectively; and for the analogy of the romanized variant, both of them would be producing sound for *s* but former one has low whistle sound in comparison to the later one, where whistle sound is bit higher due to dental place. Thus, with the romanized equivalents of these (Hindi) words we can get a higher distance through the standard edit distances (see table III); whereas, with the proposed model the PED will give results very closely. Further for reference, the tables I and II show the articulatory features of Urdu word 'صاحب' [sɑː.ɦɪb] and Hindi 'सहब ' [s̪ɑː.ɦəb], where in the table the type *c* indicates the alphabet is consonant otherwise it vowel (for which the PED is not working); similarly, NA shows the feature is not applicable on the very alphabet.

D. Computing Lexical Similarity

The flow of finding lexical similarity is described in this section as; The word lists created from UD corpora, then these word lists converted into respective IPA codes, after this IPA codes enriched with phonetic articulatory features and in last the lexical similarity computed based on these phonetic features between the languages using the proposed method and the pseudo-code is presented in Algorithm 3.

Universal Dependencies website[5] provide the corpora for all languages in CoNLL-U format with tagged PoS. The tagged

[5]https://universaldependencies.org/u/pos/index.html

dependency structure includes word lemma and the Universal Parts-of-Speech (UPoS). We used lemma in computation rather than words, for which there exists an inflectional nature.

After extraction of word-lists, we converted word-lists into IPA strings. There are many online platforms and dictionaries which converts words into respective IPA strings. Keeping in mind that the chosen languages hold short-vowels, izafat-letters, and diacritics causing issues of conversion for such platforms [2]. Underlying this, we have created our mapper of words-to-IPA (in fact script-to-IPA) for both languages.

Further, the articulatory features from the IPA phonetic chart used for a one-to-one mapping of words. The IPA chart is the standard chart for phonetic level weight-age of letters in any word. Based on these features, each word is compared with other words. These articulatory features are described in section III-B and III-C.

Algorithm 3 Computing Lexical Similarity between two Languages.

1: **procedure** LEXSIM($\mathbf{Lang_1}, \mathbf{Lang_2}$) ▷ $\mathbf{Lang_1}$ and $\mathbf{Lang_2}$ are the list of words.
2: $\Phi_{\text{Tot}} \leftarrow 0$
3: **for** every word a in $\mathbf{Lang_1}$ **do**
4: **for** every word b in $\mathbf{Lang_2}$ **do**
5: $\text{x} \leftarrow \text{PED}(a, b) \cdot \frac{1}{\max(\|a\|, \|b\|)}$
6: [Tot] ← least value as key.
7: $_{\text{Tot}} \leftarrow _{\text{Tot}} + \text{x}$
8: $\beta \leftarrow \frac{_{\text{Tot}}}{\|\mathbf{Lang_1}\|}$
9: **return** β

Finally, the lexical similarity (LS) for Urdu and Hindi is computed on all PoS. Let's take the word-lists of $\mathbf{Lang_1}$ and $\mathbf{Lang_2}$; $\mathbf{Lang_1} \neq \mathbf{Lang_2}$ including the same PoS lists. In Algorithm 3 for the comparison of languages $\mathbf{Lang_1}$ and $\mathbf{Lang_2}$, supposing words a of $\mathbf{Lang_1}$ are compared with all words b of $\mathbf{Lang_2}$ in the step-3 and step-4. Also, we have normalized the PED result value with a maximum length of the word for the comparative words in step-5. Otherwise, the smaller words will be getting less value of phonetic distance.

Here in algorithm 3, every word 'a' compared with 'b' and the minimum value recorded in edit distance, this aggregates the overall edit distance; and β is the average distance per letter for both lists. if the value of β is equal to zero '0' then both lists (languages) are similar as identical but if the value is more closer to '1' then words are different in sounds and vice versa.

IV. Results

We have computed the lexical similarity for Urdu and Hindi with the proposed method on articulatory features PoS-wise. The results are shown in Figure 5. It is identified that these languages have cognates and genetic affinity. Urdu and Hindi are quite similar in spoken level but Hindi is written in Devanagari script which is entirely different from Urdu script. The similarity index shows that the Hindi and Urdu are top similar in auxiliaries, determiners, articles, coordinating conjunctions, and pronouns PoS.

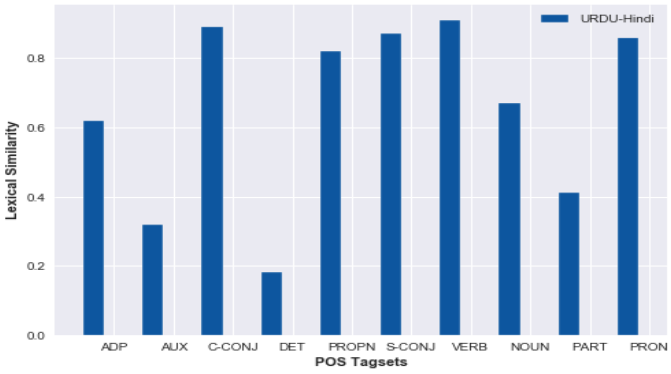


Figure 5: Lexical Similarity PoS-Wise Hindi-Urdu.

It is analyzed from our experiment that Urdu and Hindi on average 67.8% similar languages despite different scripts. This average similarity is computed from the similarity of each PoS employed in this study. In all PoS-wise comparisons, most of the results are comprehensible the only determiner is less similar as shown in Figure 5. Similarly, few PoS have shown high similarity that is due to the unbalanced size of words in those PoS like coordinating conjunction, subordinating conjunction, and verbs. The point which is important to discuss here is the low similarity in adpositions and determiners; this is due to the erroneous tagging of Hindi and Urdu PoS tags in UD dataset. Thus, if specifically the tagging for adposition and determiner is done properly, then, the results would support the fact with more emphasize that the Urdu and Hindi are the same languages though they are the victim of language split.

V. Conclusion

In this paper, we have introduced a modified algorithm for the lexical similarity of Urdu and Hindi languages based on articulatory features. This algorithm has identified the intelligibility, cognates, and borrowed words despite the spelling, script, and phonetic difference. In the conducted experiment with the proposed algorithm, the majority of similarity pairs of PoS are in agreement as per their genetic affinity. The proposed algorithm has given better and understandable results which are far better than the simple string matching with standard edit distance on such a phonetic level parameter. The proposed method is also found effective under the situation where a speaker does not qualify or unable to pronounce a certain alphabet of other languages (for example Arabs cannot pronounce the sound 'p' and 'ch'); so for these situations, they have to look into the similar or near-by sounds for substitution. In such scenarios, PED will give minute results edit distance in comparison to standard edit distance.

The $\approx$ 67.8% similarity is fair enough to stay positive on the question that whether Urdu and Hindi languages are mutually intelligible or not? Since the similarity under the phonetic aspect is high, therefore, we maintain that, within the context of the speech, it is very rightful to term both of these languages as 'Hindustani'; however, the difference of script may produce a very trivial excuse to differentiate either one of them as 'Hindi' or 'Urdu.'

Though there are some limitations in the used UD corpora (variation in the size of languages, format errors, and basic processing) and the issues itself in the languages like silent letters, diacritics, and short-long vowels. It could be improved by using digital lexicographic resources and dictionaries rather than letters to the IPA scheme. In the future, a comprehensive work could be done for the lexical similarity of the whole family of Indo-Aryan languages by extending and enriching the proposed algorithm with vowels along with consonants.

References

[1] A. K. Dutt, C. C. Khan, and C. Sangwan, "Spatial pattern of languages in india: A culture-historical analysis," *GeoJournal*, vol. 10, no. 1, pp. 51–74, 1985.

[2] M. Y. Khan, M. A. Rao, S. Wasi, T. A. Minai, and S. M. K.-u.-R. Raazi, "Edit distance-based search approach for retrieving element-wiseprosody/rhymes in hindi-urdu poetry," *Indian Journal of Science and Technology*, vol. 13, no. 39, pp. 4189–4201, 2020.

[3] M. Y. Khan and M. S. Nizami, "Urdu sentiment corpus (v1. 0): Linguistic exploration and visualization of labeled dataset for urdu sentiment analysis," in *2020 International Conference on Information Science and Communication Technology (ICISCT)*. IEEE, 2020, pp. 1–15.

[4] M. Y. Khan, S. M. Emaduddin, and K. N. Junejo, "Harnessing english sentiment lexicons for polarity detection in urdu tweets: A baseline approach," in *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*. IEEE, 2017, pp. 242–249.

[5] R. Parekh, "Urdu's origin: it's not a 'camp language," *Dawn*, Dec 2011. [Online]. Available: https://www.dawn.com/news/681263/urdus-origin-its-not-a-camp-language

[6] M. Y. Khan and T. Ahmed, "Pseudo transfer learning by exploiting monolingual corpus: An experiment on roman urdu transliteration," in *International Conference on Intelligent Technologies and Applications*. Springer, 2019, pp. 422–431.

[7] M. S. Nizami, M. Y. Khan, and T. Ahmed, "Towards a generic approach for pos-tagwise lexical similarity of languages," in *International Conference on Intelligent Technologies and Applications*. Springer, 2019, pp. 493–501.

[8] M. Smith, K. T. Cunningham, and K. L. Haley, "Automating error frequency analysis via the phonemic edit distance ratio," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 6, pp. 1719–1723, 2019.

[9] A. J. Lait and B. Randell, "An assessment of name matching algorithms," *Technical Report Series-University of Newcastle Upon Tyne Computing Science*, 1996.

[10] J. U. Rekha, "Approximate multiple string matching algorithm," *Journal of Theoretical and Applied Information Technology*, vol. 98, no. 11, 2020.

[11] S. Mojsilovic and A. Ukkonen, "Relative distance comparisons with confidence judgements," in *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, 2019, pp. 459–467.

[12] V. Gautam, A. Pipal, and M. Arora, "Soundex algorithm revisited for indian language," in *International Conference on Innovative Computing and Communications*. Springer, 2019, pp. 47–55.

[13] L. Xu and J. Chamberlain, "Cipher: A prototype game-with-a-purpose for detecting errors in text," in *Workshop on Games and Natural Language Processing*, 2020, pp. 17–25.

[14] I. Mustafin, M.-C. Frunza, and J. Lee, "Multilingual entity matching," in *International Conference on Advanced Information Networking and Applications*. Springer, 2019, pp. 810–820.

[15] M. Llompart Garcia, "Bridging the gap between phonetic abilities and the lexicon in second language learning," Ph.D. dissertation, lmu, 2019.

[16] R. Janssen, S. R. Moisik, and D. Dediu, "The effects of larynx height on vowel production are mitigated by the active control of articulators," *Journal of Phonetics*, vol. 74, pp. 1–17, 2019.

[17] K. R. Kluender, C. E. Stilp, and F. L. Lucas, "Long-standing problems in speech perception dissolve within an information-theoretic perspective," *Attention, Perception, & Psychophysics*, vol. 81, no. 4, pp. 861–883, 2019.

[18] J. Howard, ""pig" or "fig?": Grimm's law, phonemic difference, and linguistic agency in alice's adventures in wonderland," *The Explicator*, vol. 78, no. 1, pp. 41–43, 2020.

[19] A. V. Botsman and O. V. Dmytruk, "Germanic preterite-present verbs and their morphological and semantic peculiarities," *Current issues of Ukrainian linguistics: theory and practice*, no. 39, pp. 74–88, 2019.

[20] J. Nerbonne, W. Heeringa, J. Prokic, and M. Wieling, "Dialectology for computational linguists," 2019.

[21] D. Kanojia, M. Kulkarni, P. Bhattacharyya, and G. Haffari, "Cognate identification to improve phylogenetic trees for indian languages," in *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, 2019, pp. 297–300.

[22] T. Vakare, K. Verma, and V. Jain, "Sentence semantic similarity using dependency parsing," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2019, pp. 1–4.

[23] H. Kaur and R. Maini, "Assessing lexical similarity between short sentences of source code based on granularity," *International Journal of Information Technology*, vol. 11, no. 3, pp. 599–614, 2019.

[24] W. H. Gomaa, "A multi-layer system for semantic relatedness evaluation," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 23, 2019.

[25] H. Carrasco-Ortiz, M. Amengual, and S. T. Gries, "Cross-language effects of phonological and orthographic similarity in cognate word recognition: the role of language dominance," *Linguistic Approaches to Bilingualism*, 2019.

[26] E. Lefever, S. Labat, and P. Singh, "Identifying cognates in english-dutch and french-dutch by means of orthographic information and cross-lingual word embeddings," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 4096–4101.

[27] A. A. Khan, "Lexical affinities between urdu and french," *Eurasian Journal of Humanities Vol*, vol. 1, no. 1, 2015.

[28] C. S. Siew and M. S. Vitevitch, "The phonographic language network: Using network science to investigate the phonological and orthographic similarity structure of language." *Journal of Experimental Psychology: General*, vol. 148, no. 3, p. 475, 2019.

[29] G. Cormode and S. Muthukrishnan, "The string edit distance matching problem with moves," *ACM Transactions on Algorithms (TALG)*, vol. 3, no. 1, pp. 1–19, 2007.

[30] M. Kamran Malik, T. Ahmed, S. Sulger, T. Bögel, A. Gulzar, G. Raza, S. Hussain, and M. Butt, "Transliterating urdu for a broad-coverage urdu/hindi lfg grammar," in *LREC 2010, Seventh International Conference on Language Resources and Evaluation*, 2010, pp. 2921–2927.

A Novel Band Selection Approach for Hyperspectral Image Classification using the Kolmogorov Variational Distance

Mohammed LAHLIMI[1], Mounir Ait KERROUM[2] and Youssef FAKHRI[3]
Laboratory of Research in Computer Science, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco

***Abstract*—In this paper, we introduce a novel band selection approach based on the Kolmogorov Variational Distance (KoVD) for Hyperspectral image classification. The main reason we are taking interest in KoVD is its unique relation to the classification error. Our previous works on band selection using the Mutual Information (MI), the Divergence Distance (DD), or the Bhattacharyya Distance (BD) inspire this study; thus, we are particularly interested in finding out how KoVD performs against these distances in terms of the numbers of band retained and the classification accuracy. All the distances in this study are modeled with the Gaussian Mixture Model (GMM) using the Bayes Information Criterion (BIC) / Robust Expectation-Maximization (REM). The experiments are carried on four benchmark Hyperspectral images: Kennedy Space Center, Salinas, Botswana, and Indian Pines (92AV3C). The results show that band selection based on the Kolmogorov Variational Distance performs better than BD and DD, meanwhile against MI the results were too close.**

Keywords*—Band selection; Bayes Information Criterion (BIC); Bhattacharyya Distance; divergence distance; hyperspectral imaging; Kolmogorov Variation Distance; Gaussian Mixture Model (GMM); Robust Expectation Maximization (REM); remote sensing

I. Introduction

In hyperspectral imaging, sensors record data from hundreds of contiguous bands of the electromagnetic spectrum. However, the Hughes phenomenon [1] [2] [3] and computational complexity [4] are two problems that appear during the classification process. Due to the small sample size problem [5] and the large number of bands acquired from the sensors, the classifier won't be properly trained [3]. Therefore, dimensionality reduction is needed.

Two approaches for dimensionality reduction can be found in literature, band selection [6] [3] [7] [8] and band extraction [9] [10] [11] [12]. The aim for band extraction is to create a new reduced dataset from the existing one using a linear/non-linear transformation [6]. The Principal Component Analysis, Projection Pursuit, Independent Component Analysis, Orthogonal Subspace Projection, Segmented Principal Component Analysis, and others [13] [14] have been used to reduce the data volume. However, due to the linear/non-linear transformation, the original data are replaced by a new set of variables with no actual physical meaning [6] which can be a disadvantage in some applications. Band selection, on the other hand, tries to find an optimal subset from the original pool by only selecting relevant bands with valuable information for the classifier, through maximizing a class separability criterion [6]. Between band extraction and band selection, in this study, the later is the preferred one since, with band selection, the data remains unchanged and the physical meaning is preserved [15].

Band selection techniques can be broadly classified into two categories: wrapper and filter techniques. The wrapper approach [6] take advantage of the classifier itself and use it as the criterion for band selection [16], the result is a subset with a high classification score, however, the drawback of this technique is that the bias toward the used classifier. Unlike the wrapper approach, the filter [6][16] deploy metrics and distances to evaluate the bands without involving the classifier. In theory, the best criterion to measure the pertinence of a band is the Bayes error. However, the calculation of the Bayes error is, in general, a very complex problem [17]. Therefore, some approach seeks an upper bound of the error probability such as the Chernoff and Bhattacharyya bounds.

A new band selection approach is introduced in this paper, based on the Kolmogorov Variational Distance for Hyperspectral image classification. This work is a sequel on our previous research on band selection with Mutual Information [18], Bhattacharyya Distance [8] and Divergence Distance [19]. The primary interest in KoVD is the fact that is uniquely related to the classification error [20] [6], which is often difficult to estimate [17]. KoVD has been used in other fields such as signal selection, communication and radar systems [20] [21] but not in the hyperspectral imaging context.

To model the Kolmogorov Variational Distance, the Gaussian Mixture Model is used with The Expectation-Maximization (EM) algorithm [9], however, with the EM algorithm we face two issues: The first one is the choice of the number of components K as it can affect the estimation of the covariance matrix [8] and the second issue is the sensitivity to the initial values choice [22]. With a bad choice of K, we can easily end up with the Curse of Dimensionality. As a solution two approaches are proposed; a GMM based on the Bayes Information Criterion (BIC) and a Robust Expectation-Maximization (REM) algorithm [22].

Our main contributions in this study is a novel band selection approach with the Kolmogorov Variational Distance modeled with GMM-REM and GMM-BIC. To assess the performances of KoVD two criteria are being used: the numbers of the retained band and the classification accuracy. The experiments are performed on four hyperspectral benchmark datasets: The scene Indian Pines (92AV3C), Botswana scene, Kennedy space center scene, and Salinas scene.

This paper is structured as follows: Sections II and III

describe the fundamentals and the proposed band selection algorithm. Section IV discusses the experimental results. Finally the conclusion in Section V.

II. Background

A. Kolmogorov Variational Distance

The Kolmogorov Variational Distance (KoVD) is the integral of the absolute difference between two posterior probabilities. It expresses the distance between the densities [6]. The main advantage for KoVD is its direct relation to the classification error [20] [6]. KoVD is expressed as follows [6]:

$$J_{KoVD}(\omega_1, \omega_2) = \int |P(\omega_1|x) - P(\omega_2|x)|P(x)\mathrm{d}x \quad (1)$$

KoVD provides an indication of the amount of probability mass by which the two distributions differ. If the classes ω_1 and ω_2 are similar, $P(\omega_1|x) = P(\omega_2|x)$ then J_{KoVD} will equal zero, and if the classes ω_1 and ω_2 are disjoint $P(\omega_1|x) = 0$ and $P(\omega_1|x) \neq 0$, J_{KoVD} will attains its maximum value [6].

In the case of multi-class problem, between each pairwise class (ω_i, ω_j), KoVD is computed as the average cost function.

$$J = \sum_i \sum_j P(\omega_i)P(\omega_j)J_{KoVD}(\omega_i, \omega_j) \quad (2)$$

B. Mutual Information

Given X and Y, two discrete random variables, the Mutual Information (MI) is the defined as [18]:

$$I(X;Y) = H(X)\text{–}H(X/Y) \quad (3)$$

$I(X;Y)$ expresses the information we gain by decreasing the uncertainty contained in the random variable X after knowing Y. With The entropy $H(X)$ of a random variable X and $H(X/Y)$ the conditional entropy of X given Y [18] [23].

C. Divergence Distance

The divergence distance (DD) [19] is a probabilistic distance that measure of the similarity between two classes ω_1 and ω_2 often used in information theory. DD is the sum of the two Kullback-Leibler divergences. Given $P(x|\omega_1)$ and $P(x|\omega_2)$, DD is defined as [6]:

$$J_{DD}(\omega_1, \omega_2) = \int [p(x|\omega_1) - p(x|\omega_2)] \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} \mathrm{d}x \quad (4)$$

DD distance is interpreted as the amount of information necessary to change the prior probability distribution into posterior probability distribution [24]. In the case of multi-class problem, between each pairwise class (ω_i, ω_j), DD is computed as the average cost function according to equation (2).

D. Bhattacharyya Distance

The Bhattacharyya distance (BD) [8] is a similarity measurement of the scatter degree of two classes ω_1 and ω_2. The bhattacharyya distance is expressed as [6]:

$$J_{BD}(\omega_1, \omega_2) = -\log \int (p(x|\omega_1)p(x|\omega_2))^{\frac{1}{2}} \mathrm{d}x \quad (5)$$

In the case of multi-class problem, between each pairwise class (ω_i, ω_j), BD is computed as the average cost function according to equation (2).

E. Gaussian Mixture Model

The Gaussian Mixture Model (GMM) models the density as the sum of one or more weighted Gaussian components [25] [8]. For a GMM, a probability density function is the sum of K Gaussian components:

$$p(x|\omega) = \sum_{k=1}^{K} \pi_k p(x|\mu_k, \Sigma_k) \quad (6)$$

where K the number of mixture component, π_k the mixing weight ($0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$) and $p(x|\mu_k, \Sigma_k)$ a d-dimensional gaussian distribution

$$p(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right] \quad (7)$$

With μ_k the mean and Σ_k the covariance matrix for the k^{th} component.
The parameters $\{\pi c, \mu_c, \Sigma_c\}$ are usually estimated by the EM algorithm [9].

III. Band Selection by Kolmogorov Variational Distance

Given a set of band $F = \{b_i\}_{i=1}^{d}$, the goal is to find an optimal subset $S = \{b'_i\}_{i=1}^{d'}, S \subset F, d' \leq d$ that only keeps the relevant bands that contribute to the classification task while discarding any redundancy. An exhaustive search for the optimal subset S can be impractical from a computational viewpoint, and the Sequential forward selection (SFS) is one of the simplest search strategy [26] [18]. With an empty set of bands S at the beginning, we start to add sequentially the bands that maximizes the KoVD cost function until the desired number of band is achieved, or no longer maximize the cost-function. SFS algorithm have a relatively low computational burden [27].

The algorithm (Fig. 1) is the same as [18] [28] [29] except the computation of the Mutual Information as a cost function between multiple variables. Instead, KoVD is used as a criterion between multiple bands to select the salient ones for hyperspectral image classification.

A. Bayes Error

In theory, the best criterion to measure the pertinence of a band is the Bayes error. The lower the error the better. However, the calculation of the Bayes error is, in general, a very complex problem [17] and it is often difficult to evaluate its probability.

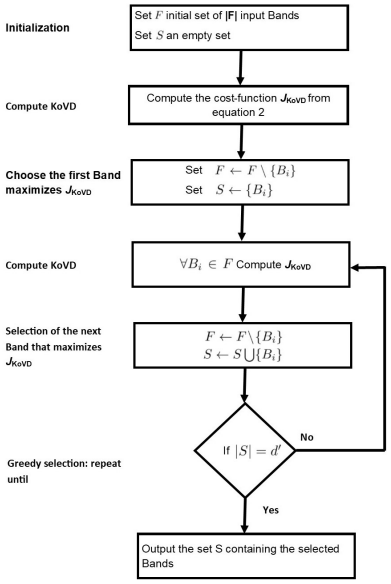


Fig. 1. The Band Selection Algorithm by Kolmogorov Variational Distance

In m-class case, the Bayes Error Probability e is given as [30]:

$$e = \int [1 - max_i P(\omega_i|x)]P(x)\mathrm{d}x \tag{8}$$

where the posterior probability

$$\begin{aligned} P(\omega_i|x) &= \frac{P(x|\omega_i) \times P(\omega_i)}{P(x)} \\ P(x) &= \sum_{c=1}^{C} P(x|\omega_c) \times P(\omega_c) \end{aligned} \tag{9}$$

$P(x)$ is the mixture density function and $P(\omega_i)$ is the prior probability

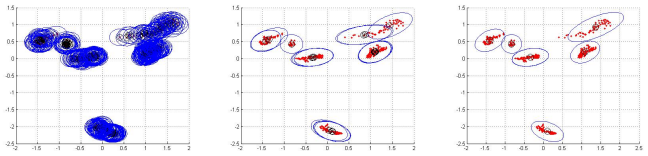
A direct calculation of equation (8) in general is often impossible or impractical [17]. In two class case, the Error Probability can be expressed as:

$$e = \frac{1}{2}\left\{1 - \int |P(\omega_1|x) - P(\omega_2|x)|P(x)\mathrm{d}x\right\} \tag{10}$$

The Kolmogorov Variational Distance is the integral from the equation (10) . From equation (1) and (10), the Error Probability e can be expressed as:

$$e = \frac{1}{2}\left\{1 - J_{KoVD}(\omega_1, \omega_2)\right\} \tag{11}$$

From equation (11) we can notice that KoVD can be expressed in terms of classification error. It has a direct relation to Bayes Error Probability, which is its main advantage unlike other probabilistic distances that only provides a bound on the error. However, KoVD requires an estimate of a probability density function and its numerical integration, which can restricts its usefulness in many practical situations [6].



(a) Iter 1 (b) Iter 10 (c) Iter 30

Fig. 2. Robust Expectation Maximization Implementation on Synthetic Data (a) All Data Points Initialization of the REM Algorithm using -(b) Reducing the Number of Clusters -(c) Finding the Optimum Number of Clusters $k = 6$.

B. KoVD Based on Gaussian Mixture Model

The Kolmogorov Variational distance based on Gaussian Mixture Model can be expressed as follows:

$$J_{KoVD}(\omega_i, \omega_j) = \sum |P(\omega_i|x) - P(\omega_j|x)|P(x) \tag{12}$$

With the equation (6) and the equation (9), the KoVD can be expressed as:

$$\begin{aligned} J_{KoVD}(\omega_i, \omega_j) &= \sum |P(x|\omega_i) \times P(\omega_i) - P(x|\omega_j) \times P(\omega_j)| \\ &= \sum |\sum_{ki=1}^{Ki} \pi_{ki} p(x|\mu_{ki}, \Sigma_{ki}) \times P(\omega_i) \\ &- \sum_{kj=1}^{Kj} \pi_{kj} p(x|\mu_{kj}, \Sigma_{kj}) \times P(\omega_j)| \end{aligned} \tag{13}$$

To compute our cost-function $J_{KoVD}(\omega_1, \omega_2)$ from equation (13) we need to estimate the following parameters: the number of clusters K, the covariance matrix Σ, the mean μ and the mixing coefficient π.

With GMM the challenge is the estimation of the parameters π, μ, Σ, K, the first three the parameters can be estimated with the Expectation-Maximization (EM) algorithm [16]. And K the number of components, the fourth parameter, is user-defined and has to be given a priori. Choosing the right value for the number of components K is crucial since it has a direct effect on the estimation of the covariance matrix. With a bad choice of K, ill-conditioned covariance matrices can be formed and the Curse of Dimensionality then can't be avoided [5] [2].

To overcame this challenge, we pursuit two approaches in order to define the optimal value for the parameter K. The first one is based on the Bayes Information Criterion (BIC) [31]. BIC is a popular measure for comparing maximum likelihood models and the model with the smallest value is the preferred one [32] [33]. BIC was introduced by [34], and defined as:

$$BIC = -2 \times \ln(likelihood) + \ln(N) \times k \tag{14}$$

With k the number of parameters estimated and N the number of observations.

The second approach is the Robust Expectation-Maximization (REM) algorithm [22]. The main advantage of this algorithm is its ability to find an optimal number of clusters K automatically, thus the number of the component

will no longer have to be defined a priori. REM also solves the issue of the initial value of the standard EM algorithm, the problem of choosing cluster centers. At first, the Robust Expectation-Maximization algorithm uses all data points as centers, and from there try to automatically reach an optimal number of clusters by discarding the clusters that do not meet the required criteria (see Fig. 2); For more detail about the algorithm, see [22].

C. Regularization Problem

For the estimation of the covariance matrix in hyperspectral imaging, the "Hughes phenomenon" and the singularity problems [25] are usually caused by the small sample size datasets. And by partitioning the already small dataset we can easily end up with an ill-conditioned mixture model [35]. For each component the sample size must not be less than the dimensionality of the data [25], since the covariance matrix should be invertible in order to compute equation (7). For Gaussian Mixture Model, the curse of dimentionality is primarily related to the estimation of the covariance matrix [36], and regularization techniques are one way around this problem:

1) Leave One Out Covariance (LOOC) : To avoid the singularity problem the LOOC estimator can be used to regulate the covariance matrix [37] [25] [3] [36]. Let S and $diag(S)$ be respectively the covariance matrix its diagonal version:

$$S_i^{looc}(\alpha_i) = \begin{cases} (1-\alpha_i)diag(S_i) + \alpha_i S_i & \text{if } 0 \leq \alpha_i \leq 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S & \text{if } 1 \leq \alpha_i \leq 2 \\ (3-\alpha_i)S + (\alpha_i - 2)diag(S) & \text{if } 2 \leq \alpha_i \leq 3 \end{cases} \quad (15)$$

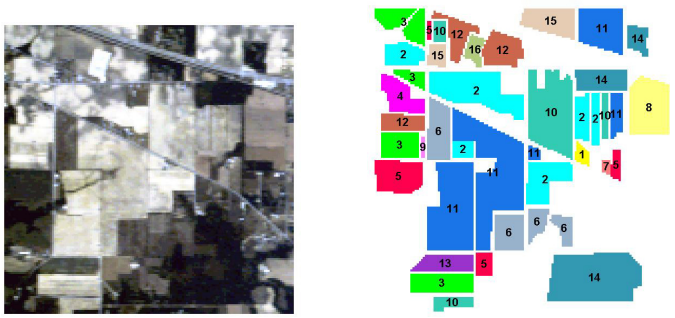
The LOOC estimator evaluate several values of α_i, and the value that maximize the average log likelihood of the Gaussian density is the optimal choice [37]. In our case, since we are using an iterative approach to select bands, using this regularization techniques as described by equation (15), did add to the complexity of the algorithm and to the computation time.

2) Maximum Entropy Covariance Selection (MECS): The MECS method deals directly with singular and unstable covariance matrices; rather than optimizing the group likelihood or the classification accuracy, MECS maximize the information under an incomplete and consequently uncertain context [38]. We are particularly interested in this method since according to [38], an optimization procedure isn't required, whenever covariance matrices are ill-posed or poorly estimated, and finally it has a much lower computational cost while performing as well as any other method.

IV. Experimental Study

A. Dataset

1) Indian Pines dataset: This scene was firstly used by David Landgrebe and his students [25] [39] [37] [2] and since become a benchmark dataset. Indian Pines dataset also known as 92AV3C dataset is a 145×145 pixels by 224 bands hyperspectral image scene captured over the Indian Pines test site in North-western Indiana on June 12, 1992, by AVIRIS sensor, with a spatial resolution of $18m$. Fig. 3 is a false-color composite of the Indian Pines scene and its ground truth map, and Table I describe the dataset.



(a) (b)

Fig. 3. (a) False Color Composite Image of Indian Pines Dataset and (b) Ground Truth.

2) Botswana dataset: In 2001-2004, over the Okavango Delta Botswana, the NASA EO-1 satellite with the Hyperion sensor gathered a sequence of data over a strip of $7.7km$. Fig. 4 is a false-color composite of the Botswana dataset and its ground truth map. The UT Center for Space Research Preprocessed the data, 14 classes were identified from the observations as described in Table II, and 145 bands were retained $[10-55, 82-97, 102-119, 134-164, 187-220]$ after removing the uncalibrated and noisy bands.

(a) (b)

Fig. 4. (a)False Color Composite Image of the Botswana Dataset and (b)Ground Truth Map.

3) Kennedy Space Center (KSC): This scene was acquired by the NASA AVIRIS, on March 23, 1996, over the Kennedy Space Center (KSC), Florida. The dataset got 176 bands after

TABLE I. DATA DESCRIPTION OF THE INDIAN PINES 16 CLASS FULL SCENE

Class number	Class name	Total samples	Training samples	Test samples
1	Alfalfa	54	29	25
2	Corn-notill	1434	719	715
3	Corn-mintill	834	419	415
4	Corn	234	117	117
5	Grass-pasture	497	249	248
6	Grass-trees	747	374	373
7	Grass-pasture-mowed	26	13	13
8	Hay-windrowed	489	243	246
9	Oats	20	10	10
10	Soybean-notill	968	483	485
11	Soybean-mintill	2468	1234	1234
12	Soybean-clean	614	304	310
13	Wheat	212	108	104
14	Woods	1294	644	650
15	Buildings-Grass-Trees-Drives	380	190	190
16	Stone-Steel-Towers	95	46	49

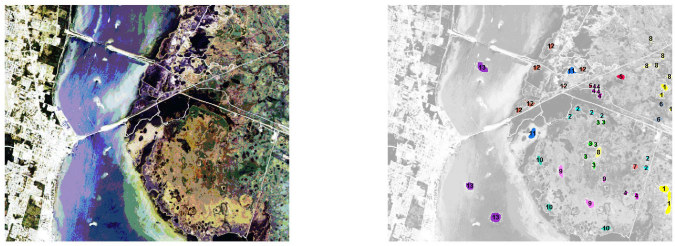
TABLE II. DATA DESCRIPTION OF THE BOTSWANA DATASET

Class number	Total samples	Training samples	Test samples
1	270	133	137
2	101	52	49
3	251	126	125
4	215	106	109
5	269	134	135
6	269	133	136
7	259	132	127
8	203	104	99
9	314	158	156
10	248	127	121
11	305	152	153
12	181	89	92
13	268	137	131
14	95	46	49

TABLE III. DATA DESCRIPTION OF THE KENNEDY SPACE CENTER DATASET

Class number	Total samples	Training samples	Test samples
1	761	373	388
2	243	122	121
3	256	128	128
4	252	127	125
5	161	79	82
6	229	116	113
7	105	51	54
8	431	214	217
9	520	259	261
10	404	203	201
11	419	211	208
12	503	250	253
13	927	465	462

removing low SNR bands with 18 m spatial resolution, and 13 classes of various types of land cover. Fig. 5 is a false-color composite of the Kennedy Space Center (KSC) scene and its ground truth map, and further details of the dataset are given in Table III



(a) (b)

Fig. 5. The Kennedy Space Center Data Set. (a) False Color Composite Image and (b) Ground Truth.

4) Salinas: Over Salinas Valley in California, the AVIRIS sensor collected 512×217 pixels by 224 bands with a spatial resolution as high as 3.7-meter per pixel. Including vineyard fields, bare soils, and vegetables, 16 classes were defined in the dataset and 20 water absorption bands were removed [108-112], [154-167], 224. Fig. 6 is a false-color composite of Salinas scene and its ground truth map.

B. Experimental Setup

The band selection approach using the Kolmogorov Variational Distance was tested using the following hardware setup: a 64-bit PC (i7-2.20GHz) with 6 GB RAM and Matlab (R2014a). The experiment was run on four benchmark hyperspectral images: the Indian Pine (92AV3C), Salinas, Kennedy Space Center, and Botswana datasets. For classification purposes, the dataset is split into two halves of training/testing. The selected bands are fed to classifiers in order de show their classification performances. The used classifier is SVM through the LIBSVM library with RBF as kernel function and the grid search technique to find the C and γ parameters [40].

C. Results and Discussions

To evaluate our proposed approach, tests were run on the benchmark dataset Indian Pine, as this scene has been often used in various studies such as [25] [39] [37] [2]. In the first

(a) (b)

Fig. 6. Salinas Data Set. (a) False Color Composite Image and (b) Ground Truth.

experiment, we measure each band independently from the rest and see how it ranks in terms of class separability according to our KoVD cost-function. The higher the value the more the classes are separable on that band. Fig. 7 we do notice that band region $170 \sim 190$ have the highest value.

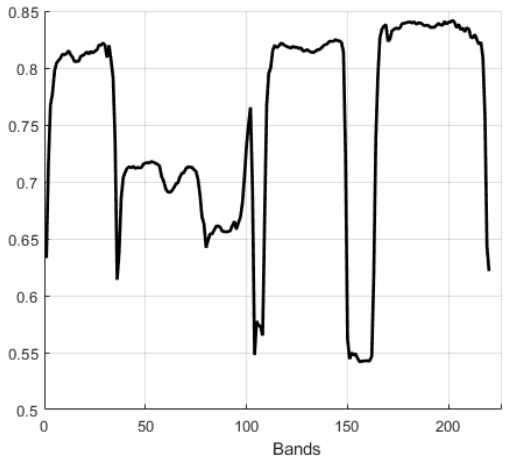
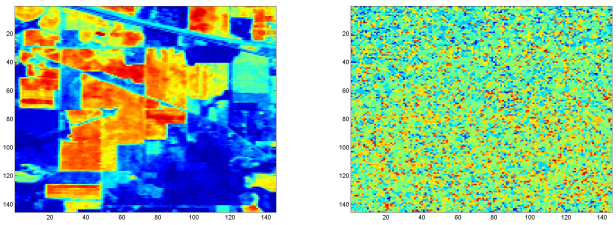


Fig. 7. KoVD Score for Each Band for 92AV3C Dataset.

In previous studies on the Indian Pine dataset [39] [41] [42] the bands $[104-108, 150-163, 220]$ were reported to be in the water absorption region with no useful information just noise as it can be seen in Fig. 7 and Fig. 8 as they got the lowest value. Hence, band selection with KoVD can successfully measure the pertinence of a band and discard those with no valuable information from the selection process.

For the second experiment, the goal is to measure the performance of the KoVD band selection algorithm with just the first two selected bands and to answer the question of whether KoVD modeled with GMM can separate classes successfully or not. For easier visual inspection the experiment was carried out on a portion of the benchmark dataset Indian Pine working



(a) (b)

Fig. 8. (a) Scene at Band Number 168 in 92AV3C Dataset, (b) Scene at Band Number 153 in 92AV3C Dataset.

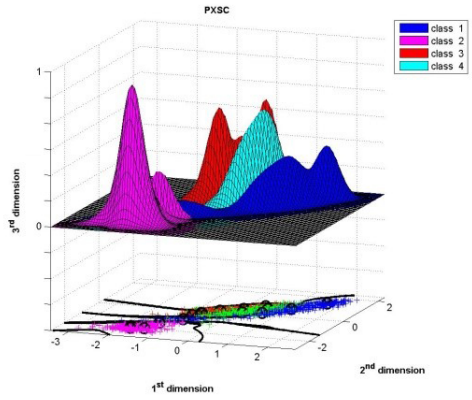
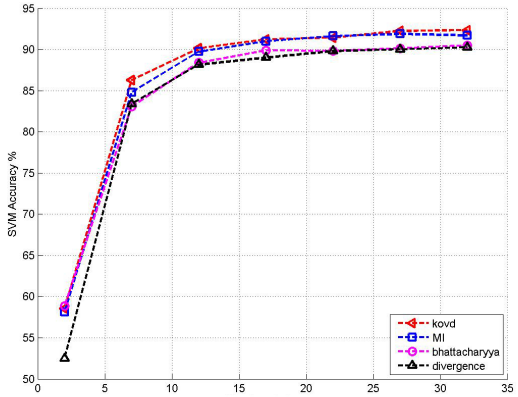


Fig. 9. Density Estimation by GMM and the Decesion Boundry that Separate between the 4 Classes.

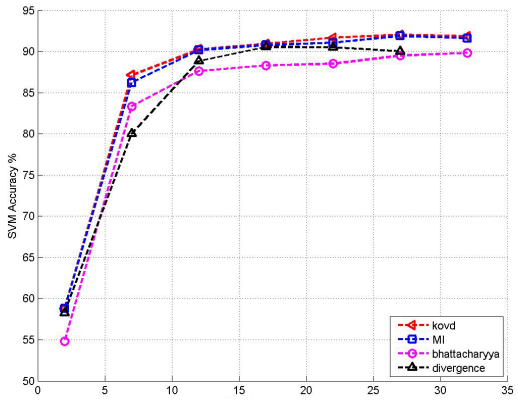
with 4 classes instead of 16 similar to [39] [41]. The data as seen in Fig. 9 is highly correlated, nonetheless, we were able to separate one class from the rest with just 2 bands out of 220 with an SVM classification score of 81.92%. On the other hand, the other classes are still correlated thus the need to add more bands to achieve the desired result. For this sub-scene, a classification score accuracy of 93.81% with SVM is achieved with only five bands and a classification score of 97.04% at dimension thirty-six. Thus, the KoVD criterion modeled with GMM can be used as a class separability measurement for band selection.

In the next step, we are going to compare the performances of KoVD against its peers - the mutual information, the divergence, and Bhattacharyya distances - in terms of classification score and the number of retained bands. Due to the complexity of the dataset, all the probabilistic distances were computed through the Gaussian mixture model. The probability estimation is computed with GMM-BIC and the GMM-REM approach, meanwhile, the SVM is used as a classifier.

In Fig. 10, 11, 12 and 13 we do notice that, for the Indian Pine dataset, the SVM classification Score for the selected bands with KoVD performs better than the ones selected with the Bhattacharyya and Divergence distances. Meanwhile, compared to the mutual information, in terms of classification Accuracy, KoVD is slightly better, in fact, the curves almost overlap each other.

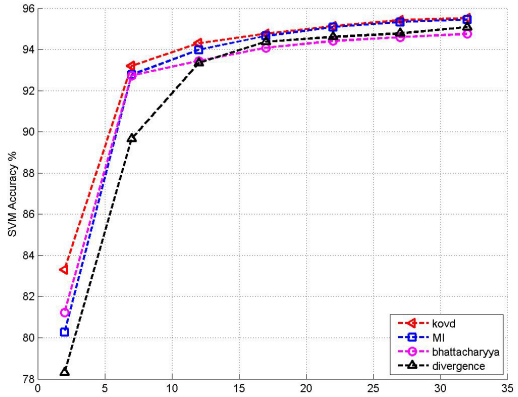


(a) GMM-BIC

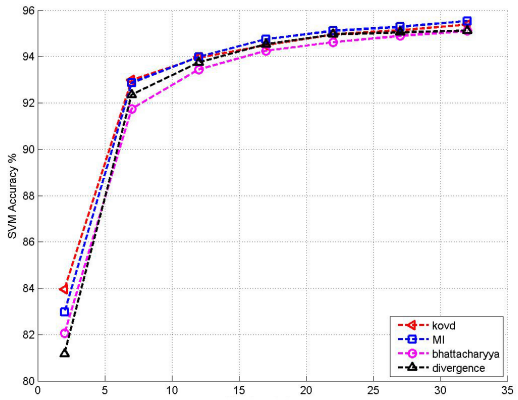


(b) GMM-REM

Fig. 10. Overall Classification Accuracy of the Selected Bands for Dataset 92AV3C using (a) GMM-BIC, (b) GMM-REM



(a) GMM-BIC



(b) GMM-REM

Fig. 11. Overall classification Accuracy of the Selected Bands for Dataset Salinas using (a) GMM-BIC, (b) GMM-REM.

Depending on the number of the selected bands, on how well the GMM was estimated, on how well the classifier parameter was chosen and on the data set itself how correlated it is and how its post-treatment was to deal with the outliers, we do notice that KoVD performs the best at times and others times the MI. According to Fig. 10, 11, 12 and 13, the results are close to each other and the margin between the classification curves of the selected bands with both distances is not wide enough to concur on the superiority of one on the others. Therefore, it is hard to decide which one of the distances is the best. Thus, we can conclude that in our setup, the KoVD performs as well as the MI and both of them perform better than the Divergence and Bhattacharyya Distances.

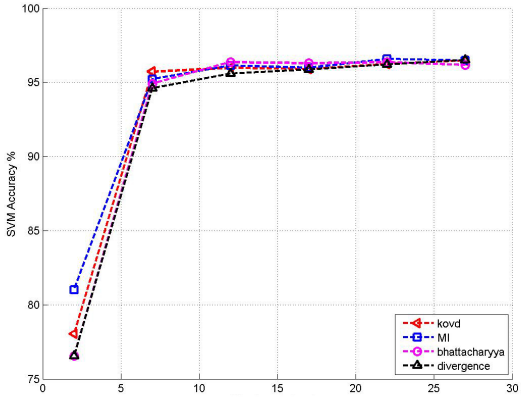
V. Conclusion

In this paper, a novel band selection approach based on the Kolmogorov Variational Distance for Hyperspectral image classification was introduced. The first experiment performed on the Indian Pine dataset have proved the efficiency and reliability of the KoVD criterion as a similarity measure. KoVD can measure the pertinence of a band, thus given a hyperspectral image dataset we can cluster the optimal bands while discarding those with no relevant information. This study was inspired by our previous work on the MI, BD, and DD. Thus, we were particularly interested in finding out how KoVD performs against these distances in terms of the numbers of bands retained and the classification accuracy. The experimental study showed that KoVD performs better than BD and DD, meanwhile against MI the results were too close; therefore, in the current setup, it is hard to decide which one is the best. Thus we can conclude that the KoVD performs as well as the MI.

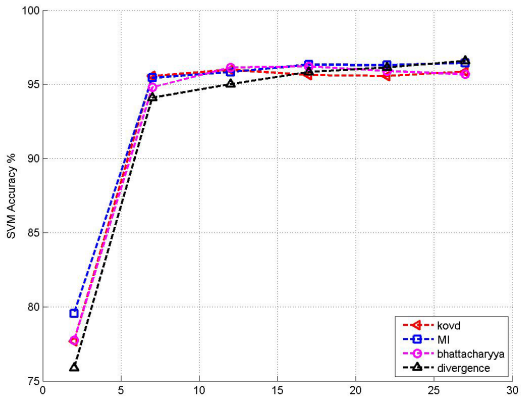
References

[1] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE transactions on information theory*, vol. 14, no. 1, pp. 55–63, 1968.

[2] L. O. Jimenez and D. A. Landgrebe, "Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 28, no. 1, pp. 39–54, Feb 1998.

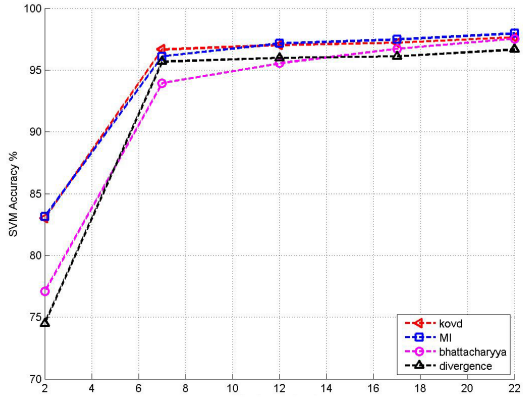


(a) GMM-BIC

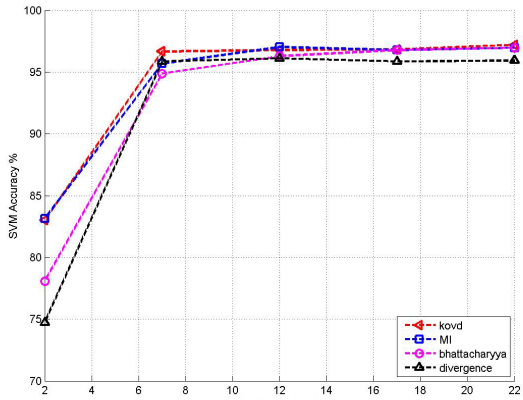


(b) GMM-REM

Fig. 12. Overall Classification Accuracy of the Selected Bands for Dataset KSC using (a) GMM-BIC, (b) GMM-REM



(a) GMM-BIC



(b) GMM-REM

Fig. 13. Overall Classification Accuracy of the Selected Bands for Dataset Botswana using (a) GMM-BIC, (b) GMM-REM

[3] J. Richards, *Remote Sensing Digital Image Analysis: An Introduction*. Springer Berlin Heidelberg, 2012. [Online]. Available: https://books.google.co.ma/books?id=ETfwQnBMP4UC

[4] A. Datta, S. Ghosh, and A. Ghosh, "Band elimination of hyperspectral imagery using partitioned band image correlation and capacitory discrimination," *International Journal of Remote Sensing*, vol. 35, no. 2, pp. 554–577, 2014. [Online]. Available: https://doi.org/10.1080/01431161.2013.871392

[5] B. M. Shahshahani and D. A. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 5, pp. 1087–1095, Sep 1994.

[6] A. R. Webb, *Statistical pattern recognition*, 2nd ed. John Wiley & Sons, 2003.

[7] J. Wang, X. Wang, K. Zhang, K. Madani, and C. Sabourin, "Morphological band selection for hyperspectral imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 8, pp. 1259–1263, Aug 2018.

[8] M. Lahlimi, M. Ait Kerroum, and Y. Fakhri, "Band selection with bhattacharyya distance based on the gaussian mixture model for hyperspectral image classification," in *Recent Advances in Electrical and Information Technologies for Sustainable Development*, S. El Hani and M. Essaaidi, Eds. Cham: Springer International Publishing, 2019, pp. 87–94.

[9] W. L. Martinez and A. R. Martinez, *Computational statistics handbook with MATLAB*. CRC press, 2007, vol. 22.

[10] M. Imani and H. Ghassemian, "Two dimensional linear discriminant analyses for hyperspectral data," *Photogrammetric Engineering & Remote Sensing*, vol. 81, no. 10, pp. 777–786, 2015.

[11] A. Datta, S. Ghosh, and A. Ghosh, "Unsupervised band extraction for hyperspectral images using clustering and kernel principal component analysis," *International Journal of Remote Sensing*, vol. 38, no. 3, pp. 850–873, 2017. [Online]. Available: https://doi.org/10.1080/01431161.2016.1271470

[12] M. P. Uddin, M. A. Mamun, and M. A. Hossain, "Feature extraction for hyperspectral image classification," in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, Dec 2017, pp. 379–382.

[13] P. K. Varshney and M. K. Arora, *Advanced image processing techniques for remotely sensed hyperspectral data*. Springer Science & Business Media, 2004.

[14] K. Burgers, Y. Fessehatsion, S. Rahmani, J. Seo, and T. Wittman, "A comparative analysis of dimension reduction algorithms on hyperspectral data," *LAMDA Research Group*, pp. 1–23, 2009.

[15] C. Lee, D. Landgrebe *et al.*, "Feature extraction based on decision boundaries," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 15, no. 4, pp. 388–400, 1993.

[16] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[17] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, ser. Computer science and scientific computing. Elsevier Science, 2013. [On-

line]. Available: https://books.google.co.ma/books?id=BIJZTGjTxBgC

[18] M. Ait Kerroum, A. Hammouch, and D. Aboutajdine, "Textural feature selection by joint mutual information based on gaussian mixture model for multispectral image classification," *Pattern Recogn. Lett.*, vol. 31, no. 10, pp. 1168–1174, Jul. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2009.11.010

[19] M. LAHLIMI, "Band selection by divergence distance based on gaussian mixture model for hyperspectral image classification," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, pp. 2330–2338, 10 2019.

[20] C. Chen, "Theoretical comparison of a class of feature selection criteria in pattern recognition," *IEEE Transactions on Computers*, no. 9, pp. 1054–1056, 1971.

[21] T. Kailath, "The divergence and bhattacharyya distance measures in signal selection," *Communication Technology, IEEE Transactions on*, vol. 15, no. 1, pp. 52–60, 1967.

[22] M.-S. Yang, C.-Y. Lai, and C.-Y. Lin, "A robust em clustering algorithm for gaussian mixture models," *Pattern Recognition*, vol. 45, no. 11, pp. 3950–3961, 2012.

[23] A. Elmaizi, H. Nhaila, E. Sarhrouni, A. Hammouch, and N. Chafik, "A novel approach for dimensionality reduction and classification of hyperspectral images based on normalized synergy," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 8, 2019. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2019.0100831

[24] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[25] B.-C. Kuo and D. A. Landgrebe, "A robust classification procedure based on mixture classifiers and nonparametric weighted feature extraction," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 40, no. 11, pp. 2486–2494, 2002.

[26] C. Chang, *Hyperspectral Data Exploitation: Theory and Applications*. Wiley, 2007. [Online]. Available: https://books.google.co.ma/books?id=NwVncgNwtI4C

[27] L. Burrell, O. Smart, G. K. Georgoulas, E. Marsh, and G. J. Vachtsevanos, "Evaluation of feature selection techniques for analysis of functional mri and eeg." in *DMIN*, 2007, pp. 256–262.

[28] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, July 1994.

[29] N. Kwak and Chong-Ho Choi, "Input feature selection for classification problems," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143–159, Jan 2002.

[30] U. Maulik, S. Bandyopadhyay, and J. Wang, *Computational Intelligence and Pattern Analysis in Biology Informatics*, ser. Wiley Series in Bioinformatics. Wiley, 2011. [Online]. Available: https://books.google.co.ma/books?id=9CBeyg2AQ4EC

[31] W. Li, S. Prasad, and J. E. Fowler, "Hyperspectral image classification using gaussian mixture models and markov random fields," *Geoscience and Remote Sensing Letters, IEEE*, vol. 11, no. 1, pp. 153–157, 2014.

[32] J. Chen and Z. Chen, "Extended bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.

[33] H. D.-G. Acquah, "Comparison of akaike information criterion (aic) and bayesian information criterion (bic) in selection of an asymmetric price relationship," *Journal of Development and Agricultural Economics*, vol. 2, no. 1, pp. 001–006, 2010.

[34] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[35] M. M. Dundar and D. Landgrebe, "A model-based mixture-supervised classification approach in hyperspectral data analysis," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 40, no. 12, pp. 2692–2699, 2002.

[36] M. Fauvel, C. Dechesne, A. Zullo, and F. Ferraty, "Fast forward feature selection for the nonlinear classification of hyperspectral images," *arXiv preprint arXiv:1501.00857*, 2015.

[37] S. Tadjudin and D. A. Landgrebe, "Covariance estimation with limited training samples," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 4, pp. 2113–2118, July 1999.

[38] C. E. Thomaz, D. F. Gillies, and R. Q. Feitosa, "A new covariance estimate for bayesian classifiers in biometric recognition," *IEEE Transactions on circuits and systems for video technology*, vol. 14, no. 2, pp. 214–223, 2004.

[39] S. Tadjudin and D. A. Landgrebe, "Robust parameter estimation for mixture model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 1, pp. 439–445, 2000.

[40] S. M and G. Sadashivappa, "Hyperspectral image classification using support vector machine with guided image filter," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 10, 2019. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2019.0101038

[41] G. Camps-Valls and L. Bruzzone, *Kernel methods for remote sensing data analysis*. John Wiley & Sons, 2009.

[42] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, Oct 2016.

Data Augmentation using Generative Adversarial Network for Gastrointestinal Parasite Microscopy Image Classification

Mila Yoselyn Pacompia Machaca[1], Milagros Lizet Mayta Rosas[2], Eveling Castro-Gutierrez[3],
Henry Abraham Talavera Díaz[4] and Victor Luis Vasquez Huerta[5]
[1]2345Universidad Nacional de San Agustín de Arequipa
Arequipa, Perú

***Abstract*—Gastrointestinal parasitic diseases represent a latent problem in developing countries; it is necessary to create a support tools for the medical diagnosis of these diseases, it is required to automate tasks such as the classification of samples of the causative parasites obtained through the microscope using methods like deep learning. However, these methods require large amounts of data. Currently, collecting these images represents a complex procedure, significant consumption of resources, and long periods. Therefore it is necessary to propose a computational solution to this problem. In this work, an approach for generating sets of synthetic images of 8 species of parasites is presented, using Deep Convolutional Adversarial Generative Networks (DCGAN). Also, looking for better results, image enhancement techniques were applied. These synthetic datasets (SD) were evaluated in a series of combinations with the real datasets (RD) using the classification task, where the highest accuracy was obtained with the pre-trained Resnet50 model (99,2%), showing that increasing the RD with SD obtained from DCGAN helps to achieve greater accuracy.**

***Keywords*—Generative Adversarial Network (GAN); Deep Convolutional Generative Adversaria Network (DCGAN); gastrointestinal parasites; classification; deep learning**

I. Introduction

Diseases caused by parasites are a public health problem on a global scale; they can be of high risk and high prevalence, as shown by their incidence rates in the population. According to the World Health Organization (WHO), the malaria disease caused by the Plasmodium parasite causes 400000 deaths per year [1], and more than 2 billion people are infected with soil-transmitted helminthiases. [2].

According to the National Institute of Health of Peru (INS), intestinal parasitism increased its prevalence rate among sectors with fewer resources [3]. Although the cases registered in the south of the country with a diagnosis of intestinal parasitism were more frequent in children or people in school age 41.75%, this problem also affects people in adulthood or youth, who also presented a considerable percentage of incidence in this disease with 20.45% and 35.09%, respectively [4].

Here is the importance of addressing this problem using computational methods to implement support systems for medical diagnosis, recognition, and classification of images obtained through the microscope. However, methods based on deep learning that have had excellent results in similar applications require large datasets. The collection of medical data involves a lengthy procedure that may require applying different protocols and the intervention of a specialist. Therefore it is a task that represents a significant consumption of time and resources.

In this work, the application of two techniques for image enhancement is presented, Wiener and Wavelet, in order to obtain an improvement in the quality of the images; also, an approach to increase data is presented for the generation of synthetic training samples of microscopy images of eight species of gastrointestinal parasites, using DCGAN a variation of GAN [5], from a reduced initial dataset. In total, three sets of augmented data are generated resulting from the training of DCGAN, both for the dataset resulting from the application of image enhancement with Wiener filter, the dataset resulting from image enhancement with Wavelet denoising, and the original dataset without enhancement. The three datasets are used for training in classification models pre-trained by transfer learning, Resnet34, Resnet50, among others, independently to verify a performance improvement and compare the results obtained.

The article's structure is explained below: In Section II, related works are addressed, Section III describes the methodology used, to finish with the results and conclusions in Sections IV and V, respectively.

II. Related Works

According to the literature, various authors have applied deep learning in analyzing medical images [6], [7]. And in particular for the treatment of microscopy images of parasites [8], [9].

However, RD collection usually has a difficult and time-consuming procedure; this generates insufficient data for experimentation with deep learning methods for tasks such as image segmentation and classification. For this reason, some authors focus on increasing datasets for better results.

To address this problem in classifying in microscopy images of parasites, some authors used: The Python Augmentor library, [10], or traditional data augmentation techniques, [11]. On the other hand, due to the problems encountered when working with layered laser scanning microscopy image data, in [12] M. Bloice et. al, present their own software package: Augmentor which is a stochastic pipeline-based augmentation

library of images and includes features relevant to the domain of biomedical images, in order to to ensure that the new data is meaningful.

In recent works, GANs have been used for the task of recognizing microscopic images of parasites [13], where through transfer learning, the authors intend to identify the parasite Toxoplasma gondii using Fuzzy Cycle Generative Adversarial Network (FCGAN) and in [10], where V. Fomene proposes a classifier for the diagnosis of malaria in rural areas using MobileNet and GAN.

GAN has even been used to the transfer between different modalities of microscopy images, as presented in [14], the authors investigate a conditional GAN in order to use the monitoring techniques of cells through the microscope: Phase Contrast (PC) and Differential Interference Contrast (DIC), passing from one to the other through the proposed algorithm. The focus of this work is to use GANs for the generation of data augmentation in order to face the problem of data deficiency for the application of deep learning models in tasks such as the segmentation [15], [16], and classification [17], [18], of microscopy images.

GANs have also been used to segmentation microscopic images of pluripotent retinal pigmented epithelial stem cells [19], where M. Majurski et al., use GAN in one of their approaches to optimize the coefficients of their Convolutional Neural Network (CNN). Some authors used Conditional GAN (CGAN), a conditional version of GAN, in which auxiliary information is fed to both the discriminator and the generator as an additional input layer [20]. This version was used to increase the dataset made up of polycrystalline iron images [15], where they also presented the transfer learning application with data fusion simulations obtained with the Monte Carlo Potts model and the image style information obtained from real images. CGAN was also used to increase the dataset of microscopic images of red blood cells [17]. The architecture that was proposed in [16], for the generation of the nuclei cell image-mask pair, consists of 2 stages where they first use a GAN to learn the data distribution of the masks and generate a synthesized binary mask and then incorporate this mask synthesized in the second GAN which also learns a mapping of the random noise vector to perform a conditional generation of the synthesized image.

On the other hand, Deep Convolutional GAN (DCGAN) has also been considered for data augmentation. DCGAN, which is a variation of GAN that uses convolutional layers in the discriminator and convolutional-transpose in the generator, where also the discriminator is composed of stridden convolution layers, batch norm layers, and LeakyReLU activations, and the generator is composed of convolutional-transpose layers, batch norm layers, and ReLU activations, [21].

R. Verma et al., present in [18], use DCGAN to generate synthetic samples of 5 classes of proteins, which were used in the classification task both before and after the increase of SD, comparing DCGAN results with those of traditional methods. In search of better results, some authors have modified the original structure of the GANs by including two discriminators [14]; it has also been considered to include U-net structural elements in their GANs [14], [19].

It has also tried to improve the quality of microscopic images as preprocessing before data augmentation, such as contrast enhancement with Histogram Equalization (HE), Adaptive Histogram Equalization (AHE), and Contrast Limited Adaptive Histogram Equalization (CLAHE) [18].

Some of the resulting augmented datasets have been evaluated in pre-trained deep learning models under a transfer learning approach, such as VGG, Resnet, NasNet, Inception, MobileNet. Because these have become popular in medical image classification and segmentation work [11], [13], [18].

The review of the related works shows that the use of GANs for data augmentation is competitive, performs well, or even outperforms traditional augmentation methods.

This work's contribution is a) the improvement of the quality of the dataset of microscopy images of eight species of gastrointestinal parasites with the Wiener and Wavelet filters, b) the use of DCGAN to augment the dataset by generating SD, and c) the resulting datasets for the training of deep learning pre-trained classification models by transfer learning and finally the comparison of the results obtained.

III. Materials and Methods

For the development of this experimental research, a process that is applied in steps was defined. The proposed methodology consists of image preprocessing with improvement techniques, use of DCGAN for data augmentation, and evaluating synthetic datasets generated by classifying images with previously trained models using the transfer-based learning approach. In Fig. 1, the methodology of the approach for this work is detailed.

A. Dataset

The dataset consists of a vector of characteristics of each image to be trained. Microscopy images of parasites were used. There are a total of 954 images:

- Ascaris.
- Hookworms.
- Trichuris trichiura.
- Hymenolepis nana.
- Diphyllobothrium pacificum.
- Taenia solium.
- Fasciola hepatica.
- Enterobius vermicularis.

Each image from ground truth has 1200 x 1600 (height by width) and has three channels. A zoomed sample of each species of these parasites is presented in Fig. 2.

B. Image Enhancement

Two techniques were applied independently, Wiener filter and Wavelet denoising to improve the set of images.

1) Wiener filter: Wiener filter has been applied for image enhancement to reduce the additive noise and add variations in images. Wiener filter is one of the most known filters used for image enhancement [22].

2) Wavelet denoising: Wavelet denoising has been applied with the purpose to vary characteristics in the images [23].

Fig. 3, shows a sample of the results in terms of image enhancement.

C. Generative Adversarial Network

CNNs are designed for data with spatial structures, and they are composed of many filters, which convolve or slide through the data and produce an activation at each slide position. These activations produce a feature map representing how much the data in that region activated the filter.

On the other hand, GANs are a type of generative model because they learn to copy the data distribution of the data they give them. Therefore, they can generate novel images that look alike. A GAN is called an "adversary" because it involves two competing networks (adversaries) trying to outwit each other.

The generator (G) is a neural network, which takes a vector of random variables (Latent Space), and produces an image *igenerator*.

The discriminator (D) is also a neural network, which takes an image, *I*, and produces a single output *p*, deciding

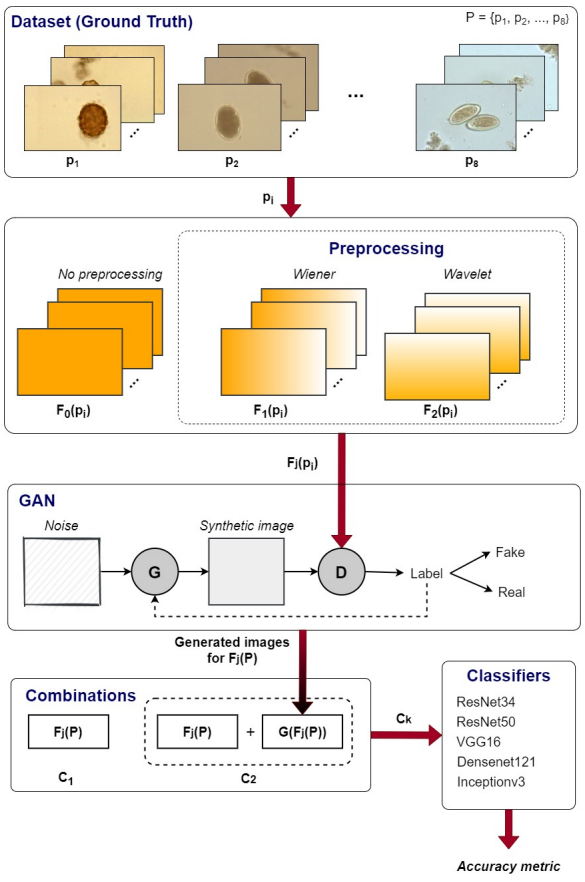


Fig. 1. Methodology. Top images should correspond to a specific species of parasite; then, it could be applied filter wiener or wavelet denoising as a preprocessing technique. After this, a DCGAN is able to generate images with one generator and one discriminator. Finally, ground truth and generated images are evaluated with classifiers Resnet34, Resnet50, and VGG. The process of data augmentation is applied independently for each parasite species.

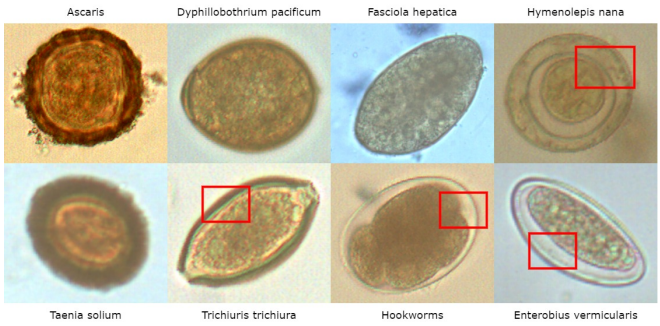


Fig. 2. Parasite species: A sample of each parasite species that is addressed in this work is presented, and special characteristics of the edges of these are highlighted in the red boxes. Hymenolepis nana has (from outside to inside) a wide rim, a semitransparent middle part, and an inner part with a texture similar to the outer ring. Trichiuris trichiura has different color intensities on edge. Hookworms have a semitransparent membrane with an irregular internal shape. Enterobius vermicularis has a semitransparent membrane with an internal oval shape.

the probability that the image is real. When p = 1p = 1, the discriminator strongly believes that the image is real, and when p = 0p = 0, the discriminator strongly believes that the image is false.

The discriminator (D) receives *igenerator*, and is taught that the image is false. In more concrete terms, the discriminator maximizes $log(1-pgenerator)$. The discriminator receives a real image, *ireal*, and is taught that the image is real, or maximizes *log(preal)*.

The generator tries to do the exact opposite; it also tries to make the discriminator maximize the probability that it believes that the false image is real, so the generator will be trying to maximize *log(pgenerator)* [5].

As mentioned in Section II, a DCGAN uses convolutional layers in the discriminator (D) and convolutional-transpose layers in the generator (G). The discriminator, composed of stridden convolution layers, batch norm layers, and LeakyReLU activations, receives an image and returns a scalar probability that it is from the distribution of real data.

The generator, consisting of convolutional-transpose layers, batch norm layers, and ReLU activations, receives a latent vector, z, which is extracted from a standard normal distribution and returns an RGB image [21].

D. Data Augmentation

DCGAN was used as part of the proposed methodology for the generation of synthetic images for data augmentation. For the DCGAN, binary cross-entropy is used as a loss function. And for discriminator and generator, Adam optimizers are applied with a learning rate of 0.0002. In Table I, DCGAN structure used for this work is detailed. A class called GAN is created. It imports the relevant classes and initializes the variables. A generator model was created with the following layers:

- Convolutional transpose 2D layer with an input of *100x100* image with 3 channels.
- Batch normalization: normalizes the data.

Fig. 3. Enhanced image samples. (a) Microscopy image of Fasciola hepatica parasite corresponding to the original dataset, (b) Image enhanced with Wiener filter, (c) Image enhanced with Wavelet denoising.

- ReLU.

These last three layers are repeated and make a block of four CBR (Convolutional Transposed 2D - Batch Normalization - ReLU) Finally, the following layers are applied:

- Convolutional transpose 2D
- Hyperbolic Tangent

A discriminator can also become a sequential model by going in the opposite direction. The discriminator was built as a model that is governed by the following directives:

- The first layer is for applying a convolution to a 64x64 image with 3 channels.
- Add a Leaky ReLU activation feature.

Then, the following three layers are applied:

- Convolution 2D
- Batch normalization 2D
- Leaky ReLU.

Finally, the following layers are applied to get the output decision:

- Convolution 2D
- Sigmoid

E. Evaluation

Classification task is defined to evaluate the synthetic microscopy image datasets generated by DCGAN. For this, the pre-trained deep learning models Resnet34, Resnet50, VGG16, Densenet121, and Inceptionv3 are used under a transfer learning approach.

1) Accuracy: The metric that was chosen to evaluate the models is Accuracy1.

$$Accuracy = \frac{\sum_c TP_c + FN_c}{\sum_c TP_c + TN_c + FP_c + FN_c} \quad (1)$$

TABLE I. DCGAN STRUCTURE

LAYER	DESCRIPTION
ConvTranspose2d	Applies a 2D transposed convolution operator over the image.
BatchNorm2d	Applies Batch Normalization over a 4D input.
ReLU	Applies the rectified linear unit function element-wise.
Tanh	Applies hyperbolic tangent function element-wise.
Conv2d	Applies a 2D convolution over an input signal composed of several input planes.
LeakyReLU	Applies leaky rectified linear unit function element-wise.
Sigmoid	Applies sigmoid function element-wise.

IV. EXPERIMENTATION AND RESULTS

A. Tools and technological infrastructure

For the experimentation, the following software tools were used: Python (libraries such as Pytorch, Numpy, Scipy, Scikit Image) and Cuda. The preprocessing, training and classification tasks were performed in a workstation with a processor Intel Xeon Gold 5115 CPU, Memory 128 GB, and a Video card Quadro P5000 16GB.

B. Data Augmentation Results

DCGAN was trained separately in isolation for each of the image sets for each mentioned parasite species. The average execution time for the datasets without any image enhancement considering 15000 epochs was 11 hours; the average execution time for the datasets improved with the Wiener filter, considering 15000 epochs was 8 hours. As for the data sets improved with the Wavelet denoising. Initially, 15000 epochs were considered. However, it had to be reduced to only 3000 epochs due to the long execution time that represented a high computational cost and the small number of images generated, obtaining an average of 7 hours, Fig. 4.

The numbers of synthetic images obtained in these runs were 612 for the datasets without image enhancement, 470 for the datasets enhanced with the Wiener filter, and 159 for the datasets enhanced with Wavelet denoising. These results

represent an increase of 65.88%, 50.59%, and 17.12%, over the original dataset, respectively. Fig. 5, shows the most significant amounts of synthetic images generated for each species of parasite.

According to [16], large-scale GAN imaging and training stability is challenging. The original size of the images 1200 x 1600 was kept for the training in DCGAN; the synthetic images obtained have allowed magnifying the domain, providing a greater training ground for the classification of parasites' species. Fig. 6, shows a data grid with 64 generates synthetic images.

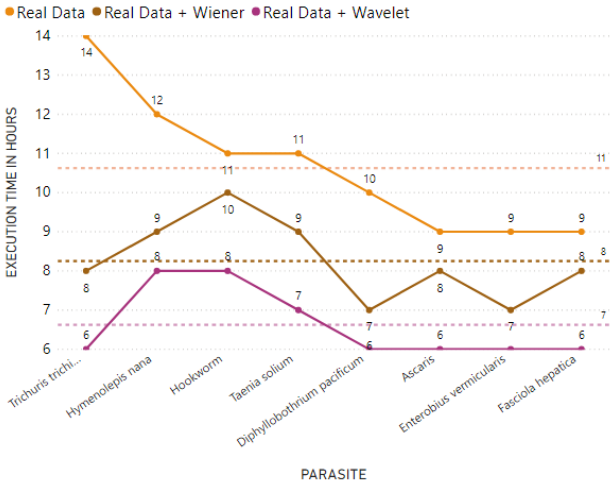


Fig. 4. Execution time for generating synthetic images in DCGAN

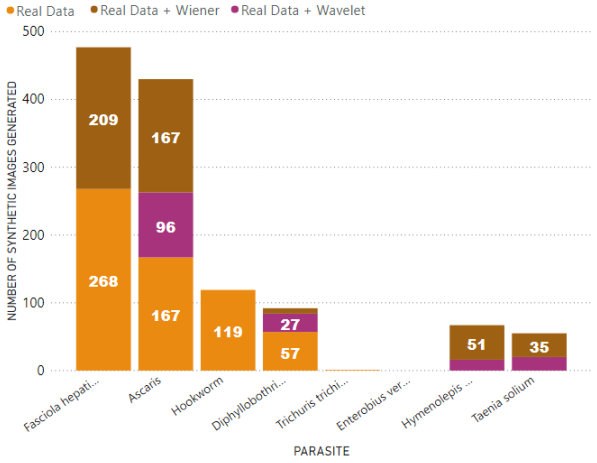


Fig. 5. Increment of the dataset. Real Data: Images generated from real data without enhancement. Real Data + Wiener: Images generated from real data enhanced with Wiener. Real Data + Wavelet: Images generated from real data enhanced with Wavelet.

C. Results and Discussion

It was decided to train model using Resnet34, Resnet50, VGG16, Densenet121, and Inceptionv3, independently, resizing the images from 1200 x 1600 to 64 x 64 to improve the results and to evaluate the datasets and get a comparative view of the accuracy metric, considering only 50 epochs, with the datasets: First, only with the original RD without any preprocessing or image enhancement applied. Second, with the original RD without any preprocessing or image enhancement combined with the SD generated. Third, only with the RD obtained after applying the Wiener filter. Fourth, with the RD obtained after applying the Wiener filter combined with the SD generated from it. Fifth, only with the RD obtained after applying the Wavelet denoising and sixth, with the RD obtained after applying the Wavelet denoising combined with the SD generated from it.

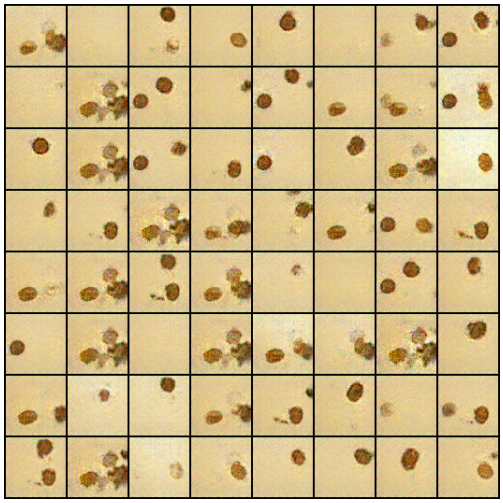


Fig. 6. Grid of synthetic images generated after 6000 epochs, parasite: Ascaris without image enhancement before DCGAN training.

The results can be seen in Table II. The best accuracy: 0.992 was obtained with the Resnet50 model, and the dataset: SD obtained after applying the Wiener filter + the RD. This shows that using SD in the image classification task improves the accuracy results.

According to the literature, no data augmentation works have been found for classification of microscopy images of the eight species of gastrointestinal parasites using DCGAN.

However, the use of different techniques was found, for objectives similar to those that have been raised in the present work. M. Roder et. al, present in [24], the use of Restricted Boltzmann Machines for data augmentation and the classification of helminth eggs using Deep Belief Networks, obtaining a balanced accuracy of 92.09%, the authors worked with grayscale microscopy images , while in this work RGB scale images were used.

The work of R. Verma et al. [18] is considered a similar methodology to that described in the present work. Verma used DCGAN for the generation of SD of 5 classes of proteins and also classification models based on transfer learning with which they obtained a result of 0.924 for the accuracy measure in VGG16 training. The best result in classification was obtained with the Resnet50 model with 0.992 for the accuracy measure in this work. It should be noted that in their work, the authors used different preprocessing techniques and datasets.

V. Conclusions

DCGAN was used to increase synthetic data, and the results generated were compared through the classification

TABLE II. CLASIFICATION RESULTS.

Model	**Cassification Metric: Accuracy**					
	No Dataset Enhancement		Dataset with Wiener filter		Dataset with Wavelet denoising	
	RD	RD + SD	RD	RD + SD	RD	RD + SD
ResNet34	0.619	0.974	0.978	0.978	0.935	0.990
ResNet50	0.989	0.961	0.967	**0.992**	0.978	0.930
VGG16	0.869	0.902	0.826	0.892	0.847	0.888
Densenet121	0.967	0.987	0.967	0.985	0.956	0.907
Inceptiov3	0.967	0.980	0.956	0.956	0.956	0.990

task so that greater precision was obtained with the Resnet50 model, with an accuracy of 0.992, and the dataset: RD + SD obtained after applying the Wiener filter. Other pr-trained models also showed similar results, such as Resnet34 and Inceptionv3, with an accuracy of 0.990, both with the dataset: RD + SD obtained after applying Wavelet denoising. Therefore, it is shown that the use of RD + SD provides greater accuracy in the classification compared to using only the RD.

The combinations of datasets that gave the best results with an average accuracy of 0.961 were: RD + SD obtained from the set of images without improvements and RD + SD obtained from the set of images improved with the Wiener filter.

The highest amount of SD generated from RD was obtained without improvements (612). On the other hand, the amount of SD obtained after applying both the Wiener filter (470) and the Wavelet denoising (159) was lower due to the characteristics of the images as it highlights impurities present in the samples.

Despite the complexity and computational cost of the DCGAN training stage, an accuracy of 99.2% was achieved. However, there is still room for improvement. In the future, it is intended to find preprocessing techniques that allow improving all image sets in order to generate quality synthetic images for the eight species of parasites. It is recommended to use techniques other than Wiener filter and Wavelet denoising for Uncinaria, Trichiuris trichiura, and Enterobius vermicularis parasites that highlight their morphological characteristics to obtain better samples of synthetic data.

ACKNOWLEDGMENT

The authors would like to thank the support and subvention of the UNIVERSIDAD NACIONAL DE SAN AGUSTÍN DE AREQUIPA to the Project called: Assistance in the diagnosis of gastrointestinal parasitosis through prevalence rates and micrographs, using high performance computational and computer vision applied to the Arequipa region with the contract No IBA-BIOM-2018-1. Thanks to the CiTeSoft Contract: EC-0003-2017-UNSA for the equipment and the resources bring to the project.

REFERENCES

[1] World Health Organization and others, "Malaria eradication: benefits, future scenarios and feasibility," 2020.

[2] World Health Organization and others, "Soil-transmitted helminthiases: eliminating as public health problem soil-transmitted helminthiases in children: progress report 2001-2010 and strategic plan 2011-2020," 2012.

[3] M. Beltrán, R. Tello, and C. Náquira, "Manual de procedimientos de laboratorio para el diagnóstico de los parásitos intestinales del hombre," 2003.

[4] G. Pajuelo Camacho, D. Lujan Roca, and B. Paredes Perez, "Estudio de enteroparásitos en el hospital de emergencias pediátricas, lima-perú." *Revista Médica Herediana*, vol. 16, no. 3, pp. 178–183, 2005.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[7] O. Z. Kraus, J. L. Ba, and B. J. Frey, "Classifying and segmenting microscopy images with deep multiple instance learning," *Bioinformatics*, vol. 32, no. 12, p. 52–59, 2016.

[8] A. Peixinho, S. Martins, J. Vargas, A. Falcao, J. Gomes, and C. Suzuki, "Diagnosis of human intestinal parasites by deep learning," *Computational Vision and Medical Image Processing V: Proceedings of the 5th Eccomas Thematic Conference on Computational Vision and Medical Image Processing (VipIMAGE 2015, Tenerife, Spain*, p. 107, 2015.

[9] J. A. Quinn, R. Nakasi, P. K. Mugagga, P. Byanyima, W. Lubega, and A. Andama, "Deep convolutional neural networks for microscopy-based point of care diagnostics," in *Machine Learning for Healthcare Conference*, 2016, pp. 271–281.

[10] V. Fomene, "Developing a machine learning model for malaria diagnosis in rural areas," 2018.

[11] M. M. Aladago, "Classification and quantification of malaria parasites using convolutional neural networks," 2018.

[12] M. D. Bloice, P. M. Roth, and A. Holzinger, "Biomedical image augmentation using augmentor," *Bioinformatics*, vol. 35, no. 21, pp. 4522–4524, 2019.

[13] S. Li, A. Li, D. A. M. Lara, J. E. G. Marín, M. Juhas, and Y. Zhang, "A novel transfer learning approach for toxoplasma gondii microscopic image recognition by fuzzy cycle generative adversarial network," *bioRxiv*, pp. 567–891, 2019.

[14] L. Han and Z. Yin, "Transferring microscopy image modalities with conditional generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 99–107.

[15] B. Ma, X. Wei, C. Liu, X. Ban, H. Huang, H. Wang, W. Xue, S. Wu, M. Gao, Q. Shen *et al.*, "Data augmentation in microscopic images for material data mining," *npj Computational Materials*, vol. 6, no. 1, pp. 1–9, 2020.

[16] S. Pandey, P. R. Singh, and J. Tian, "An image augmentation approach using two-stage generative adversarial network for nuclei image segmentation," *Biomedical Signal Processing and Control*, vol. 57, p. 101782, 2020.

[17] O. Bailo, D. Ham, and Y. Min Shin, "Red blood cell image generation for data augmentation using conditional generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[18] R. Verma, R. Mehrotra, C. Rane, R. Tiwari, and A. K. Agariya, "Synthetic image augmentation with generative adversarial network for enhanced performance in protein classification," *Biomedical Engineering Letters*, vol. 10, no. 3, pp. 443–452, 2020.

[19] M. Majurski, P. Manescu, S. Padi, N. Schaub, N. Hotaling, C. Simon Jr, and P. Bajcsy, "Cell image segmentation using generative adversarial networks, transfer learning, and augmentations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[20] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[21] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[22] R. K. B. Vijayalakshmi A, "Deep learning approach to detect malaria from microscopic images." *Multimedia Tools and Applications*, vol. 79, p. 15297–15317, 2019.

[23] A. M. Wink and J. B. T. M. Roerdink, "Denoising functional mr images: a comparison of wavelet denoising and gaussian smoothing," *IEEE Transactions on Medical Imaging*, vol. 23, no. 3, pp. 374–387, 2004.

[24] M. Roder, L. A. Passos, L. C. F. Ribeiro, B. C. Benato, A. X. Falcão, and J. P. Papa, "Intestinal parasites classification using deep belief networks," in *International Conference on Artificial Intelligence and Soft Computing*. Springer, 2020, pp. 242–251.

Comparative Analysis of Threat Modeling Methods for Cloud Computing towards Healthcare Security Practice

Prosper K. Yeng[1]
Department of Information Security and Communication Technology
NTNU
Gjøvik, Norway

Stephen D. Wulthusen[2]
Department of Information Security and Communication Technology
NTNU
Gjøvik, Norway
School of Mathematics and Information Security
Royal Holloway, University of London
Egham, United Kingdom

Bian Yang[3]
Department of Information Security and Communication Technology
NTNU
Gjøvik, Norway

***Abstract*—Healthcare organizations consist of unique activities including collaborating on patients care and emergency care. The sector also accumulates high sensitive multifaceted patients' data such as text reports, radiology images and pathological slides. The large volume of the data is often stored as Electronic Health Records (EHR) which must be frequently updated while ensuring higher percentage up-time for constant availability of patients' records. Healthcare as a critical infrastructure also needs highly skilled IT personnel, Information and Communication Technology (ICT) and infrastructure with regular maintenance culture. Fortunately, cloud computing can provide these necessary services at a lower cost. But with all thees enormous benefits of cloud computing, it is characterized with various information security issues which is not enticing to healthcare. Amid many threat modelling methods, which of them is suitable for identifying cloud related threats towards the adoption of cloud computing for healthcare? This paper compared threat modelling methods to determine their suitability for identifying and managing healthcare related threats in cloud computing. Threat modelling in pervasive computing (TMP) was identified to be suitable and can be combined with Attack Tree (AT), Attack Graph (AG) and Practical Threat Analysis (PTA) or STRIDE (spoofing, tampering, repudiation, information disclosure, denial of service and elevation of privilege). Also Attack Tree (AT) could be complemented with TMP, AG and STRIDE or PTA. Healthcare IT security professionals can hence rely on these methods in their security practices, to identify cloud related threats for healthcare. Essentially, privacy related threat modeling methods such as LINDDUN framework, need to be included in these synergy of cloud related threat modelling methods towards enhancing security and privacy for healthcare needs.**

Keywords*—Cloud computing; healthcare; threat modelling; security practice; data privacy

I. Introduction

Flexibilities of cloud computing has provided huge benefit to variety of users. So, individuals, governmental and non-governmental organizations, SMEs and other companies are adopting cloud services such as Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS). Adopting cloud computing can be very useful for healthcare organizations. It can enable healthcare organizations to focus on their core business of therapeutic services while maximizing the various benefits such as easy collaboration and data sharing, mobility and cost reduction on ICT services [1]. Cloud computing is a kind of distributed system aimed at providing unlimited shared pool computing resources (hardware or software) to registered users [2], [4].The resources can be scaled up or down to meet each clients' need [4]. Cloud Service Providers (CSP) mostly host services such as applications (SaaS), application development platforms and tools (PaaS) or servers, storage and other virtualized computing resources (IaaS) and makes them available to their clients on the internet. Some of these CSP include Microsoft Azure Services (MAS), Amazon Web Services (AWS), Google Cloud Platform (GCP), IBM, NetSuite and Salesforce [2].

Users who do not have the capacity to acquire and own such systems, can basically adopt to these cloud services as tenants to the providers, at a much lower cost [5]. Institutions which require temporal resources such as storage, processing and development platforms can also leverage on the capabilities of cloud computing. In addition to the lower costs, the tenants of the cloud service companies could benefit from scalability, business continuity, collaboration efficiency, flexibility and strategic values [5], [6]. Based on perceived demand for cloud-based systems, IBM and Active Health Management developed "Collaborative Care Solution" which was implemented in 2010 to support medical staff in accessing healthcare data from different sources [1]. Additionally, the General Electric (GE) also came up with "centricity practice" cloud-based healthcare system [7]. The system was patient centered which enables self-service of patients such that the patient could be able to communicate with their healthcare providers at remote locations but in a secured manner. Similarly, Dell's cloud based solution focused on EHR for small and medium scale healthcare organizations [8]. Additionally, the National Health Service (NHS) in the UK proposed cloud-based solutions for financial relieves. In fact, the need for cloud-based systems cannot be over emphasized but security

and privacy hurdles need to be clarified. This study therefore compared threat modeling for cloud computing to assess their suitability for assessing healthcare related threats in cloud computing towards countermeasures. The remaining of the paper consists of the Scope, Research Problem and Contribution. This is followed by the background section which describes security challenges in cloud computing for healthcare, Overview of Cloud Computing, features and Models of cloud Computing, Healthcare related information security threats, threat modelling methods and their related characteristics. The method used in the review, and the findings were provided under the method and findings sections respectively. These findings were subsequently discussed and concluded.

II. Scope, Research Problem and Contribution

Different type of threat modeling methods have been analyzed in regards to their suitability for threat modeling cloud computing environment. But with the need for healthcare to adopt to extra security measures for protecting sensitive data while making it available in a timely way, dedicated threat modeling methods are required.

Therefore, this study reviewed threat modelling methods for cloud computing towards healthcare security practice. To enable healthcare sector to successfully adopt cloud computing, some key issues relating to security, need to be addressed. For instance, what threat modelling methods can be efficiently used by healthcare IT professionals to determine comprehensive threats in cloud computing towards mitigation? How can healthcare staffs' security practice be effectively analysed in the context of big data in cloud computing towards enhancing security?

Though the above questions have not been addressed by existing studies, there have instead been varying opinions regarding effective threat modelling methods for cloud computing [6], [14], [26]–[29]. Additionally, threat modeling methods are being used in their isolated ways in threat modeling healthcare systems. Example, Abomhara et al adopted STRIDE-Based threat modeling method for telehealth systems [56]. Similarly, threat modeling methods were individually used for threat modelling mobile health systems [57], electronic health records system [58] and home care system in the cloud [59]. However, the further question is whether the isolation use of these threat models are effective enough to cover the relevant threats of cloud computing needs for healthcare. This study therefore answered the above research questions, having compared cloud related threat modelling methods and assessed their advantages and disadvantages with respect to threat modelling characteristics for cloud computing and challenges in healthcare.

III. Background

A. Security Challenges in Cloud Computing for Healthcare

According to Shostack et al., cloud computing security challenges can originate from various sources [10]. Using the attacker grouping approach, threats can emanate from both CSP and tenants' sides. The CSP related threats could include insiders who are staffs of the CSP thrust boundary. These insiders may intentionally or accidentally attack tenants or become victim of an attack. Security issues can also originate from all the tenants and other users of the cloud system. Tenants malicious behavior can result in blacklisting effect [9], [10]. Tenants with certain user privileges (such as IaaS and PaaS) could execute malicious codes and the consequences can affect co-tenants and CSP. Additionally, the CSP could be directly targeted by some tenants. Further to this, CSP can face compliance issues. For instance, for CSP to host sensitive applications such as health related or Payment Card Industry (PCI) applications, it is suggested that the CSP must comply with these organizations' requirements. Tenants may also face litigation related issues. For instance, if a tenant requires to know some information about their data which has been held in cloud computing systems for some purposes, they might not be able to get precisely what they want. Data stored on private cloud is more legally protected than data with third party CSP. Forensic response can also be an issue for the tenants similarly to legal related issues. Another source of issue in cloud computing is the usage of mobile devices much like other computing devices [10], [11], [66]. Device loss and the possibility for an adversary to access resources illegitimately is deemed to be a major concern.

According to Cloud Security Alliance (CSA) [11], the security guidance for key areas in cloud computing include data lose or leakage, account or service hijacking, insecure interface, denial of service [12] and malicious insider. Other areas of cloud computing which require attention are data breaches, abuse of cloud services [13], [14], insufficient due diligence and insecure VM migration [14], [15]. Cloud computing is also characterized with various vulnerabilities relating to both technological and human aspects. Some of the common vulnerabilities in cloud computing are session riding, virtual machine escape [14], obsolete cryptography, unauthorized access to management interface [7], internet protocol and data recovery. Additionally, metering and billing Systems and Vendor lock-in are some of the security concerns in cloud computing [14]. In vendor lock-in, for instance, a healthcare organisation could move its IT operations to a cloud provider and subsequently realised it can cannot easily move in the future to a different provider without substantial costs, legal constraints, or technical incompatibilities [62], [66].

In a dynamic and distributed network environment characterized with many users, resources and omnipresent electronic devices, there is a need to adopt to appropriate threat modelling methods to adequately identify related threats and vulnerabilities for efficient measures. The aim of this study was therefore to present the state-of-the-art threat modelling methods that can be used to effectively analyse and identify information security related threats in cloud computing. In this section, the overview of cloud computing and the research problem was presented. This was followed by a presentation of threat modelling methods in Section 2 as the state-of-the-arts. The Section 3 presented the the methods used. In Section 4, the findings and gap analysis of the cloud related methods which were found in the state-of-the-art, were presented and compared. A discussion and conclusion on the state-of-the-art were presented in Section 5.

B. Overview of Cloud Computing

Cloud Computing arose from parallel computing, through grid and utility computing [2], [3], [16]. Parallel computing involves the simultaneous use of multiple homogeneous process-

ing elements in solving a scientific problem. The problems are usually broken down into smaller tasks which are then solved at the same time with multiple processors [2], [17], [18]. The purpose is to safe time, money and to overcome complex tasks while efficiently utilizing the computing power. Application areas include military, energy exploration, data bases and data mining, real time simulation of systems, advanced graphics, augmented reality and virtual reality [2], [17], [18].

Grid computing is a form of parallel computing which uses a network of computers with many CPU cores spread across multiple locations to execute a task instead of the usage of many CPU cores on a single machine [2], [19]. Grid computing is a decentralized service, involving multiple computers with heterogeneous operating systems at different physical locations [2], [19]. Utility computing aimed towards providing resources to clients in a scalable fashion based on the clients' demands and this translates into corresponding scalable pricing [2], [20]. Basically, utility computing maximizes resource usage while minimizing cost of service provision [2], [20]. SaaS is mostly suitable for SMEs to use advanced technologies at lower costs [2], [20]. SaaS involve delivering application software over the internet at flexible packaged payments for license and maintenance fees [2]. Within cloud computing, edge computing is a distributed paradigm in which data storage and computing power is moved closer to the devices or data sources [60] while fog computing is a form of cashing which enables devices to access and process data within the local network when internet is unstable [61].

C. Features and Models of Cloud Computing

The features of cloud computing can be categorized into physical and operational or functional features. The physical features include client side, internet, distributed servers and data-centers [2], while the operational or functional features include on-demand self-service, resource pooling, elasticity, measured service and ubiquitous network access [21]–[23].

The clients include computing resources, such as hardware or software, which are dependent on cloud computing and are being used by the end users for service delivery [2], [22]. The clients can be specifically designed for the cloud and therefore becomes useless without the cloud computing.

The computers may include thin clients, mobile devices and thick clients. The software which are being used by the cloud users are usually hosted in several servers known as the data center [2], [22]. A data center consists of a large group of networked servers, either in a room or building, housing the servers for remote processing and storage or distribution of large amount of data [2]. A server can contain many virtual machines (VM) which the number of VM per server depends on the speed and memory size of the host among others. For resilience such as reliability, availability and fault tolerance, servers can be distributed across geographical locations. The distributed server feature of the cloud computing also helps in scalability [2], [22].

The operational features include elasticity, measured services, ubiquity, resource pooling and on-demand self-service [2], [22]. Elasticity feature defines the property of cloud computing which enables scaling up or down of the unlimited resources to meet the needs of the users. Measured Service is the ability to measure exactly, the cloud services usage per clients despite the shared pooled resources by many clients. Ubiquitous Network Access feature provides access to the cloud computing resource on the network which can be accessed by different type of clients such as mobile phones, laptops and desktops at different location. Resource Pooling enables cloud provider to provide services to the subscribers through a multi-tenant type. Cloud computing resources are assigned and reassigned, following the subscriber needs. On-Demand Self-Service enables a cloud user to use cloud services as needed and without human interference.

Cloud computing models can be categorized into business models and deployment models [2], [22]. The business models include SaaS, IaaS and PaaS while the deployment models include public, private, community and hybrid cloud models. SaaS model enables the provider to provide software services to their clients or subscribers. The clients can rent the software from the providers via internet and the software should be able to react to the clients' interface to appear as if the software belong to him only despite other client are using the same software. PaaS enables users to develop their own applications by renting the development environment and toolkits from the cloud computing providers. The development toolkits are usually accessed from the cloud by the developers through the web browser. Other resources such as operating system, processors, memory and storage of the application development files are provided by the cloud computing provider [2], [24]. CSP also provide the infrastructure as a services (IaaS) to clients and the usage time are quantified per CPU utilization per hour, usage of storage and data transfer rates [25]. There exist other cloud services such as app's stores, online games and electronic books but this study focused on the IaaS, PaaS and SaaS models.

Each of the different deployment models of cloud computing (such as the public, private, community and hybrid) have its related security repercussions. The public cloud computing model is usually patronized by the general public as pay per use or use for free arrangement. Example of such cloud providers include Amazon's AWS, Microsoft Azure and Rackspace Cloud Suite [2], [4]. Private cloud computing model is usually provided by an organization behind its firewall for accesses of members of the organization only. The services of the private cloud computing model are usually restricted from public access and the IT Network administrators of the organization's data center are usually the cloud providers. The community cloud computing model is a kind of public and private model, which is setup for a specific purpose. Essentially, the advantages of cloud computing are transforming the way cloud computing users work. Users can globally access all their programs and documents from any computer via the internet such that the user is no longer tied to a single computing device such as a particular laptop computer to access data and resources. In addition, cloud computing enhances group collaborations, since all group members can flexibly access the shared resource in the cloud from wherever they are located but effective security measures are required [24].

D. Healthcare Related Information Security Threats, Threat Modelling Methods and their Related Characteristics

Threat modeling is the use of methods to help in thinking, identifying and enumerating possible risks and threats [6], [14], [26]–[29]. Threat modelling also helps in the identification of the lack of security controls for mitigating risks [6], [14], [26]–[29].

Various type of threat modelling which were identified in this study include Attack Tree (AT), Attack Graph (AG),Attack Surface (AS), Practical Threat Analysis (PTA) , Threat Model Framework for Personal Network (PN), STRIDE, Threat modeling in pervasive computing paradigm (TMP), [6], [14], [26]–[29], [50] and LINDDUN (Likability, Identifiability, Non-Repudiation, Detectability, Disclosure of Information, Unawareness and Non-Compliance) [30] as defined below:

Attack trees provide a structured way of describing [10] threats and vulnerabilities of a system, with regards to varying attacks and shows the possible attack paths which attackers can follow to be able to compromise the system [26].The attacks are represented against the system in the form of a tree. The goals are placed as root nodes while the leaves nodes represent the different ways of achieving the goals. Attack trees are widely used for security modelling and analysis in the cloud computing. The general steps involves in the attack tree include [26]:

- Specifying the attacker's main goal and this is termed as root node. An Example could be for the attacker to access assets such as data in the cloud.
- Decompose the main goal into sub-goal which is termed as leave nodes: Example include attack repository, get credentials through social engineering; attack certificate or Attack Web portal.
- Continue to divide the sub goals into stepwise subtasks.
- Assign the leave nodes with attribute values.
- The security of the goal can then be computed after all nodes are computed

Attack graph adopts to graphic view of all attack paths including attack point to attack target [27]. Attack graph assesses network configuration and vulnerability information of network by obtaining the entire dependency interactions of the information. Attack surface captures software features that have the tendency to contribute to vulnerabilities. Some of such features includes entry and exit points communication channels, and untrusted data items.

TAM is based on business objectives [14] which involves the business needs and issues that must be met by the system. This is followed with system components. The components include application architecture, user roles, trust level, authentication mechanism and external dependencies. The third level involves generating threats and classifying them into CIA traits. Finally, each threat is assessed based on business need.

PTA process involve identifying assets and their related financial values. The PTA provides tools to enable providers to obtain the value of assets and potential level of damage that can be caused by the adversaries. This is followed with the identification of vulnerabilities , the assessment of risks of threats and specifying the risks mitigation strategies of the system. The identification of vulnerabilities is based on the assessment of the architecture of the system and different type of users.

PN is a user centric network which consists of users' network devices known as personal devices. The network composed of applications, telecommunication, environment and services for the users. The first step in PN framework focused on describing a use case to include everything in the network from the users' perspective. The second step involve gathering network requirements from use case diagrams, network architecture, environments and technologies. In the third stage, data flows are clarified using UML sequence diagrams and the determination of actors and devices involves. Asset identification and determination of all threats were subsequently followed. Vulnerabilities were then identified, and the risks were therefore determined from the threat and vulnerabilities profiled. Finally, the threats and vulnerabilities were rated.

STRIDE is a threat modelling framework which is focused on identifying threats relating to spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service and Elevation of privileges [14]. Various sources [10], [31]have indicated that STRIDE framework can be used based on per-element of the cloud system or per-interaction with the cloud system. STRIDE-per-Element involves enumerating various elements of the system and assessing for the specified threats around each element. This approach simplifies the finding and identification of threats which are associated with each element. But STRIDE-per-Interaction identifies threats based on the origin, destination and interaction attributes of elements, which also makes it easier to understand related threats. STRIDE approach generally provides guidelines as what threat to look for and where they can be found. STRIDE seeks to identify and mitigate threats and vulnerabilities through the reduction of the cost of entire development process. The STRIDE based modelling follows five systematic steps beginning from, classifying assets, obtaining the overview of the system by creating DFD. Threats are then modeled and identified. The identified issues are addressed and the threats are eventually ranked by using Damage, Reproducibility, Exploitability, Affected Users, and Discoverability (DREAD).

With Threat modeling in TMP, all cloud computing users' roles are identified in relation to their service usage and authentication mechanism in the first step. The second stage involve identifying security domains to understand how users interact with applications within the domain. Trust levels were subsequently identified so that users can access resources based on their level of trust level. Vulnerability identification is done in the next level so that known vulnerabilities can be mitigated while unknown vulnerabilities are managed in a manner that would protect the system. Risk evaluation are performed in the next stage to provide knowledge for appropriate risk management strategies.

In the area of privacy, the LINDDUN (Likability, Identifiability, Non-Repudiation, Detectability, Disclosure of Information, Unawareness and Non-Compliance) privacy threat modeling framework has recently gained attention in the privacy

community [30]. LINDDUN provides systematic support to elicit and mitigate privacy threats. The strength of LINDDUN is its systematic approach in guiding the analyst through the privacy assessment exercise effort, combined with its extensive privacy knowledge base.

Within the healthcare domain, it is critically important to correctly identify patients and correctly map them to their health related records [55]. This compels healthcare organizations to collect and store detailed amount of personal identifiable healthcare data from each patient, making it rich for committing identity theft [55]. Therefore, in the development of health information systems, security practice relating to threat modeling should adequately be adopted to rigorously protect the systems. Some of the dimensions of threat areas in healthcare have been identified to include malicious users, misuse of information system resource, communication interference, damages, failures, errors and threats relating to theft. Others include repudiation and attribution, misuse of system resources, legal and regulatory requirements as shown in Table I. These related threats are not unique to health information systems, however, threat impact of the aforementioned can have life-threatening impact on the vulnerable patients [55].

IT staff in healthcare who have the responsibility of managing the development of healthcare systems will need to comply with various security practices [54] including threat modeling to identify attackers, resources or assets can be compromised and their mitigation strategy.

In summary, threat modelling methods for cloud computing in healthcare should have the outlined abilities [10], [14], [32] as shown in Fig. 1. The characteristics includes:

- Identifying and classifying assets: IT assets within healthcare includes but not limited to data, software or hardware which are being used by the healthcare providers. All these assets need to be identified and categorize for efficient security management. For instance, assets can be classified based on different values, damage costs and trust levels to enable prioritization for countermeasures.
- Identifying users and threat agents: For each user domain, different users with different access controls have different trust. Example, different healthcare staffs such as administrators, all authorized users and unregistered users have different trust levels [31]. Additionally, a threat can adversely act on assets. Threat agents includes unauthorized entities, system administrators and other authorized users. Natural events including flood, earthquake, and fire can also be a threat agent. All these need to be identified for effective security management [35].
- Establishing trust level and User's Role: Establishing thrust levels and linking it to established healthcare professionals' roles with authentication, authorization and access control mechanisms, enhances the confidentiality, integrity and availability of the assets [36].
- Identifying Security Domain: Different user domains have different security levels with their respective different kind of information types. Therefore, it is important to separate domains for security reasons and isolate risks based on the identified security domains [37].
- Identifying Threats and Vulnerabilities: In adopting cloud computing, healthcare organizations are inadvertently outsourcing their computational resources on virtual domains. The huge and sensitive amount of data can be damaged by threats from different resources including employees' activities and malicious attacks. So threat and vulnerabilities to the data need to be assessed [38].
- Ranking and measuring vulnerabilities: This enables organizations to identify various weaknesses around their information systems and provides them with the knowledge to prioritized the implementation of countermeasures [39], [40]. The Common Vulnerability Scoring System (CVSS) [41] is one of the methods for rating IT vulnerabilities. The CVSS has base metric, temporal metric and environment metric for evaluating security vulnerabilities. Base metric assesses the basic attribute of vulnerabilities. The environment metric evaluates vulnerability metrics that are associated with the environment and temporal metric considers dynamic aspect of the vulnerabilities.
- Ranking and measuring threats: Ranking and measuring threats is used to assess the risk posed by the identified threats having taken various factors into consideration. The first step is to consider the risk level which is posed by each threat. So, Microsoft Threat Modeling introduces Damage potential, Reproducibility, Exploitability, Affected users and Discoverability (DREAD) as a type of threat ranking and measuring method as follows: The equation that is used to compute a risk value has been shown as follows [32].

$$Riskvalue = (D + R + E + A + D)/5 \quad (1)$$

The risk value is between the range of 0 and 10, so that the risk increases with higher risk values. Similarly, the Open Web Application Security Project (OWASP) measures risks based on the likelihood of the threat and its related impact as follows:

$$Risk = Likelihood * Impact \quad (2)$$

Applying Common Vulnerability Scoring System to a business process cannot fully deliver optimal security measure in dynamic environments [33]. VRank and hybrid ranking have been developed to satisfy business requirements. The VRank [34] as a dynamic framework is able to integrate existing vulnerability databases with the specific detail context of business process. Accordingly, the VRank provides more specific vulnerability assessment for SOA. The hybrid ranking model proposes a combination of CVSS rating and numerical estimation of vulnerability that influence the global network [33]. By developing dynamic environments estimating vulnerability in an influential level and aggregating other rating models with high-level rating technics provides a hybrid model to precise measurement of vulnerabilities more economic and efficient.

TABLE I. POSSIBLE THREATS TO HEALTHCARE [55]

Related Threats	**Description and Possible harm**	**Instance**
Threats resulting from errors	These include operation errors, maintenance errors, user error and accidental miss-configurations or miss-routing. Operation errors can result in intentional disclosures of confidential information. Maintenance errors such as miss-configurations of software can be committed by staff members or third party employees who are responsible in maintaining the systems. If the miss-configuration has to do with authentication and authorization, serious consequences can be uncured. Legitimate accesses for therapeutic purposes can be denied and this can lead to further worsening of health conditions and lost of lives. Conversely, miss-configuration of authorization and authentication mechanisms can enable broad access to large and wide number of unauthorized users and this can lead to huge compromise of privacy and security over the cloud computing.	According to [68], [69], errors such as misconfiguration due to human error in cloud infrastructures has increased by 424%. Example, in February 2019, the University of Washington Medicine exposed the information of approximately one million patients due to accidental removal of website server protections [68], [69].
Communication Interference	This includes both the interception and infiltration of the communication of healthcare information systems. Communication infiltration occurs when the adversary tempers with the normal flow of communication and this can lead to denial of service. Additionally, if the message in the communication channel is not protected during transmission, the confidentiality of patients records that are involved, will be compromised if the message is intercepted. Since access to healthcare records for therapeutic purposes is time sensitive, denial of service can also have life-threatening consequences.	In a recent ransomware attack at Duesseldorf University Clinic in Germany, the medical records of a patient were not timely available in an emergency case and the patient loss her life as a result [63].
Unauthorised users	These include insider and outsider masquerades and other unauthorised users who illegally accesses healthcare information and breaches confidentiality, integrity and availability of the system [43]–[45], [52]. . Cloud computing tend to host a number of tenants who are their internal users. Not all internal users of the cloud are necessarily, the internal users of the healthcare facility and their security indiscipline can tend to negatively affect healthcare. Clearly, the healthcare faces broader scope of threats from unauthorised users. Example include all the internal users of the cloud computing system and the exclusive users of the healthcare organization such as a patient who overtakes an unattended workstation in a physician office and accessing the data. Also, when a healthcare professional is to takeover a shift from a colleague, due to inconveniences, the first user may fail to logout in order for the colleague who is taking over to continue work without going through the pain of logging in.	Insider masquerade include an instance in the UK, where DR. Harold Shipman tried hiding a number of records of his patient who is a notorious murder [45].
Threats relating to damage	This includes willful damages by insiders, outsiders, terrorists, and the introduction of damaging and disruptive software or malicious codes. Example include threat of information security systems by co-tenants, disgruntled staff, patients or relatives which leads to availability problem of the CIA.	Terrorists for instance can target large installations of healthcare systems of which the impact could be huge. Example include data breaches in Helse Sør-Øst RHF (Health South-East) of Norway of which the focused was on patient records and the health service's relationship with Norway's armed forces [65]
Threats of failures	Failures include connection failures, technical failures of hosts, storage systems, and network infrastructure, network software failure, application software failure. Others include environmental support failures (eg power failure, failure from natural disasters and man made disasters), and staff failures or shortage. Such impact could compromise with the CIA of the information system in various ways	For example, it was suspected that, the failure to update legacy windows XP resulted in a compromise of about 3million patients records [65]
Theft	Threat of theft include insiders and outsiders who can steal equipment or data in order to sell or disclose to others. Compromising confidentiality and integrity	A laptop of a vendor of Health Share was stolen. The laptop contained 654,000 patients records consisting of names, contact details, date of birth and medical ID numbers of patients [64].
Repudiation and attribution	When there are issues, the ability to determine in time whether the issue is originating from the cloud provider, or the hospital end in a timely manner is a concern.	Example includes a medical record alteration by a cloud maintenance officer who uses the account of a healthcare staff. As the logs of accesses are controlled by the cloud provider, their timely access by the hospital could be a problem. Also, how to ensure non-repudiation even if the logs were provided to the hospital is problematic.
Misuse of system resources	This include scenarios where users tend to use information systems and services for personal purposes. Such activities include misuse of internet and computing resources which an tend to threaten the availability of these systems for healthcare functions.	Example, healthcare workers may tend to be downloading or watching large files of videos on the hospital's network, slowing down the network access for healthcare purposes [52].
Legal and regulatory requirement	Regulatory requirement such as GDPR or HIPAA need to be considered. Example, if sensitive data are to be processed outside EU under the GDPR, there can be a constrain if some of the servers and data processing are carried out outside EU.	Example, if healthcare data from EU is to be stored in third countries, those countries have to be under Adequacy Decision or adopt to additional safeguards to comply with the GDPR else, the DGPR will be violated [67]–[69].

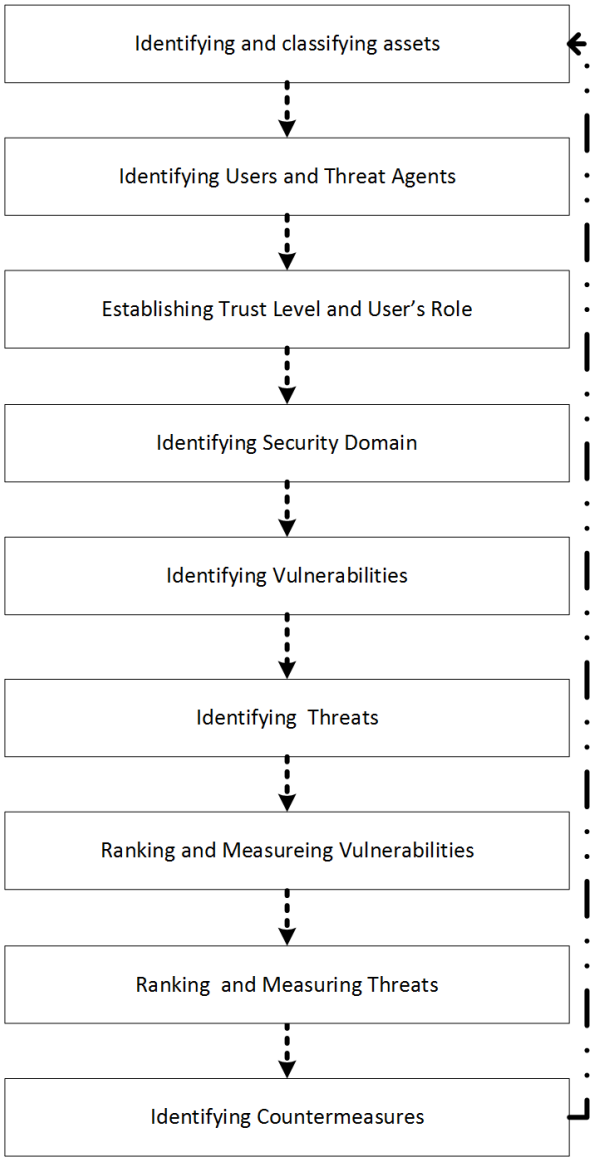


Fig. 1. Characteristics of Threat Modelling for Cloud Computing [31], [32]

- Identifying countermeasures: Information security countermeasures include actions, devices, procedures, techniques and other measures which are adopted reducing vulnerabilities towards protecting assets against threats [42]. The appropriate countermeasures are identified based on the measured threats and vulnerabilities. Identifying new assets, vulnerabilities and Threats [31].

Threat modeling and its countermeasure should not be a one-time event but an ongoing process in cloud computing in healthcare. This will enable the identification of new assets and associated threats and vulnerabilities as shown in Fig. 1. Additionally, users of the system ought to be conscious of new vulnerabilities and their countermeasure. Mitigation procedures need to be adopted for the identified new vulnerabilities to keep the system secure against the new threats [31], [32]. Ideally, the threat modeling methods for cloud computing should systematically cover all the outlined characteristics as shown in fig. 1, from identifying and classifying assets, to identifying countermeasures. Furthermore, the process of detecting new assets, threats and vulnerabilities should be a looping process since cloud computing environment is a dynamic system with evolving potential set of new threats and vulnerabilities.

In a related study, some of the threat modelling methods which were assessed include Microsoft threat modeling with STRIDE, TAM, PTA, Personal network threat modeling (PN) and pervasive computing threat modeling. None of the reviewed threat modeling methods was found to be fully suitable for cloud computing systems [14]. But pervasive computing threat modeling had most of the threat modeling traits (as shown in Figure 1) of cloud computing.

Similarly, attack surface, attack trees and attack graphs, were used to undertake threat modeling exercises in cloud computing infrastructures [26], [28]. Additionally, Cheng et

al., generated attack graph model [27] of a cloud computing system and subsequently suggested two security evaluation metrics after combining Markov Chain with attack graph. In the study [27], various assumptions relating to cloud computing system were made.

Based on the effective way of how STRIDE framework categorizes threats, [6] Hong et al. identified cloud related threat through literature survey and categorized them with the STRIDE framework [6]. Basically, the various attacks on cloud computing were categorized with OWASPs attack categories. The attacks were further categorized into cloud computing related threats with the STRIDE framework and the cloud components were mapped to the possible threats. The study was used to propose traceability method for identifying cloud related threats. Furthermore, Yahya et al. adopted STRIDE with a three step threat identification features thus characterizing the system, assets and access identification and threat identification [29], to analyse threats in a cloud storage scenario. In this scenario, the adversary's main target was on the software related services such as SaaS cloud resources. Cloud computing related threats were subsequently identified and classified with the STRIDE method [29]. The threats which were identified in the study were mapped to security requirements objectives.

IV. Method

A literature search was conducted in Google scholar, IEEE Explore, ACM Digital Library and Elsevier for cloud related threat modeling methods. Only threat modeling methods which were assessed for cloud computing, were included in the study. The threat modeling methods were then assessed for their suitability towards effective identification and analysis of threats in cloud computing in healthcare context. Additionally, the threat modeling characteristics were also identified from related studies [10], [14], [32] and these were analysed in healthcare scenarios as shown in table II. The identified characteristics were mapped to cloud relating threats in healthcare context which have been identified in table I. This depicted the role of each of the threat modeling characteristics , in healthcare scenarios as shown in Table II. These characteristics were used as benchmark in the comparison of the identified cloud related threat modeling methods in this study as shown in Fig. I and Tables II, III and IV. The findings and gaps are presented under Section IV.

In the review assessment, each of the threat modeling methods was reviewed against the identified cloud computing threat modelling traits as shown in Table II, Table III and Table IV. Each of the methods that supports any of the particular cloud modeling characteristics, was assigned with a value of "Yes" in its cell in Table IV. However, in the review, if a particular threat modeling method does not support any of the outlined cloud computing threat modelling characteristics, then a value of "NO" is assigned to that threat model. Findings of the entire assessment is presented in Table IV.

V. Findings

The suitability of the identified threat modelling for assessing threats and vulnerabilities in cloud computing were compared as shown in Table IV.

The threat modelling methods were compared against threat modelling characteristics of cloud computing as shown in Table IV. On Table IV, 'Yes' means the method supports the threat modelling traits, 'No' means the mettahod does not support the threat modelling characteristics. In summary, TMP supports almost all the provision for all cloud computing threat modelling characteristics except establishing user's role, scanning domain security and ranking and measuring vulnerabilities [14] and followed by AT as shown indicated in Table IV. On the contrarily, TAM methods provide for only identification and classification of assets and ranking and measuring vulnerabilities [14]. Similarly, STRIDE method has provision for assets identification and classification of assets, identify threats and identify vulnerabilities [6].

A. Gap Analysis

The various studies in threat modeling methods which were explored for cloud computing in healthcare, are shown in Table III and Table IV. In following the characteristics of cloud computing models, there are various gaps in the threat modeling methods. For instance, attack trees and graph mostly support in identifying a comprehensive attack related threats to a system, however, there are no systematic methods to determine parameter values for each node, especially in an Attack Tree [26], [27]. Additionally, attack tree and attack graphs methods are deemed challenging task, particularly for large sized networked systems [6], [27], [46]. This is because, the number of possible attacks grow exponentially with the growth rate of the number of hosts [46]. Also, attack tree is still a relatively high-level concept, without details about specific ways for exploiting a resource. Similarly, Attack surface heavily relies on experts' knowledge of the system features and knowledge past attacks on the system, using these features [50], [51].

Furthermore, TAM framework does not have features for assets determination and identification of vulnerabilities [14], [46]. Meanwhile, cloud computing have multiple assets from the side of both cloud provider and tenants. Therefore the lack of identification of asset and vulnerability assessment is deemed to be a major shortfall for threat molding cloud computing environment.

Additional, PTA model has no provision to estimate the cost of vulnerabilities [10], [14]. Cost of vulnerabilities is computed with cost of down time, replacement and systems' downtime. The lack of cost of vulnerability hides knowledge and idea of avoiding different kind of threats. Due to pervasiveness feature of cloud computing all known and unknown vulnerabilities need to be assessed and planned for but there is a lack of vulnerability ranking mechanism in PTA. Although, PTA rank threats, but it does so only on the bases of financial value and not functional, technical, strategic or reputational value. In cloud computing, tenants can be users of multiple domains which require the need to comprehensively rank threats from different factors and levels of security such as user roles but not only financial bases.

Furthermore, there is a lack of update model for new threats and vulnerabilities in PN model [14]. But in cloud computing, it is vital to frequently assess new threats and vulnerabilities since users of cloud computing have access to

TABLE II. CLOUD COMPUTING THREAT MODELING CHARACTERISTICS IN HEALTHCARE SCENARIOS

TM Traits	Healthcare Context	
	Scenario	**Remarks**
Identifying and classifying assets	In EHR scencario, the main assets include the sensitive healthcare records [70]. Others include mobile devices, user credentials, and the network of the EHR system.	These assets face the related threats [70] as shown in Table I.
Identifying Users and Threat agents	The users include the healthcare professionals who accesses the system for therapeutic purposes, the paramedical staff, temporal staff, service providers such as contracted service personnel eg system software and hardware engineers [55]. Threat agents include insider masquerades, hackers, and natural disasters, patients or subjects of care.	For instance, a subject of care can access an unattended workstation system thereby, compromising the CIA of the system [55].
Establishing Trust level and user's role	Low trust level require minimum or optional, to no security protection mechanism [71]. However, if the highest level of trust is compromised, it involves, loss of data resulting in long-term and permanent damage to the hospital, patients, or groups of individuals. Therefore, high trust level needs the incorporation of protection mechanisms designed into the system to be commensurate with the expected risk of exposure [71].	Example, a software engineer who have access to the entire EHR system in production have access to the highest trust level, requiring for appropriate countermeasures [71].
Identifying security domains	Security domains are list of objects with similar security requirements that can be access by objects including healthcare professionals and end devices, known as users [72]	Nurses may be in the same security domain which is different from the security domain of healthcare application developers [72].
Identifying threats	All possible threats as outlined in Table I are then identified.	Instances of related threats are shown in Table I.
Identifying vulnerability	Possible weaknesses that can lead to attacks are identified	These include web application vulnerabilities (such as SQL injections, Cross site scripting [72], [73]), cloud related vulnerabilities (malicious insiders such as cloud maintenance engineers) and vulnerabilities relating to end users and devices.
Ranking and measuring threats	Metrics of the threats are assessed and ranked for prioritization regarding counter measures.	
Ranking and measuring vulnerabilities.	Vulnerabilities are also assessed and prioritised.	
Identifying counter measures	Possible mitigation are kept identified [70].	Possible countermeasures include multi-layer countermeasures, multi-factor authentication, access revocations etc, fail safe default, mechanism, least privileges [70].
Defining new assets, threats or vulnerabilities	Due to updates, addition of resources and system upgrades, new assets, threats and vulnerabilities need to be identified	For instance, if a new module or department such as radiology, is added, the process need to be repeated to determine new assets and threats.

TABLE III. THREAT MODELING METHOD

No.	Threat Modelling Method	Reference
1	Attack Tree (AT)	[26]
2	Attack Graph (AG)	[26], [27]
3	Attack Surface (AS)	[26]
4	Microsoft's threat analysis and modelling (TAM)	[14]
5	Practical Threat Analysis (PTA)	[14]
6	Threat Model Framework for Personal Network (PN)	[14]
7	STRIDE	[6], [14]
8	Threat modelling in pervasive computing (TMP)	[14]
9	LINDDUN	[14]

TABLE IV. COMPARISON OF THREAT MODELLING METHODS FOR CLOUD COMPUTING

TM Traits	Threat modelling methods								
	AT	**AG**	**AS**	**TAM**	**PTA**	**PN**	**STRIDE**	**TMP**	**LINDDUN**
Identifying and classifying assets	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No
Identifying Users and Threat agents	No	No	No	No	Yes	No	Yes	No	No
Establishing Trust level and user's role	No	No	No	No	Yes	No	Yes	No	Yes
Identifying security domains	No	No	No	No	No	No	No	Yes	No
Identifying threats	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Identifying vulnerability	Yes	Yes	Yes	No	No	Yes	No	Yes	No
Ranking and measuring threats	Yes	No	No	No	No	Yes	No	Yes	No
Ranking and measuring vulnerabilities	Yes	No	No	Yes	No	Yes	Yes	No	No
Identifying counter measures	Yes	No	No	Yes	No	No	No	Yes	No
Defining new assets threats or vulnerabilities	Yes	No	No	Yes	No	No	No	Yes	No

different services and domain which exposes them to different type of attacks. Also, in TMP, cloud based assets management methods need to be considered instead of the traditional asset management approach in threat modeling because assets and resources in the cloud are dynamic.

STRIDE was also assessed to be limited for its effective application in cloud computing [6], [14]. Cloud computing is dynamic and pervasive, and can not be threat modeled with static frameworks such as STRIDE. Furthermore, defining new threats, identifying vulnerabilities and privacy related issues have not also been considered in STRIDE framework thus limiting its applicability in cloud computing. According to

TABLE V. THE FOLLOWING ABBREVIATIONS ARE USED IN THIS MANUSCRIPT:

AT	Attack Tree
AG	Attack Graph
AS	Attack Surface
CIA	Confidentiality, Integrity and Availability
CSP	Cloud service provider
CVSS	The Common Vulnerability Scoring289System
TAM	Directory of open access journals
TMP	Threat modelling in pervasive computing
PN	Threat Model Framework for Personal Network
PTA	Practical Threat Analysis
STRIDE	spoofing, tampering, repudiation, information disclosure, denial1 of service and elevation of privilege
LINDDUN	Likability, Identifiability, Non-Repudiation, Detectability, Disclosure of Information,Unawareness and Non-Compliance

Shostack et al., STRIDE only gathers high level security requirement and not details of what can go wrong in the system [2], [10]. Though, LINDDUN model fill in the gap on privacy requirement assessment, it's focus is generally narrowed to obtaining privacy requirement by mapping data flow diagrams of the system to its application scenarios and related threats [30], [50].

VI. DISCUSSION

Though cloud computing paradigm is providing crucial services [5], [6], its distributed nature (ubiquity, elasticity, many user types and resources) has made it susceptible to attacks. To enable healthcare sector to be able to adopt to the usage of cloud computing services, there is the need to establish mechanisms towards identifying and mitigating its related threats. Based on this, threat modeling methods for cloud computing were explored for their fitness for healthcare purpose as presented in Table II and Table IV. Threat modeling methods which were found to be commonly used for cloud computing include Attack Tree (AT), Attack Graph (AG) and Attack Surface (AS), S Microsoft's threat analysis and modeling framework (TAM), Practical Threat Analysis (PTA), TRIDE, Threat Model Framework for Personal Network (PN), TMP [6], [14], [26]–[29] and LINDDUN. From the findings as shown in Table IV, none of the existing threat modelling methods in the review could be used to completely assess threats and vulnerabilities to meet all the characteristics of threat modelling methods for cloud computing. A related study which assessed a limited number of threat modelling methods (without assessing AT,AS and AG) also observed similar results [14]. This poses a major deficiency in identifying and mitigating cloud computing related threats for healthcare sector. Table IV clearly depicts that if any of the identified threat modelling methods is used for cloud related threat modelling, some of the threats and vulnerabilities relating to the outlined characteristics in Table IV would not be covered. For instance, AG, AS, TAM, PTA, PN and STRIDE have no provision in their framework to determine new assets threats and vulnerabilities in cloud computing. This drawback does not support the Confidentiality, Integrity and Availability (CIA) of cloud computing since cloud computing is characterized with omnipresence, elasticity and dynamic environment.

Furthermore, the specification of threats in STRIDE makes it limited for its effective application in cloud computing [13]. Cloud computing is dynamic and pervasive and cannot be threat modelled with static frameworks such as STRIDE. Additionally, defining new threats, identifying vulnerabilities and privacy related issues have not also been considered in STRIDE framework thus limiting its applicability in cloud computing. Therefore, cloud computing requires threat modelling methods which can help identify new assets and their related threats and vulnerabilities in a dynamic network such as cloud computing. Other cloud threat modelling characteristics which were less considered by most of the various frameworks include establishing user's role and estimating trustworthiness as shown in Table IV. Trustworthiness is a critical characteristic in distributed system which need to be estimated by cloud computing related threat modelling methods. In distributed computing paradigm, two or more systems with heterogeneous features and environment are combined to accomplish a given task. Assessing the security status of the individual systems to integrate them into cloud computing system would contribute towards enhancing the CIA of cloud computing [14], [47] for healthcare.

The flexibility, ubiquity and reliance of cloud computing can enable healthcare staff to access various healthcare services at different times and places. Each user sometimes has multiple roles such as an administrator with high level privileges in one domain. In another domain that user can be an ordinary user with low level privileges. Therefore, proper authorization, authentication and access control mechanisms are required to enhance the CIA of cloud computing [14], [48]. This can also contribute to complexities in scenarios where the cloud computing logs are to be analysed for users' security practices. Aside this, in the context of public cloud computing where access logs are controlled by the cloud service providers (CSP), the flexibilities for healthcare providers, as tenants, to analyse their access logs for security measures would be highly limited. In general, if tenants such as healthcare providers, require to know some information about their data which has been held in third party CSP , they might not be able to get precisely what they want at the appropriate time and this might undermined critical decision making in healthcare. From Table IV, threat modeling in pervasive computing (TMP) and Attack Tree (AT), comparatively support a higher number of cloud computing threat modelling characteristics. TMP framework support all cloud computing modeling characteristics except establishing user" role, scanning domain security, ranking and measuring vulnerabilities. TMP framework exhibited high efficiency for cloud computing based on the background that it was designed to incorporate pervasive computing environment issues [14]. For TMP to fully address cloud computing related threat modelling characteristics, it needs to be complemented with AT, AG and PTA or STRIDE as shown in Table IV. Similarly, Attack Tree (AT) also addresses more of the cloud computing

threat modelling concerns but require to be completed with TMP, AG and STRIDE or PTA as shown in Table IV.

Various attack paths which can possibly be followed by attackers, can be identified by attack trees however, it does not provide detailed specific ways of exploiting resources [25, 26]. Attack graphs can be used to augment for this shortfall. TMP can also complement AT to be able to establish trustworthiness and development of counter measures [26], [27] as shown in Table IV. According to the ISO standard for healthcare security (ISO 27799:2016), healthcare organizations usually collect detailed personal information due to the ultimate importance to perfectly identify patients and correctly match them to their health records [52]. Even though other critical sectors such as the financial industries, collect demographic data, personal data, payment card details and social security numbers of their clients, the healthcare sector, additionally, collects healthcare information such as medical history, diagnosis and treatment, insurance claims and payments, bio-data and medical prescriptions. This makes the healthcare sectors' records to be greatly richer and more sensitive than other sectors including the banking and financial industry [53]. This deeply raises privacy concerns in the adoption of cloud computing in healthcare which most of the identified cloud threat modeling methods have not addressed. While adopting a synergy of methods to address cloud related issues for healthcare, privacy requirement methods such as LINDDUN should be incorporated towards identifying and mitigating security and privacy issues.

In terms of the deployment models of cloud computing, private cloud is more secure and currently fit for healthcare since the services of the private cloud computing model are usually restricted from public access and the IT Network administrators of the organization's data center are usually the cloud providers.

A. Spatial Consideration in Healthcare Domain

Current requirement in healthcare is to access patients records anywhere and at any time of which cloud computing has the ultimate solution [75]. Cloud computing can fulfil this special need thereby, providing more effectiveness and efficiency in the healthcare sector at a relatively lower cost [74], [75]. For example, cloud computing services can support hospitals in sharing EHR, doctor's references, prescriptions, insurance information, and test results [68], [69], [74], [75]. Due to huge radiology data and sharing needs, many radiology departments are already adopting cloud related methods to lower their storage costs while efficiently providing exchange of images [74], [75]. However, regulatory obstacles, privacy and security challenges are some of the barriers to adopting cloud computing in healthcare as outlined in Table I. Healthcare data specially has strict privacy and security concerns as specified in popular regulations such as HIPAA and GDPR [74], [75]. These regulatory concerns must be completely fulfilled when sensitive healthcare data is to be entrusted onto a third party such as the cloud system. To prevent exposing sensitive healthcare information to unauthorized persons, an effective and efficient security measures should be considered in the aspect of access controls, authentication, authorization, security relating to transmission and storage as outlined in Table I. That is why the threat modeling methods that should be used to threat model cloud computing for healthcare need to have all the outlined characteristics as shown in Table III, to help in identifying detailed threats and vulnerabilities in cloud computing towards providing effective and efficient counter measures. So, based on this work, various threat modeling methods such as AT, TMP, AG STRIDE or TAM can be synergiesed to cover all the threat modeling characteristics to identify detailed threats and vulnerabilities for countermeasures in cloud computing for healthcare.

B. Conclusion

It has become increasingly necessary for healthcare IT professionals to adopt to better methods of assessing the security of cloud computing towards their adoption in healthcare since healthcare data is classified among the most sensitive personal data in which the privacy and security of the data subjects cannot be taken for granted in threat modelling cloud computing for healthcare [49]. As a result, threat modelling methods for cloud computing were compared, with respect to their advantages and disadvantages. After the methods were thoroughly reviewed against cloud related threat modelling characteristics, TMP and AT threat modelling methods were found to support more threat modelling characteristics for cloud computing. Therefore, TMP could possibly be combined with AT, AG and PTA or STRIDE while Attack Tree (AT) was seen to better partner with TMP, AG and STRIDE or PTA. The challenge is that, attack tree and attack graphs are difficult to use for large sized networked systems [6], [46] because the number of possible attacks grows exponentially with the growth rate of the number of hosts [46].

In the future a development of a hybrid threat modelling framework for cloud computing in healthcare need to be considered alongside with risk identification and mitigation, and assessing the method for actual use towards enhancing healthcare security practice. Future studies need to also consider how the threat modeling methods can be incorporated into other technologies such as block-chain, towards enhancing the security and privacy of healthcare systems. Also, threat modelling methods in healthcare context need to be incorporated with privacy related threat modelling activities as found in LINDDUN [30], [50]. Meanwhile, mitigation strategies should assess some of the recent technologies [76]–[78] for adoption in cloud computing including access control measures. Additionally, cloud computing infrastructure are located across different geographical locations. This raises legal and regulatory concerns in healthcare in terms of storing healthcare sensitive data across different geographical boundaries. In the future works, there is a need to explore for the legal and regulatory ways of addressing such threats in healthcare context.

For research articles with several authors, conceptualization,Prosper Yeng Stephen Wolthusen ; methodology, Prosper Yeng and Bian Yeng; validation, Prosper Yeng., Stephen Wolthusen. and Bian Yang; formal analysis, Prosper Yeng.; writing–original draft preparation,Prosper Yeng; writing–review and editing, Prosper Yeng, Bian Yang; X.X.; supervision,Stephen Wolthusen; project administration, Bian Yang.

The authors declare no conflict of interest.

REFERENCES

[1] H.A. Aziz, A. Guled "Cloud Computing and Healthcare Services," J Biosens Bioelectron 7: 220,2006, doi: 10.4172/2155-6210.1000220

[2] M.U. Bokhari , Q. Makki , Y.K. Tamandani "A Survey on Cloud Computing". In: Aggarwal V., Bhatnagar V., Mishra D. (eds) Big Data Analytics. Advances in Intelligent Systems and Computing, vol 654.2018, Springer, Singapore

[3] Böhm, M., Leimeister, S., Riedl, C.H.R.I.S.T.O.P.H. and Krcmar, H., 2010. Cloud computing and computing evolution. Technische Universität München (TUM), Germany.

[4] P. Mell , T. Grance The NIST Definition of Cloud Computing — CSRC. NIST. 2011:1-2.

[5] IBM. Benefits of cloud computing 2019 [updated 2019-05-09; cited 2019. Available from: https://www.ibm.com/cloud/learn/benefits-of-cloud-computing.

[6] J. B. Hong ,A. Nhlabatsi , D. S. Kim, A. Hussein ,N. Fetais , K. M. Khan, "Systematic identification of threats in the cloud: A survey", Computer Networks. 2019;150:46-69.

[7] Healthcare, G. E. "Centricity practice solution going beyond meaningful use." (2010).

[8] N. Kolakowski, "Dell practice fusion to offer medical records system. 2010". Available from:https://www.eweek.com/news/dell-practice-fusion-to-offer-medical-records-system

[9] R. Daman, M.T. Manish, K.M. Saroj,"Security issues in cloud computing for healthcare." In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1231-1236. IEEE, 2016.

[10] A. Shostack, "Threat Modeling: Designing for Security: United States:", John Wiley & Sons Ltd;p. 590,2014.

[11] L. F. B. Soares, D. A. B. Fernandes, M.M. Freire, P. R. M. Inacio, editors. Secure user authentication in cloud computing management interfaces. 2013 IEEE 32nd International Performance Computing and Communications Conference (IPCCC); 2013 6-8 Dec. 2013.

[12] Modi, Krishna, and Abdul Quadir. "Detection and Prevention of DDoS Attacks on the Cloud using Double-TCP Mechanism and HMM-based Architecture." International Journal of Cloud Computing and Services Science 3.2:113,2014.

[13] Hamza, Yasir Ahmed, and Marwan Dahar Omar. "Cloud computing security: abuse and nefarious use of cloud computing." Int. J. Comput. Eng. Res 3.6: 22-27,2013

[14] A. Amini, N. Jamil, A. Ahmad, Z'aba M. R. Threat Modeling Approaches for Securing Cloud Computin. Journal of Applied Sciences.;15:953-67,2015

[15] Hashizume K., Rosado D. G., Fernández-Medina E., Fernandez E. B. An analysis of security issues for cloud computing. Journal of Internet Services and Applications,4(1):5,2013

[16] S. M. Hashemi, A. K. Bardsiri, Cloud computing vs. grid computing. ARPN journal of systems and software, 2(5), 188-194, 2012.

[17] GeeksForGeeks. Introduction to Parallel Computing - GeeksforGeeks 2018 [updated 2018-10-04; cited 2019 02.11.2019]. Available from: https://www.geeksforgeeks.org/introduction-to-parallel-computing/.

[18] Golub G. H., Ortega J. M. Scientific computing: an introduction with parallel computing. Choice Reviews Online, 30(10):30-5637-30, 1993.

[19] Y. Jadeja, K. Modi. Cloud computing-concepts, architecture and challenges. In 2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET) 2012 Mar 21 (pp. 877-880). IEEE.

[20] Ben-Yehuda O. A., Ben-Yehuda M., Schuster A., Tsafrir D. The rise of RaaS. Communications of the ACM. 2014;57(7):76-84.

[21] Lalanda P., Julie M., A., Diaconescu A. Autonomic Computing - Principles, Design and Implementation: Springer; 2013 2013-05-27. 288 p.

[22] Singh, S., Jeong, Y. S., & Park, J. H. (2016). A survey on cloud computing security: Issues, threats, and solutions. Journal of Network and Computer Applications, 75, 200-222.

[23] Baecker R. M. Readings in human-computer interaction : toward the year 2000. 2nd ed. ed. San Francisco: Morgan Kaufmann Publishers; 1995.

[24] Dillon, T., Wu, C., & Chang, E. (2010, April). Cloud computing: issues and challenges. In 2010 24th IEEE international conference on advanced information networking and applications (pp. 27-33). Ieee.

[25] hang Q., Cheng L., Boutaba R. Cloud Computing: State-of-the-art and Research Challenges. Journal of Internet Services and Applications. 2010;1:7-18.

[26] Alhebaishi N., Wang L., Singhal A. Threat Modeling for Cloud Infrastructures. ICST Transactions on Security and Safety. 2019;5:156246.

[27] Cheng Y., Du Y., Xu J., Yuan C., Xue Z., editors. Research on security evaluation of cloud computing based on attack graph. 2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems; 2012 30 Oct.-1 Nov. 2012.

[28] Zimba A., Chen H., Wang Z., editors. Attack tree analysis of Man in the Cloud attacks on client device synchronization in cloud computing. 2016 2nd IEEE International Conference on Computer and Communications (ICCC); 2016 14-17 Oct. 2016.

[29] Yahya F., Walters R. J., Wills G. B., editors. Analysing threats in cloud storage. 2015 World Congress on Internet Security (WorldCIS); 2015 19-21 Oct. 2015.

[30] Gholami A., Laure E. Advanced cloud privacy threat modeling. arXiv preprint arXiv:160101500. 2016.

[31] Amini A., Jamil N., Ahmad A. R., Z'aba M. R. Threat Modeling Approaches for Securing Cloud Computin. Journal of Applied Sciences. 2015;15(7):953-67.

[32] Malik, Nazir A., Muhammad Younus Javed, and Umar Mahmud. "Threat modeling in pervasive computing paradigm." 2008 New Technologies, Mobility and Security. IEEE, 2008.

[33] Zhao F., Huang H., Jin H., Zhang Q. A hybrid ranking approach to estimate vulnerability for dynamic attacks. Computers & Mathematics with Applications. 2011;62(12):4308-21.

[34] Jiang J, Ding L, Zhai E, Yu T. VRank: a context-aware approach to vulnerability scoring and ranking in SOA. In2012 IEEE Sixth International Conference on Software Security and Reliability 2012 Jun 20 (pp. 61-70). IEEE.

[35] Rhee K., Won D., Jang S.-W., Chae S., Park S. Threat modeling of a mobile device management system for secure smart work. Electronic Commerce Research. 2013;13(3):243-56.

[36] Ryan M. D. Cloud computing security: The scientific challenge, and a survey of solutions. Journal of Systems and Software. 2013;86(9):2263-8.

[37] GuiShan D, ZhengJun L, Dong Z. A security domain isolation and data exchange system based on VMM. In2009 3rd International Conference on Signal Processing and Communication Systems 2009 Sep 28 (pp. 1-5). IEEE.

[38] Jouini M., Rabai L. B. A., Aissa A. B. Classification of Security Threats in Information Systems. Procedia Computer Science. 2014;32:489-96.

[39] Zhao F., Huang H., Jin H., Zhang Q. A hybrid ranking approach to estimate vulnerability for dynamic attacks. Computers & Mathematics with Applications. 2011;62(12):4308-21.

[40] Yeng P. K., Yang B., Weyori B. A., Nimbe P., Solvoll T., editors. Web Vulnerability Measures for SMEs. Norwegian Information Security Conference; 2019 20.11.20—9; Narvik: NISK Journal; 2019.

[41] Scarfone K., Mell P. An analysis of CVSS version 2 vulnerability scoring. Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement. 1671289: IEEE Computer Society; 2009. p. 516-25.

[42] Laorden C., Sanz B., Alvarez G., Bringas P. G. A threat model approach to threats and vulnerabilities in on-line social networks. Computational Intelligence in Security for Information Systems 2010: Springer; 2010. p. 135-42.

[43] Yeng P., Yang B., Snekkenes E., editors. Observational Measures for Effective Profiling of Healthcare Staffs' Security Practices. 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC); 2019 15-19 July 2019.

[44] Nomen. (2020). Code of conduct for information security and data protection in the healthcare and care services sector. Retrieved from https://ehelse.no/normen/documents-in-english

[45] ISO. ISO 27799:2016(en), Health informatics Information security management in health using ISO/IEC 27002. 2016.

[46] Hong, Jin B., and Dong Seong Kim. "Performance analysis of scalable attack representation models." In IFIP International Information Security Conference, pp. 330-343. Springer, Berlin, Heidelberg, 2013.

[47] Kallath D. Trust in trusted computing - The end of security as we know it. Computer Fraud & Security. 2005;2005:4-7.

[48] Ryan M. Cloud computing security: The scientific challenge, and a survey of solutions. Journal of Systems and Software. 2013;86:2263-8.

[49] Prosper Kandabongee Yeng, Adam. S., Bian Yang, Einar Arthur Snekkenes. Framework for Healthcare Staffs' Information Security Practice Analysis: Psycho-Socio-Cultural Context. journal of Medical and Internet Research (Preprint). 2019.

[50] Deng M., Wuyts K., Scandariato R., Preneel B., Joosen W. A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. Requirements Engineering. 2011;16(1):3-32.

[51] Manadhata PK, Wing JM. A formal model for a system's attack surface. InMoving Target Defense 2011 (pp. 1-28). Springer, New York, NY.

[52] ISO, ISO 27799:2016(en), Health informatics:Information security management in health using ISO/IEC 27002. 2016.

[53] Yeng P, Nweke LO, Woldaregay AZ, Yang B, Snekkenes EA. Data-Driven and Artificial Intelligence (AI) Approach for Modelling and Analyzing Healthcare Security Practice: A Systematic Review.

[54] Whitten, D., 2008. The chief information security officer: An analysis of the skills required for success. Journal of Computer Information Systems, 48(3), pp.15-19.

[55] ISO, ISO 27799:2016(en), Health informatics Information security management in health using ISO/IEC 27002. 2016.

[56] Mohamed Abomhara, M.G., Geir M. Køien. A STRIDE-Based Threat Model for Telehealth Systems — NISK Journal. in Norsk Informasjonssikkerhetskonferanse. 2015. Ålesund, Norway: NISK Journal.

[57] Cagnazzo, M., Hertlein, M., Holz, T. and Pohlmann, N., 2018, April. Threat modeling for mobile health systems. In 2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW) (pp. 314-319). IEEE.

[58] Almulhem, A., 2012. Threat modeling for electronic health record systems. Journal of medical systems, 36(5), pp.2921-2926.

[59] Deng, M., Petkovic, M., Nalin, M. and Baroni, I., 2011, July. A Home Healthcare System in the Cloud–Addressing Security and Privacy Challenges. In 2011 IEEE 4th International Conference on Cloud Computing (pp. 549-556). IEEE.

[60] Satyanarayanan, M., 2017. The emergence of edge computing. Computer, 50(1), pp.30-39.

[61] Aazam, M. and Huh, E.N., 2014, August. Fog computing and smart gateway based communication for cloud of things. In 2014 International Conference on Future Internet of Things and Cloud (pp. 464-470). IEEE

[62] Opara-Martins, J., Sahandi, R. and Tian, F., 2016. Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective. Journal of Cloud Computing, 5(1), p.4.

[63] German Hospital Hacked, Patient Taken to Another City Dies, https://www.securityweek.com/german-hospital-hacked-patient-taken-another-city-dies, Date: 17/09/2020

[64] HEALTH SHARE OF OREGON: 654,000 PATIENTS. Accessed on 24/09/2020. https://healthitsecurity.com/news/the-10-biggest-healthcare-data-breaches-of-2020-so-far

[65] Luke Irwin,Breach at Norway's largest healthcare authority was a disaster waiting to happen Date: 01/02/2018, https://www.itgovernance.eu/blog/en/breach-at-norways-largest-healthcare-authority-was-a-disaster-waiting-to-happen

[66] Michael Usiagwu is an Entrepreneur, 6 Cloud Security Threats Healthcare Companies May Face – With Solutions, Date: July 14 2020: https://www.tripwire.com/state-of-security/featured/6-cloud-security-threats-healthcare-companies-face-solutions/

[67] European Commission, "What rules apply if my organisation transfers data outside the EU?", https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations/what-rules-apply-if-my-organisation-transfers-data-outside-eu_en

[68] Brad Taylor,"Healthcare organizations and the cloud: Benefits, risks, and security best practices",Accessed on 25/09/2020 https://www.helpnetsecurity.com/2018/01/02/healthcare-cloud-risks/

[69] COMPLIANCE, NETWORK SECURITY, NEWS, "Misconfiguration is the Most Common Cause of Healthcare System Breaches", Accessed on 28.09.2020,https://mytechdecisions.com/compliance/healthcare-system-breaches/

[70] Almulhem, A., 2012. Threat modeling for electronic health record systems. Journal of medical systems, 36(5), pp.2921-2926.

[71] Papa, S.M. and Casper, W.D., 2011. Levels of Trust.

[72] Conrad, E., Misenar, S. and Feldman, J., 2010. CISSP study guide. Syngress.

[73] Yeng, P., Yang, B., Solvoll, T., Nimbe, P. and Weyori, B.A., 2019. Web Vulnerability Measures for SMEs.

[74] Terry, K. (2012). Cloud computing in healthcare: the question is not if, but when. Retrieved from http://www.fiercehealthit.com/story/ cloud-computing-healthcare-question-not-if-when/2020-10-21

[75] Ahuja SP, Mani S, Zambrano J. A survey of the state of cloud computing in healthcare. Network and Communication Technologies. 2012 Dec 1;1(2):12.

[76] Chinnasamy P, Deepalakshmi P. A scalable multilabel-based access control as a service for the cloud (SMBACaaS). Transactions on Emerging Telecommunications Technologies. 2018 Aug;29(8):e3458.

[77] Chinnasamy P, Deepalakshmi P, Shankar K. An analysis of security access control on healthcare records in the cloud. InIntelligent Data Security Solutions for e-Health Applications 2020 Jan 1 (pp. 113-130). Academic Press.

[78] Chinnasamy P, Deepalakshmi P. Design of Secure Storage for Healthcare Cloud using Hybrid Cryptography. In2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) 2018 Apr 20 (pp. 1717-1720). IEEE.

Dense Dilated Inception Network for Medical Image Segmentation

Surayya Ado Bala[1]
Department of Computer Science and Engineering
Sharda, University, Greater Noida, India

Shri Kant[2]
Research and Technology Development Centre
Sharda University, Greater Noida, India

***Abstract*—In recent years, various encoder-decoder-based U-Net architecture has shown remarkable performance in medical image segmentation. However, these encoder-decoder U-Net has a drawback in learning multi-scale features in complex segmentation tasks and weak ability to generalize to other tasks. This paper proposed a generalize encoder-decoder model called dense dilated inception network (DDI-Net) for medical image segmentation by modifying U-Net architecture. We utilize three steps; firstly, we propose a dense path to replace the skip connection in the middle of the encoder and decoder to make the model deeper. Secondly, we replace the U-Net's basic convolution blocks with a modified inception module called multi-scale dilated inception module (MDI) to make the model wider without gradient vanish and with fewer parameters. Thirdly, data augmentation and normalization are applied to the training data to improve the model generalization. We evaluated the proposed model on three subtasks of the medical segmentation decathlon challenge. The experiment results prove that DDI-Net achieves superior performance than the compared methods with a Dice score of 0.82, 0.68, and 0.79 in brain tumor segmentation for edema, non-enhancing, and enhancing tumor. For the hippocampus segmentation, the result achieves 0.92 and 0.90 for anterior and posterior, respectively. For the heart segmentation, the method achieves 0.95 for the left atrial.**

***Keywords*—Deep learning; Dense-Net; inception network; medical image segmentation; U-Net**

I. Introduction

Accurate and automated segmentation of anatomical structures is the most critical and challenging task in analyzing medical images. Medical image segmentation extracts the region of interest for the diagnosis and treatment of various diseases [1], including brain cancer [2], cardiovascular diseases [3], liver cancer [4], pulmonary disease [5], etc., and the list goes on. Accurate and automatic segmentation of anatomical structures is the most important and demanding activity of medical imaging Medical image analysis aims to provide radiologists and clinicians with an efficient, accurate, and precise interpretation of medical images, reducing the time, cost, and error for effective diagnosis. Medical images such as magnetic resonance images (MRI) provides a variety of information (i.e., shape, size, and position) for a diagnostic which achieves multiple anatomical tomographic imaging by setting different parameters [6].

Deep learning (DL) models recently achieved huge success in segmenting medical images [7] because of their great ability to learn critical data features automatically [8][9]. Compared to traditional approaches, multi-layered DL has become the preferred solution for various complicated tasks. Motivated by its performance, multiple types of medical image segmentation research were conducted, notably using a convolutional neural network (CNN) such as brain tumor segmentation [10], heart segmentation [11], and hippocampus segmentation [12].

Over the years, many sophisticated CNN models have been proposed such as Alex Net [13], VGG [14], Google Net [15], Dense Net [16], ResNet [17], Deeplab [18], fully convolution network (FCN) [19] and U-Net [20]. Among these CNN networks, U-Net, an encoder-decoder based model, makes an outstanding achievement and becomes the most famous model in medical image segmentation tasks and computer vision at large that outperformed the existing approaches [21]. The encoder extracts the features while the decoder performs the segmentation based on the extracted features, which results in a remarkable performance on medical images. However, these encoder-decoder architecture has a drawback in learning multi-scale features in complex segmentation tasks and a weak ability to generalize to other tasks. The network structure needs to be optimized to be robust enough to make the parameter space wider and deeper to solve the problem. Although network widening and deepening increase network parameters and computational cost, which causes difficulty while training, causing the gradient to vanish during training [22]. Therefore, the challenge is to make the network wider and deeper without gradient vanishing and fewer parameters.

To overcome the above-aforementioned challenges, we propose a generalized encoder-decoder model called dense dilated inception network (DDI-Net) for medical image segmentation by modifying U-Net architecture. More specifically, we utilize three steps; firstly, we propose a dense path to replace the skip connection between the encoder and decoder to make the model deeper. Secondly, we replace the U- Net's basic convolution blocks with a modified inception module called multi-scale dilated inception module (MDI) to make the model wider without gradient vanishes and with fewer parameters. Thirdly, data augmentation and normalization was applied to the training data to improve the model generalization. We evaluated our DDI-Net on three subtasks of medical segmentation decathlon challenge (MSD) datasets [23]. The experimental results show that our proposed method outperformed the existing ones in each task. Our contribution to this paper is as follows:

- A generalized network named Dense Dilated Inception Network (DDI-Net) is proposed for medical image segmentation.
- We proposed a simple and efficient Preprocessing pipeline, which uses data normalization and augmentation for training and testing DDI-Net to improved segmentation generalization and accuracy.
- We conduct experiments with three different medical segmentation tasks to verify integrated components' performance and the overall model's generalization. The results show that our model outperforms other state-of-the-art models with fewer parameters.

The remaining part of this paper is as follows; we review the related work in Section II. In Section III represents our proposed DDI-Net. The experimental setup, including dataset preprocessing, implementation details, and evaluation, are describes in Section IV. Section V discusses the experiments to evaluate the effectiveness of our DDI-Net. Finally, we conclude in Section VI.

II. Related Work

Nowadays, many encoder-decoder based architectures have been proposed for medical image segmentation. Based on recent studies, the encoder-decoder architecture, such as U-Net, has shown excellent performance due to its flexibility and extensible structure. Several extensions of U-Net have been proposed by integrating sophisticated network blocks such as residual network [24], dense network [25], inception module [26], and dilated convolution [27] for improving segmentation accuracy. Li et al. [25] proposed a hybrid densely U-Net (H-DenseU-Net) for 3D liver and tumor segmentation. H-DenseUNet combines densely connected paths and U-Net to improve performance. Alternatively, Yang et al. [28] propose a U-Net with dilated convolution, and they called their structure DCU-Net for brain tumor segmentation.

Similarly, Chen et al. [29] embedded dense and residual blocks into a U-Net segmentation network. Ibtehaz and Rahman [30] combine a U-Net with residual inception modules for multi- scale feature extraction and perform segmentation on different modalities. Also, Wang et al. [31] integrate the inception module in U-Net architecture for segmentation of left atrial. Li and Tso [32] in cooperated inception modules and dilated inception modules in U-Net architecture for liver and tumor segmentation. Furthermore, Zang Z.et al [33] integrates the inception module with a dense connection into U-Net architecture. Jingcong L. et al. [34] replace the basic convolution block of U-Net architecture with a dilated inception block for multi-scale feature aggregation for cardiac right ventricle segmentation. Moreover, Bala S.B. and Kant S. [35] proposed a hybrid network. They combined CNN and Gated Recurrent Unit (GRU) using the U-Net structure to perform segmentation of cardiac MRI.

III. Proposed Method

In this study, inspired by U-Net, Dense-Net, Inception module, and Dilation convolution, we proposed a generalized medical segmentation model. The model was built upon U-Net based encoder-decoder architecture by integrating dense path and MDI blocks into U-Net. We modify U-Net by replacing the skip connection with the proposed dense path between the encoder and decoder and in cooperating MDI block to replace the basic convolutional block to improve the model's accuracy. Fig. 1 illustrate the proposed DDI-Net architecture. The DDI-Net comprises four dense paths, nine MDI blocks, four down sampling layers, four up sampling layers, and one output layer.

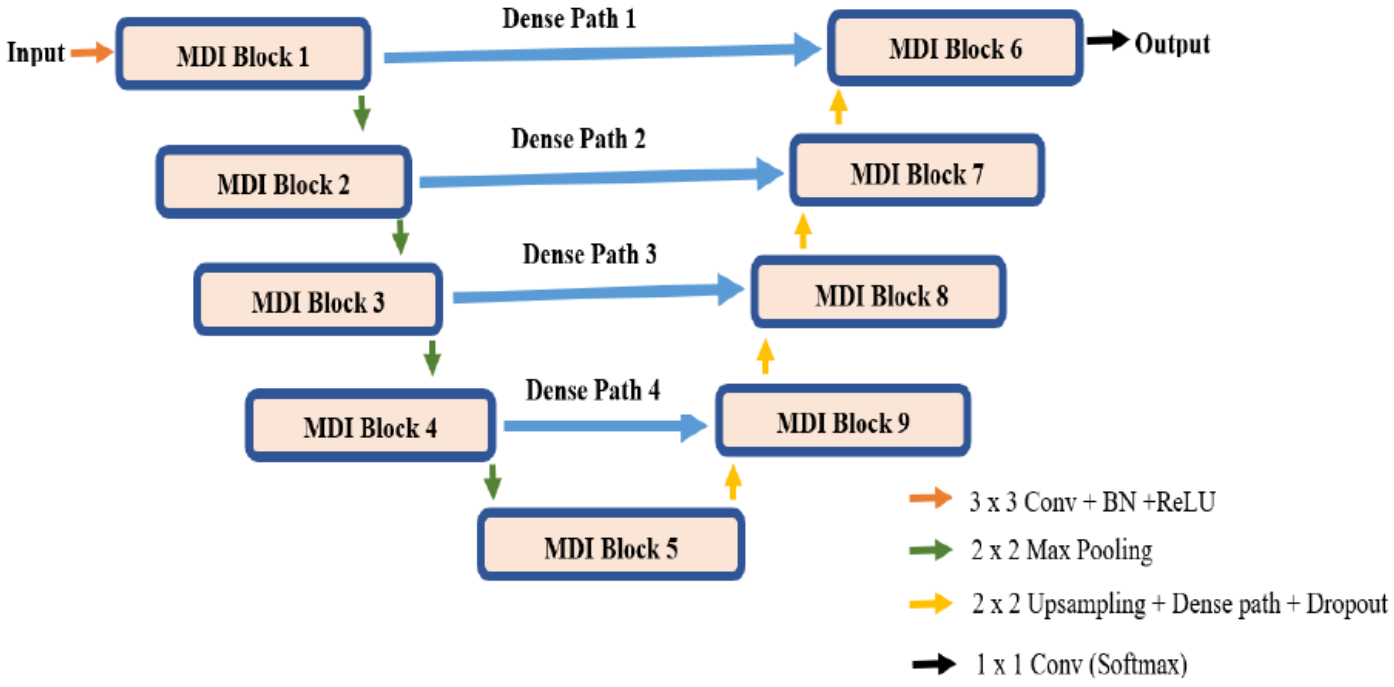


Fig. 1. Overall Architecture of the DDI-Net Model.

A. Dense Path

In U-Net, features extracted at the encoding path are pass using skip connections to their corresponding decoder path, which results in the passing of extra features forward, leading to the mortifying the exactitude of segmentation [36][37]. Also, we observe a large semantic gap between the encoder and the decoder feature map. Thus concatenation (feature fusion) of the feature maps from the encoder and the decoder will cause disparity during learning, thereby affecting segmentation prediction. Therefore to alleviate these challenges, we proposed to replace the skip connection with convolutional layers densely connected, which we referred to as dense path. Rather than merely concatenating the feature. As illustrated in Fig. 2, the dense path comprises densely connected convolution layers with 3x3 filters and a bottleneck layer. The dense path allows in-depth supervision to make the model deeper to allow the encoder to extract low-level features, thus helping the decoder recover the lost spatial information. The dense path also improved the flow of information and the gradient all over the network. This aids in alleviating the difficulty in training the network and hence reduces overfitting with its regularizing effect. Moreover, the dense path performs feature reuse to utilize the network's potential, with a resilient condensed model that is easy to train and highly parameter efficient.

B. Multi-Scale Dilated Inception Block

There are usually different scales of interest in medical image segmentation, such as tumors, lesions, and organs. Therefore, we need a network that can learn and extract multi-scale features with fewer parameter Networks models like googleNet [12] propose the inception module. The inception module consists of multiple convolutional layers with kernels of different sizes that learn multi-scale features. In each convolutional layer, the receptive field size is determined by the kernel size [38]. These kernel sizes include both small and large sizes. The small kernels are used to learn small scale features such as 1 x 1, 3 x 3, while the large scale kernel is used to learn large scale features such as 7 x 7 and 13 x 13[38]. According to [39] [40], multi-scale features improved the performance of the network model. However, large convolutional kernels used in obtaining large scale features increase the parameters and computational cost. To overcome this challenge, [39] apply dilated convolution. Dilated convolution is a convolution type that expands the receptive field to obtain large scale features using different dilation rates without increasing the parameters and computational cost. Inspired by the inception module [12] and dilated convolution [39], we propose a modified inception module by incorporating dilated convolutions called multi-scale dilated inception module (MDI). MDI module is developed to be used in the encoder as well as the decoder path to extract and aggregate the multi-scale feature maps. These feature maps are aggregated from kernels of different sizes with different dilation rates to widen the network to learn multi-scale features to improve the segmentation performance [41]. As depicted in Fig. 3, three convolutional layers with 3x3 kernels with four different dilation rates are used in the MDI module. The dilation rates are 1, 2, 4, and 6. Each convolutional kernel's feature scale is $(2l+l)^2$, where l is the kernel's dilation rate. Features extracted from the dilated convolution result produce a different scale of 3 x 3, 5 x 5, 9 x 9, and 13 x 13, as illustrated in Fig. 4. The output of the four dilated convolution layers is concatenated. Batch normalization [42] is applied to accelerate the training and enhance the model's stability, followed by a 1x1 convolution to reduce the dimension and ReLU is used as the activation function for each convolutional layer [43].

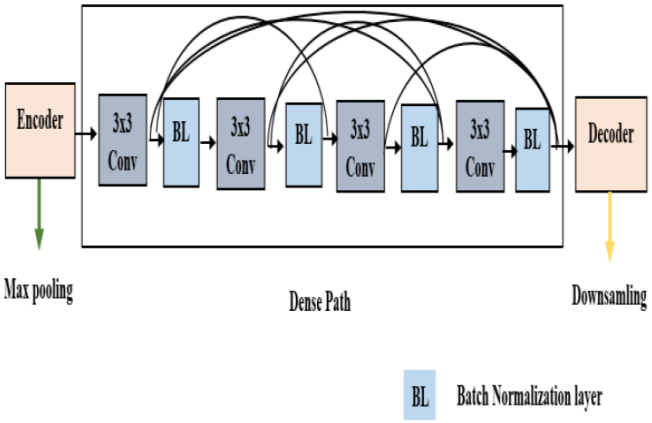


Fig. 2. The Architecture of Dense Path.

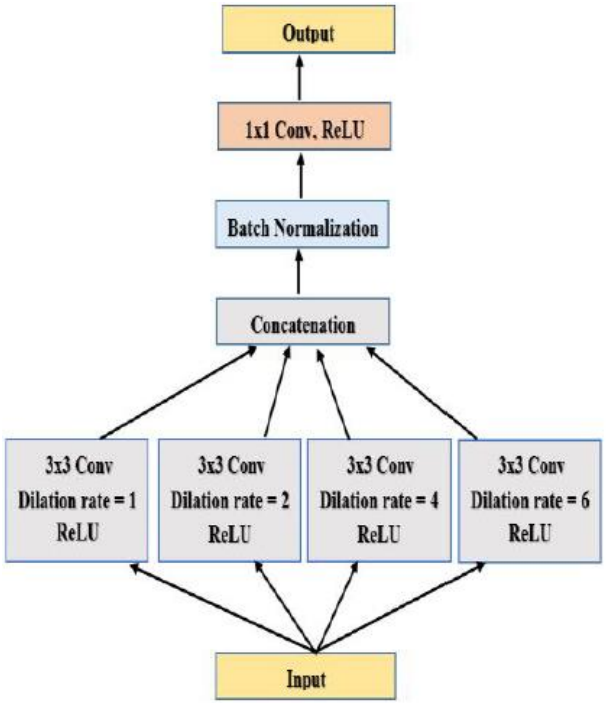


Fig. 3. The Architecture of Multi-Scale Dilated Inception (MDI) Block.

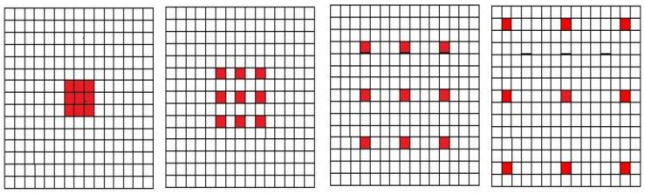


Fig. 4. Dilated Convolution with 3x3 Kernel with a Dilation Rate of 1, 2, 4, and 6.

We modified the U-Net architecture by replacing the convolution block with the proposed MDI module. Experiments verified that our proposed MDI enhanced the segmentation performance by learning more multi-scale features without any free blow up in computational complexity [37] with fewer parameters than the original inception module.

IV. Experimental Setup

The experimental setup, including the preprocessing of the dataset, implementation description, and evaluation metrics, is discussed in this section.

A. Datasets

We use three subtasks from the decathlon challenge dataset for medical segmentation. There are 484, 260, and 20 image data for brain tumors, hippocampus, and heart tasks. In Table I, the dataset is briefly outlined.

1) Preprocessing of training and testing data: Various scanners, institutions, and anatomical structures with different pixel spacing were used in collecting training and testing data used. Hence, these differences make it very important to preprocess the training and testing data before feeding our model. Fig. 5 shows the overview of the preprocessing steps followed during training and testing. Precisely, we performed image resampling to make the pixel spacing of all the images the same, and then we normalized the images. Lastly, data augmentation is applied during the training and testing process to improved generalization.

a) Image Resampling: Since the dataset used for both training and testing, the experiments are from three different datasets with pixel spacing ranging from 1mm to 1.25mm.We do image resampling to eliminate the difference. For brain MRI the pixel spacing is 1mm x 1mm x 1mm, while the hippocampus 1mm x 1mm x 1mm and heart is 1.25mm x 1.25mm x 2.70mm. Therefore, we resample the heart images to 1mm to make the spatial resolution the same as the brain and hippocampus images. After image resampling, we applied intensity normalization to the three datasets' images to normalize the image.

b) Data Normalization: We normalize the images using intensity normalization by subtracting the volume's mean and dividing by the volume standard deviation to the range of 0, 1. After normalization, we applied augmentation to increase the training data to improve model generalization and avoid overfitting.

c) Data Augmentation: Data augmentation increases the training data by artificially generating more training data to generalize the model. The training data is augmented by;

- Random rotation of angle between -5 and 5 degrees.
- Vertical flipping with a probability of 0.2 for increasing the orientation variety.
- Random image scaling with a scale factor s: s E [0.2, 0.6] to maximize the images' variance.

B. Implementation Details

The network model has been implemented using keras [44] with tensorflow [45] backend using python 3 programming languages. Our network was trained and tested on a desktop computer with NVIDIA GeForce RTX 2080Ti with 11 GB of memory and 2 graphics card. During the training, the network was initialized with the normal weight [46], 0 bias, 0.0001 learning rate, and cross-entropy as our loss function. We optimize our network with Adam optimizer [47] with Beta-1=0.90, Beta-2 = 0.99 and epsilon = 0.000001. We executed 5-fold cross-validation and trained the model for 100 epochs. After every epoch, we evaluate the model using the validation data, and then the best model is selected for evaluating the test data. For the training and validation, we use a batch size of 4. In each epoch, 4 data is transposed to the model as input. All layers use a Rectified Linear Unit (ReLU) as an activation function except the output layer that uses softmax. We use batch normalization to normalize the feature maps and stabilize the network while training.

C. Evaluation Metric

The performance of our model is to assess using the Dice score. It is evaluated as;

$$Dice\ (GT, SR) = \frac{2|GT \wedge SR|}{(\ |GT| + |SR|\ )} \quad (1)$$

GT and SR are ground truth and segmentation results, respectively. Ground truth is the segmented region extracted by experienced experts manually using standard annotation protocol. In contrast, the segmentation result is the segmented region from the evaluated method.

TABLE I. General Description of the Dataset

Name	Modality	Number of subjects	Scanners	Source	Image Spatial Resolution	Target
Brats 2016 & 2017	Multimodal Multisite MRI Data FLAIR, T1w, T1gd, T2w	484	1T to 3T	Brats 2016 & 2017 Datasets	Gliomas segmentation necrotic, active tumor and oedema	1x1x1mm^3
LASC	Mono-modal MRI	20	1.5T Achieva Scanner Philips	Kings College London	Left Atrium	1.25x1.25x1.25mm^3
Hippocampus	Mono-modal MRI	260	1.5T Achieva Scanner Philips	Vanderbilt University Medical Centre	Hippocampus Head and Body	1x1x1mm^3

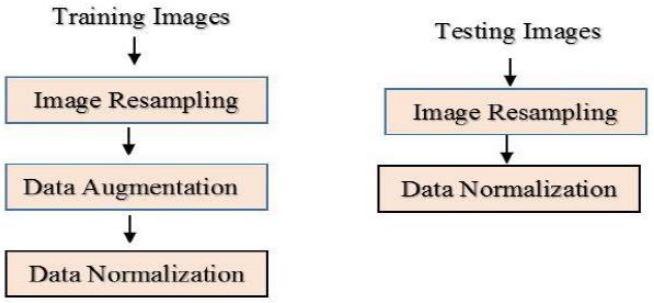


Fig. 5. Overview of Preprocessing of Images Steps during Training and Testing.

V. Experimental Result

This section evaluates our proposed model's effectiveness and generalizability on three separate segmentation tasks, including multimodal MRI segmentation of brain tumors, mono-modal MRI segmentation of the hippocampus, and MRI segmentation of the heart.

A. Brain Tumour Segmentation

We experiment with brain MRI images for brain tumor diagnosis of glioma to test our model's efficacy. The most common brain tumor found in the brain and spinal cord is a glioma. Due to the diverse and heterogeneously positioned targets shown in Fig. 6, glioma segmentation is a difficult task. This segmentation is targeted at segmenting glioma into edema, tumor non-enhancement, and tumor enhancement. 484 multi-parametric magnetic resonance imaging (MRI) scans from patients diagnosed with glioblastoma or lower grade glioma with the same number of ground-truth images are included in the brain dataset given. The proposed method uses all four sequences to segment brain MRI images, comprising volumes of Native T1-weighted (T1), Post-contrast T1-weighted (T1-Gd), T2-weighted (T2), and T2-fluid attenuation inversion recovery (FLAIR). 70 % of the data in this experiment is used for training, 15 % for validation, and 15% for testing. To get an accurate and stable model, we performed a 5-cross validation. DDI-Net results were contrasted with two recently published state-of-the-art models, and the outcome is shown in Table II. The results of the dice score obtained from DDI-Net demonstrated superior performance over the existing models.

B. Hippocampus Segmentation

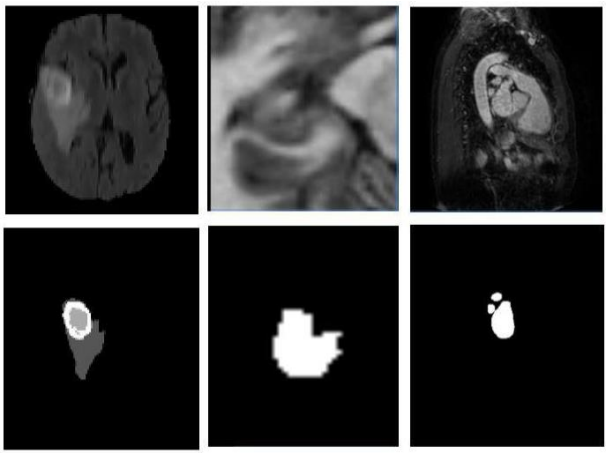
The hippocampus is a complex organ of the brain embedded deep in the temporal lobe. In learning and memory, it has the most responsible function. For Alzheimer's disease (AD) diagnosis, hippocampus segmentation is essential. As shown in Fig. 6, a complicated task is hippocampus segmentation. It has two adjacent tiny structures with high precision. The data set consisted of 260 stable adults and adults with non-affective psychotic illness, taken from the Vanderbilt University Medical Center phenotype data repository. 70% of the data in this experiment is used for training, 15% for validation, and 15% for testing.

To get an accurate and stable model, we performed a 5-cross validation. The hippocampus's entire MRI is used as the input to the network, as shown in Table III. Compared to the other two art method states, our proposed method gets the highest result.

C. Heart Segmentation

The heart is one of the human body's vital organs that pump blood throughout the body. Segmentation of the Left atrial from the heart plays a vital role in diagnosing atrial fibrillation (AF).

Segmentation of the left atrial from the heart is challenging because of the small training dataset with considerable variability, as shown in Fig. 6. The provided dataset consists of 20 MRI images from the left atrial segmentation challenge (LASC), Kings College Kingdom, London, United Kingdom. We use the whole MRI of the heart as input to the network. As shown in Table IV, the best result compared to other method states is obtained by our proposed method.



(a) (b) (c)

Fig. 6. A Sample of the Three Tasks of Segmentation. The First Row Demonstrates the Original Images and the Ground Truths was Shown in the Second Row. (a) Brain Tumor. (b) Hippocampus. (c) Heart.

TABLE II. Comparisons Result on Brain Tumor Segmentation

Networks	Edema	Non-enhancing	Enhancing
NDN	0.71	0.60	0.72
nnU-Net	0.68	0.48	0.68
DDI-Net (ours)	**0.82**	**0.68**	**0.79**

TABLE III. Comparisons Result on Hippocampus Segmentation

Networks	Anterior	Posterior
NDN	0.88	0.89
nnU-Net	0.90	0.89
DDI-Net (ours)	**0.92**	**0.90**

TABLE IV. Comparisons Result on Heart Segmentation

Networks	Left Atrial
NDN	0.85
nnU-Net	0.93
DDI-Net (ours)	**0.95**

D. Ablation Studies

We propose and introduce dense paths and the MDI blocks to improve the baseline encoder-decoder-based U-Net model's segmentation accuracy in the proposed method.

To verify these introduced modules' effectiveness, we conduct the following ablation studies to investigate their contributions to the overall DDI-Net performance. We use the heart dataset for the ablation studies because it is the most challenging dataset used in our experiment. Hence, we make a comparison among the U-Net, U-net with dense paths (U-Net + dense path), and U-net with MDI blocks (U-Net +MDI) and the DDI-Net (U-Net+ Dense path +MDI). We initially start with the baseline U-Net and then assess the dense path and MDI block's effect on the results.

1) Ablation study for replacing the skip connection with the dense path: To verify the dense path's effectiveness, we replaced the skip connection with the proposed dense path. Table V illustrates the segmentation result. The results show that we achieved 0.9 on the dice score compare to 0.89 in the original U-Net. Our result signifies that the dense path proposed has improved the segmentation accuracy, making the network deeper and without a vanishing gradient. The dense path also alleviates the semantic gap between the encoder and the decoder by adding more blocks of convolutional operation and dense connection, which aids in a proper fusion of the feature maps.

2) Ablation study for replacing the convolutional layer with MDI blocks: To verify MDI blocks' effectiveness, we replaced the basic convolutional blocks with MDI blocks. Table VI depicted the segmentation results. The results illustrate that we achieved 0.93 on the dice score compare to 0.89 in the original U-Net. We observed that MDI blocks make the network wider; this aid in extracting multi-scale features from different scales. This indicates that using a filter of different sizes allowed the network to capture multi-scale features and improved the segmentation result.

3) Ablation study for the proposed DDI-Net: To verify the effectiveness of DDI-Net, We experimented with dense path and MDI blocks together. The results of the comparison are depicted in Table VII. Our results show that we achieved 0.95 on the dice score compare to 0.89 in the original U-Net. Table VII shows that the DDI-Net contributes to improving medical image segmentation's performance and accuracy. The cooperation between these two proposed modules into U-Net has yielded the best result from the segmentation results.

E. Evaluating the Effect of Data Normalization and Data Augmentation on DDI-Net Generalization

Using two data normalization and three data augmentation techniques, including image resampling, intensity normalization, rotation, flipping, and scaling, this section verifies the efficacy of data normalization and data augmentation on DDI-Net generalizability. We trained DDI-Net using all three datasets with the same setting to analyze the impact of data normalization and augmentation in model generalization. Firstly, we experiment with data normalization only. Secondly, we experimented with data augmentation and experimented with normalization and augmentation of data, as seen in Table VIII. From Table VIII, it indicates that the data normalization and augmentation increase Dice score result. By integrating data normalization and augmentation operations, the best segmentation efficiency is obtained for all three datasets.

F. Algorithm Run-Time

Table IX shows the training and testing time for all the models in each experiment. It can be found that in both segmentation tasks, the proposed model requires less time for training and testing compared to nnU-Net and NDN. Besides, brain data requires more time than the hippocampus and heart dataset for training and research.

G. Comparison with State-of-the-Art Methods

To verify the effectiveness of our proposed improvements with the state of the art methods. We compare our method with two proposed methods by Wang L. et al. [48] and Isensee F.et al. [49].For the brain and hippocampus dataset, the result is from the papers. For the heart dataset, Wang L.et al. do not perform implementation with the heart dataset. We obtained the result using Wang L.et al; implementation details and Isensee F. result from their paper. Tables II, III, and IV show the two methods' dice score and the DDI-Net on the three datasets. As observed visually from the tables, the proposed DDI-Net improves the segmentation's accuracy and generalizes all three datasets. Fig. 7 visually illustrates the output results of the DDI-Net proposed.

TABLE V. ABLATION STUDY FOR REPLACING THE SKIP CONNECTION WITH DENSE PATH

Networks	Heart Segmentation Left Atrial	Parameters
U-Net	0.89	2.81M
U-Net + Dense path	**0.90**	**1.75M**

TABLE VI. ABLATION STUDY FOR REPLACING THE CONVOLUTION BLOCK WITH MDI BLOCKS

Networks	Heart Segmentation Left Atrial	Parameters
U-Net	0.89	2.81M
U-Net + MDI	**0.93**	**1.10M**

TABLE VII. ABLATION STUDY FOR REPLACING THE SKIP CONNECTION AND CONVOLUTION BLOCK WITH DENSE PATH AND MDI BLOCKS

Networks	Heart Segmentation Left Atrial	Parameters
U-Net	0.89	2.81M
DDI-Net	**0.95**	**0.90M**

TABLE VIII. PERFORMANCE OF DDI-NET WITH DIFFERENT TRAINING CONFIGURATION

Configuration		Brain Segmentation			Hippocampus Segmentation		Heart Segmentation
Data Norm	Data Aug	Edema	Non-enhancing	Enhancing	Anterior	Posterior	Left Atrial
X	√	0.73	0.50	0.71	0.85	0.81	0.82
√	X	0.79	0.62	0.75	0.89	0.86	0.89
√	√	**0.82**	**0.68**	**0.79**	**0.92**	**0.90**	**0.93**

TABLE IX. MODEL TRAINING AND TESTING TIME

Networks	Brain Segmentation		Hippocampus Segmentation		Heart Segmentation	
	Training Time	Testing Time	Training Time	Testing Time	Training Time	Testing Time
NDN	8.2h	1.5sec	7.4h	0.91sec	4.7h	0.8sec
nnU-Net	7.6h	1.8sec	8.1h	1.32sec	5.2h	1.3sec
DDI-Net (ours)	**7.5h**	**0.9sec**	**6.5h**	**0.68sec**	**4.3h**	**0.7sec**

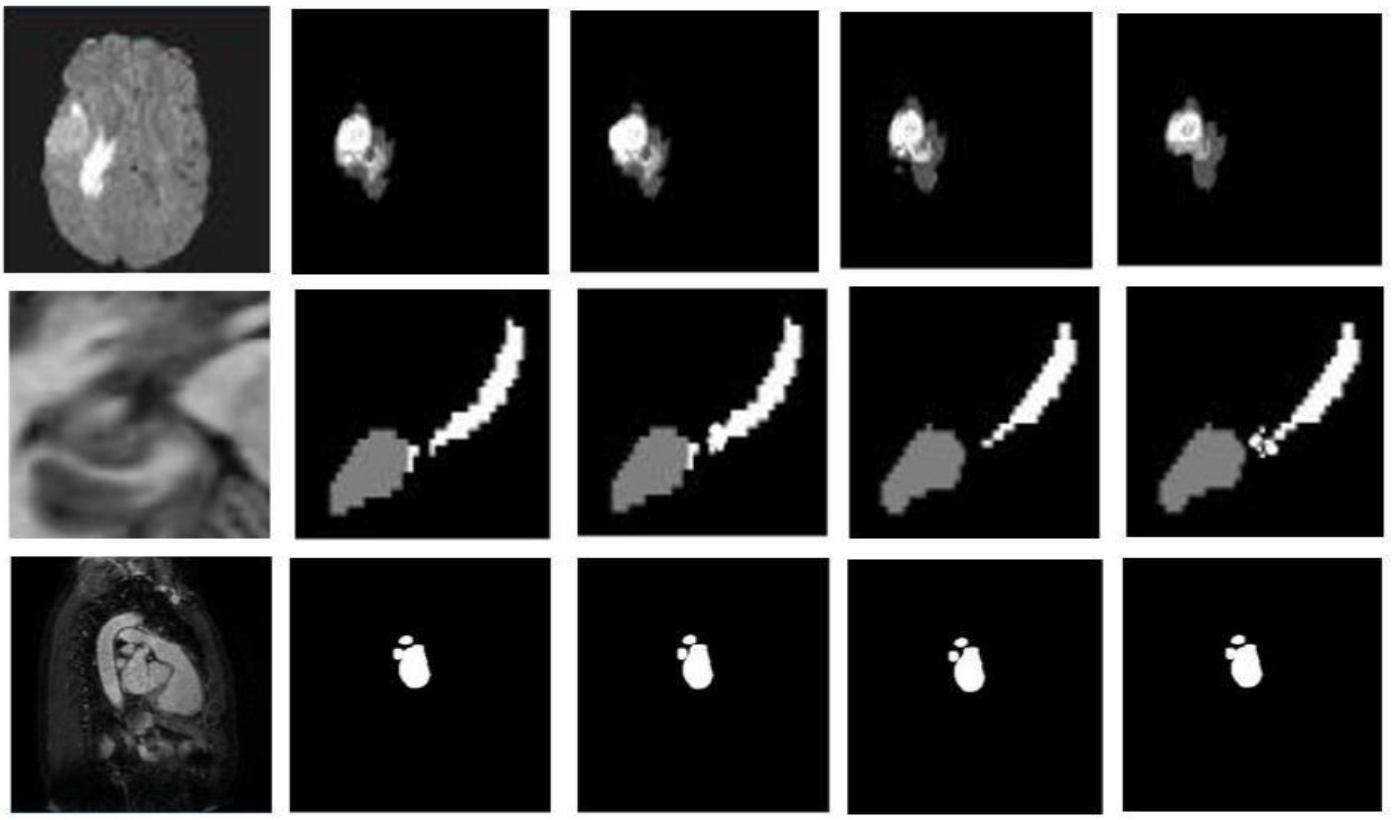


Fig. 7. Visual Illustration of Segmentation Output on Brain, Hippocampus and Heart Dataset. from Left to Right: Original Image, Ground truth, DDI-Net, NDN and nnU-Net.

VI. CONCLUSION

In this paper, by modifying the U-Net architecture using Dense-Net, Dilated Convolution, and Inception network, we propose a new encoder-decoder network called DDI-Net. There are two features on the DDI-Net, namely dense paths and MDI blocks. The dense path enables in-depth supervision to deepen the model. Low-level features can be extracted by the encoder, thus helping the decoder recover the missing spatial information. This helps to facilitate the reuse of features with a resilient simplified training path and highly efficient parameters.

The MDI block, meanwhile, makes the model wider without the gradient vanishing but with fewer parameters. Besides, using data normalization and augmentation, we propose a general training and testing process. The experiments conducted show that they play an essential role in generalizing the model across images from various tasks. To prove the DDI-Net generalization, the model is tested using

three distinct medical image segmentation tasks. The result shows that our DDI-Net performs better than the state of the art method, including nnU-Net and NDN. However, we have limited the tasks only to MRI images. Therefore, to make the network more general and efficient for all medical problems in the future, we plan to experiment with other modalities.

REFERENCES

[1] S.A.Bala, S.Kant. " Deep learning based model architectures for cardiac MRI segmentation: A Survey," International Journal of Innovative Science, Engineering & Technology, pp.129-135, 2020.

[2] N.M.Aboelenein, S.Piao, K. Anis, N.Alam, and A.Ahmed. "HTTU-Net: Hybrid Two Track U-Net for automatic brain tumor segmentation," IEEE Access, 2020.

[3] Lindsey, Tony, and J.J.Lee." Automated Cardiovascular Pathology Assessment Using Semantic Segmentation and Ensemble Learning," Journal of Digital Imaging, pp. 1-6, 2020.

[4] S.Almotairi, G. Kareem, M. Aouf, B. Almutairi, and M.A.M. Salem. "Liver tumor segmentation in CT scans using modified segnet," Sensors,5 (20), pp. 1516, 2020.

[5] C.Liu, R. Zhao, W.Xie, and M.Pang." Pathological lung segmentation based on random forest combined with deep model and multi-scale superpixels." Neural Processing Letters, pp.1-19, 2020.

[6] Lundervold, S.Alexander, and L.Arvid. "An overview of deep learning in medical imaging focusing on MRI," Zeitschrift für Medizinische Physik, 2(29), pp. 102-127, 2019.

[7] G.Litjens, et al. "A survey on deep learning in medical image analysis," Medical Image Analysis, 42, pp. 60-88, 2017.

[8] L.Yann, B.Yoshua, and H.Geoffrey. "Deep learning," Nature, 521(7553), pp. 436-444, 2015.

[9] S.A. Bala, S.Kant and K. Kumar." Impact of deep learning in medical imaging: a systematic new proposed model," International journal of recent technology and engineering, pp.112-118,2019.

[10] D.Daimary, M.B.Bora, K.Amitab, and D.Kandar. "Brain Tumor Segmentation from MRI Images using Hybrid Convolutional Neural Networks," Procedia Computer Science, 167, pp. 2419-2428,2020.

[11] S.M.Ibrahim, M.S.Ibrahim, Shakeel Muhammad, M.Usman, I.Naseem, and M.Moinuddin. "Heart Segmentation From MRI Scans Using Convolutional Neural Network," arXiv preprint arXiv:1911.09332, 2019.

[12] M.Liu, L.Fan, Y.Hao, W.Kundong, M.Yixin, and X.Mingqing. "A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease." NeuroImage 208, 2020.

[13] A. Krizhevsky, I.Sutskever, and G.E Hinton. "ImageNet classification with deep convolutional neural networks," Advances in neural information processing systems, pp.1097-1105,2012.

[14] K.Simonyan, A. Zisserman." Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556,2014.

[15] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.

[16] G.Huang, L.Zhung, M. Laurens, and K.Q Weinberger Gao. "Densely connected convolutional networks," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708 2017.

[17] K.He, Z.Xiangyu, R.Shaoqing, and S.Jian. "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition, pp.770-778, 2016.

[18] L.C.Chen, G.Papandreou, I.Kokkinos, K.Murphy, and A.L Yullie. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," IEEE transactions on pattern analysis and machine intelligence, 40 (4), pp.834-848, 2017.

[19] J.Long, E.Shelhamer, and T. Darrell. "Fully convolutional networks for semantic segmentation," Proceedings of the IEEE conference on computer vision and pattern recognition, pp.3431-3440, 2015.

[20] O.Ronneberger, P. Fischer, and T.Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, pp. 234-241 2015.

[21] L.Liu, J.Cheng, Q.Quan, F.Wu, Y.P.Wang, and J.A Wang." A survey on U-shaped networks in medical image segmentation," Neurocomputing, 409, pp.244-258, 2020.

[22] L.Bi, J.Kim, A.Kumar, M.Fulham, D.Feng." Stacked fully convolutional networks with multi-channel learning: application to medical image segmentation." The Visual Computer 33, pp.1061-1071 2017.

[23] Simpson, Amber L., et al. "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," arXiv preprint arXiv:1902.09063, 2019.

[24] M.Z.Alom, M.Hassan, C.Yakopcic, T.M.Taha,V.K.Asari. "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," arXiv preprint arXiv:1802.06955 , 2018.

[25] X.Li, H.Chen, X.Qi, Q.Dou, C.W.Fu, P.A Heng. "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes," IEEE transactions on medical imaging 37(12), pp.2663-2674, 2018.

[26] D.E.Cahall, G.Rasool, N.C.Bouaynaya, H.M.Fathallah-Shaykh. "Inception modules enhance brain tumor segmentation," Frontiers in computational neuroscience,13, pp.44, 2019.

[27] A.Tureckova and A. J. Rodríguez-Sánchez. "ISLES challenge: U-shaped convolution neural network with dilated convolution for 3D stroke lesion segmentation." International MICCAI Brainlesion Workshop. Springer, Cham, pp.319-327, 2018.

[28] T.Yang, Y.Zhou, L.Li, and C.Zhu." DCU-Net: Multi-scale U-Net for brain tumor segmentation." Journal of X-Ray Science and Technology Preprint, pp. 1-18, 2020.

[29] L.Chen, P.Bentley, K.Mori, K.Misawa, M.Fujiwara, and D.Rueckert "DRINet for medical image segmentation," IEEE transactions on Medical Imaging, 37 (11),pp.2453-2462, 2018.

[30] N.Ibtehaz and M.S. Rahman. "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," Neural Networks 121,pp.74-87,2020.

[31] C.Wang, M.Rajchi, A.D Chan, and E. Ukwatta. "An ensemble of U-Net architecture variants for left atrial segmentation," Medical Imaging 2019: Computer-Aided Diagnosis. Vol. 10950. International Society for Optics and Photonics, 2019.

[32] S.Li and G.K. Tso. "Bottleneck supervised u-net for pixel-wise liver and tumor segmentation," arXiv:1810.10331, 2018.

[33] Z.Zhang, C.Wu, S.Coleman, D.Kerr." DENSE-INception U-net for medical image segmentation," Computer Methods and Programs in Biomedicine (192),2020.

[34] J.Li, Z.L.Yu, Z.Gu, H.Liu, and Y.Li. "Dilated-inception net: multi-scale feature aggregation for cardiac right ventricle segmentation." IEEE Transactions on Biomedical Engineering 66(12), pp. 3499-3508, 2019.

[35] S.A Bala and S. Kant S. "A hybrid spatio-temporal network for cardiac MRI segmentation," International journal of advanced trends in computer science and engineering, pp.9089-9097,2020.

[36] Q.Tong, et al. "RIANet: Recurrent interleaved attention network for cardiac MRI segmentation," Computers in biology and medicine, 109,pp.290-302, 2019.

[37] Z.Zhang, X.Zhang, C.Peng, X.Xue and J.Sun." Exfuse: Enhancing feature fusion for semantic segmentation." Proceedings of the European Conference on Computer Vision (ECCV),pp.269-284,2018.

[38] W.Luo, Y.Li, R.Urtasun, and R.Zemel. "Understanding the effective receptive field in deep convolutional neural networks," In Advances in Neural Information Processing Sytems, pp. 4898–4906, 2016.

[39] A. Krizhevsky, I. Sutskeyer, and G.E Hinton. "Imagenet classification with deep convolutional neural networks," In Advances in Neural Processing Systems,pp.1097-1105,2012.

[40] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv preprint arXiv:1511.07122, 2015.

[41] C. Peng, X.Zhang, G.Yu, G.Luo, and J. Sun. "Large kernel matters-improve semantic segmentation by the global convolutional network," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.4353-4361,2017.

[42] S.Ioffe, and C. Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint arXiv:1502.03167,2015.

[43] A.F.Agarap. "Deep learning using rectified linear units (ReLU)." arXiv preprint arXiv:1803.08375, 2018.

[44] F.Chollet. "Keras: The python deep learning library," [online] Available: https://keras.io/.

[45] M.Abadi et al., "Tensorflow: Large-scale machine learning on heterogeneous systems," arXiv:1603.04467, 2015.

[46] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," Int. Conf. Learn. Represent, pp. 1-13, 2015.

[47] K. He, X. Zhang, S. Ren, J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," Proceeding of IEEE International Conference on Computer Vision, pp. 1026-1034, 2015.

[48] L.Wang, R. Chen, S.Wang, N.Zeng, X.Huang, C.Liu. "Nested Dilation Network (NDN) for Multi-Task Medical Image Segmentation." IEEE Access. 1(7) ,pp.44676-44685,2019.

[49] F. Isensee, et al. "nnU-net: Self-adapting framework for u-net-based medical image segmentation." arXiv preprint arXiv:1809.10486 ,2018.

Phishing Image Spam Classification Research Trends: Survey and Open Issues

Ovye John Abari[1], Nor Fazlida Mohd Sani[2], Fatimah Khalid[3],
Mohd Yunus Bin Sharum[4], Noor Afiza Mohd Ariffin[5]
Faculty of Computer Science & Information Technology,
Universiti Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia

***Abstract*—A phishing email is an attack that focused completely on people to circumvent existing traditional security algorithms. The email appears to be a dependable, appropriate, and solid communication medium for internet users. At present, the email is submerged with spam content, both in text-based form or undesired text planted inside the images. This study reviews articles on phishing image spam classification published from 2006 to 2020 based on spam classification application domains, datasets, features sets, spam classification methods, and the measurement metrics adopted in the existing studies. More than 50 articles, both from Web of Science and Scopus databases were picked. Achieving the study's target, we carried out a broad survey and analysis to identify the domains where spam classification was applied. Furthermore, several public data sets, features set, classification methods, and measuring metrics are found and the popular once were pinpointed. The study revealed that Personal Collection, Dredze, and Spam Archives datasets are the most commonly used datasets in image spam classification research. Low-level and image metadata are the most widely used features set. The methods of image spam classification as identified in this study are supervised machine learning, unsupervised machine learning, semi-supervised machine learning, content-based and statistical learning. Among these methods, the most commonly utilized is the Support Vector Machine (SVM) which falls under supervised machine learning. This is followed by Naïve Bayes and K-Nearest Neighbor. The commonly adopted metrics for the performance evaluation of the existing image spam classifiers are also identified and briefly discussed. We compared the performance of the state-of-the-art image spam models. Lastly, we pointed out promising directions for future research.**

***Keywords*—Phishing; spam; image spam classification; machine learning; deep learning**

I. Introduction

Phishing is a social engineering attack against people in a helpless society by controlling human beings into giving their confidential information to the cheats, called phishers. It is a criminal way of stealing internet users' private information using deceptive emails and counterfeit websites [1]. Phishing is also defined by [2] as a criminal instrument that utilizes both social engineering and specialized deception to take consumers' individual personality information and monetary account credentials. The coming of the Internet and the increasing number of its users have made email to be an important medium of communication. As of late, there has been an expanding utilization of emails and this has driven to the appearance of issues caused by phishing emails and spam. A typical email user gets around 40-50 emails per day [3].

According to [4], the entire number of phish identified in 1Q 2018 was 263,538. This was more than 45% from the 180,577 taken note in 4Q 2017. It was moreover higher than the 190,942 recorded in 3Q 2017. Likewise, the whole number of phishing identified in 2Q 2018 was 233,040, related to 263,538 in 1Q 2018. These sums are more than the 180,577 recorded in 4Q 2017 and the 190,942 watched in 3Q 2017. The phishing identified in 2Q and 3Q of 2019 were 112,163 and 122,359 respectively. Although there is a significant decrease in the phishing activities when compared with the figures of the previous years (2018 and 2017); however the request for phishing identification in our contemporary society is still a necessity to protect end-users from malicious emails. Phishing attacks are growing speedily in size and it's attacks expanding dynamically. This results in a serious economic loss around the world [1]. Fig. 1 depicts the statistics of phishing attacks in the 1Q of 2019 while Fig. 2 illustrates the most-targeted industry sectors in 2Q of 2019 [4].

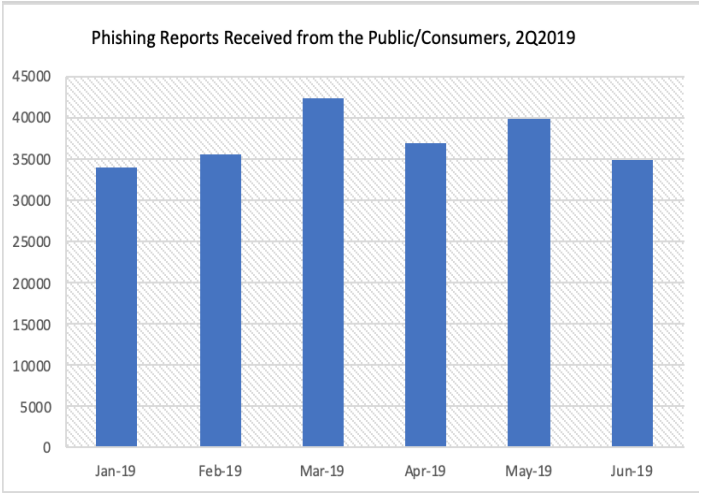


Fig. 1. Phishing Report of 2Q 2019 [4].

The past decade has seen the internet and emails to be flooded with spam content [5]. Regardless of constant awareness and the number of anti-spam algorithms emerging, spam contents are in increase [6]. Sending a large volume of spam contents at the server-side causes delays in service response, reducing the authenticity of the mail and consume a large portion of the storage space. At the user side, grouping the spam into valid and not valid, considering the large number

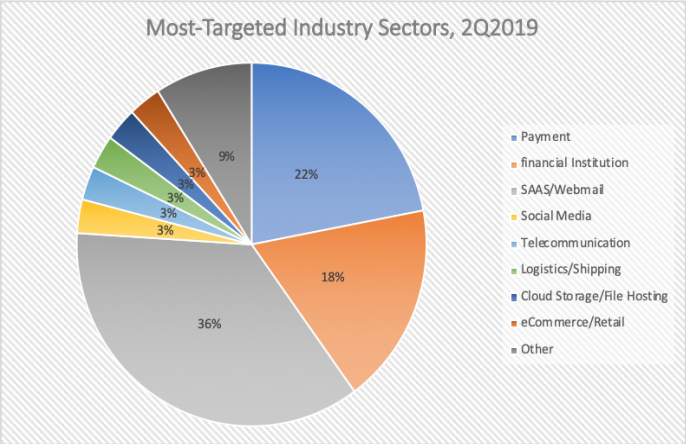


Fig. 2. Most Targeted Industry Sectors in 2Q 2019 [4].

of electronic mails that a user gets per day need devoting a substantial amount of time [7]. Spam messages are not restricted to email. Many people are exposed to spam content when they visit social networks like Telegram, Facebook, Instagram, Twitter, and so on. A study revealed that more than 70% of the total internet users use these social networks and are exposed to spam content [8].

Various algorithms have been designed to solve the problem of text-based spam. At present, spammers are sending these messages in the form of an image to confuse and possibly overpower these algorithms. Image spam is a concept that began in early 2005. More than 50% of the spam was made up of images by the end of 2006 [9], [6]. Image spam is another modern challenge in a phishing email. Image spam is email spam where a text content inserted into images to confuse conventional text-based spam channels [10]. It is a complex type of spam that is tempting and strenuous for the user to notice [5], [11]. Fig. 3 shows examples of spam images.

Fig. 3. Examples of Image Spams [10].

The objective of image spam is clearly to bypass the investigation of the content of text-based email performed by the existing spam algorithms. For this reason, spammers usually include some bogus text to the email together with the attached image such as a length of words that are persuasive or cogent to surface in genuine emails and not in spam [10].

Machine Learning (ML) is a branch of artificial intelligence that involved in creating algorithms that can modify itself using structured data without human intervention to yield expected results [12]. Examples are Linear Regression, Logistic Regression, Decision Tree, Support Vector Machine (SVM), Naïve Bayes, K-Means, and Random Forest. Deep Learning (DL) is a branch of machine learning in which algorithms are developed and function similar to those in machine learning, but there are multiple layers of these algorithms, and each providing a different meaning to the data it feeds on [12]. These algorithms include the Artificial Neural Networks (ANN), Deep Neural Network (DNN) and Convolutional Neural Network (CNN) [13]. In summary, machine learning algorithms need structural data, that is they are built to learn to do things by understanding labeled data, then use it to produce further outputs with more sets of data. However, they need to be retrained through human intervention when the actual output is not the desired one. While deep learning algorithms depend on layers of the artificial neural network. They do not require human intervention as the nested layers in the neural networks put data through hierarchies of different concepts, which eventually learn through their own errors [12].

There are different types of techniques used in classifying image spam as shown in Fig. 4 [3]. These are grouped into Supervised Machine Learning, Unsupervised Machine Learning, Semi-supervised Machine Learning, Content-based Learning, and Statistical Learning. Numerous researchers utilized these approaches for phishing email classification and detection. Depending on the nature of the data to be classified, choosing suitable and appropriate techniques is exceptionally crucial. The supervised machine learning algorithms often used from the surveyed literature are Decision Tree, Fuzzy Logic, Support Vector Machine, Neural Networks, Bayesian Network, and Genetic Algorithm. Some researchers compared two or more of these techniques to see which one produces better results [14], [15]. Deep learning approaches have not been well exploited in image spam classification since their advent [16]. They have the capability to handle large datasets and can extract image features more accurately than the existing image processing techniques [5].

Unlike other survey articles, we achieve comparisons of the performance of the existing state-of-the-art image spam models. Also, this review can help researchers working in the field of image spam classification by answering the following research questions:

(a) What are the various areas of application where image spam classification has been utilized?
(b) Which publicly available datasets can be accessed for the various areas of application of image spam classification?
(c) What are the commonly used features set in the existing image spam classification models?
(d) What performance evaluation parameters are applied to determine the effectiveness of the image spam classification algorithm?
(e) What are the challenges and research directions for future researchers working in the field of image spam classification?

The organization of the paper is as follows. Section 2 review the existing literatures or related works. Section 3

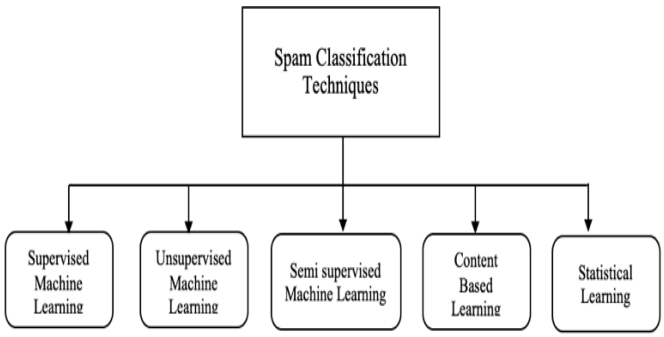


Fig. 4. Types of Techniques in Spam Classification [3].

discuss the future research directions. Section 4 gives the summary of the paper.

II. Related Works

The review of the related works is discussed under the following headings: Identification of spam classification application areas, spam classification dataset analysis and review, feature set analysis and review, and the analysis and review of spam classification techniques.

A. Identification of Spam Classification Application Areas

Basically, spams are categorized into Text-based and Image-based [5]. Spams are further divided into content-based spam and non-content-based spam. Content-based spam is the first-generation image spam [17]. This includes the spam in emails in text-based form. In this category, the extracted content from the body, headers, and keywords of emails are used by the classification algorithms to classify the images[5]. A wide range of machine learning techniques can handle this type of spam classification [6]. Non-content-based spam include complex kind of email spam and this falls into the second and third generation of image spam [17]. In this category, the undesired text is embedded in images. To classify the image spam, we can rely on the attributes of the image but recently, the advent of deep learning techniques make it possible to classify these images based on their raw byte form [5].

Images that fall in the first generation contain simple spam images hence they can be easily recognized by the optical character recognition (OCR) tools. In the second and third generation, the images contain noise and superimposing background to confuse and make them unrecognizable by the OCR. The OCR tools have the ability to partition the portions of the image that contain particular objects for the purpose of text extraction and detection [17], [5]. The background noise included with the text inside an image is a challenging task for OCR [17]. In this study, we are going to look at the application areas of spam classification under two (2) domains. Text-based and Image-based spam as shown in Table I.

B. Spam Classification Dataset Analysis and Review

This section shows the datasets that were used in spam classification and the detailed analysis. The researchers used public datasets in their works. They used one or more personal collections, Dredze, spam archive, Princeton spam corpus, image spam hunter, and so on as their datasets. For example, [33] used only Dredze dataset. The detailed analysis of data sets used in both text-based and image-based spam classification is shown in Table II and their locations in Table III.

Table II depicts the name of the datasets and sample size, the number of studies, and their references (where a specific dataset is used). This study reviewed that the Dredze dataset is the most commonly used datasets in image spam classification. This dataset consists of a total of 5789 spams (with 3239 spams and 2550 ham). Ten (10) studies adopted Dredze dataset, followed by Spam Archive dataset (with seven studies), image spam hunter (four studies), Trec07, ICDAR2003 and Char74k (two studies each), while the others datasets (Enron corpus, SMS spam, Princeton spam corpus, LingSpam, SpamAssassin and Indian corpus) have one studies each. Seventeen (17) studies used personal collection datasets from twitter. The location where the datasets can be downloaded and utilize are also presented and showed in Table III. Fig. 5 shows the name of datasets with the corresponding number of articles that adopted the datasets.

C. Feature Set Analysis and Review

This section discusses the feature sets used in all the studies under review. A feature describes the specific or distinctive attributes of image spam during processing. One of the essential steps to design efficient and accurate algorithms in spam classification is the feature extraction and selection [3]. A brief overview of these features is explained below. Table V shows the features used in image spam classification and Fig. 6 presents the graph of the number of articles versus the image features.

- Text area: This is the boundary the text occupied in an image. It is also called a text boundary. This is a way of identifying the presence of text in an image.
- Low-level (Color): These attributes are entropy values of the image RGB color, brightness, hue, and saturation. Other values include variance, skew, and the mean. The mean value represents the average pixel value of the image and it is applied to define the background of an image. In these features, there are distinct histogram attributes for a spam and ham image. Skewness is used in identifying the surfaces of an image. Spam images normally have high kurtosis values than ham images.
- Image similarity (Texture): The local binary pattern (LBP) is useful in measuring the similarity and information of adjacent pixels in an image. LBP is a powerful tool for identifying image spam which is simply text placed on a white background.
- Image region Similarity (Shape, Edge): Histogram of Oriented Gradients (HOG) determines how intensity gradient varies in an image. Edges are features used to detect spam images. It helps to identify boundaries in an image. A canny algorithm is an edge filter that is mostly used to determine the edges in an image.

TABLE I. DISTRIBUTION OF ARTICLES BASED ON THE APPLICATION DOMAINS.

Domain	No. of Studies	Reference
Text-based	9	[18], [19], [20], [21], [22], [23], [24], [25], [26]
Image-based	16	[27], [28], [29], [9], [30], [31], [32], [33], [34], [35], [17], [5], [36], [37], [38], [39]

TABLE II. DATASETS USED IN BOTH TEXT-BASED AND IMAGE-BASED SPAM CLASSIFICATION.

Dataset	No. of Studies	Sample Size	References
Spam archive	7	12053 spam (spam,=9503 & ham = 2550)	[20], [30], [34], [5], [37], [38], [39]
Dredze	10	5789 spam (spam =3239 & ham = 2550)	[30], [31], [33], [34], [40], [17], [5], [37], [28], [21]
Enron corpus	1	Not specified	[22]
SMS spam	1	Not specified	[18]
Princeton spam corpus	1	Total spam = 1004	[37]
Image spam hunter	4	1730 spam (spam =920 & ham = 810)	[21], [35], [40], [17], [5]
Trec07	2	Not specified	[21], [25]
ICDAR2003	2	11615 spam (train data =6185 & test data = 5430)	[29], [9]
Char74k	2	Total 62992 spam (train data = 44094 & test data = 18897)	[29], [9]
LingSpam	1	Not specified	[25]
SpamAssassin	1	Not specified	[25]
Personal collection & twitter	17	5326 spam (spam =3299 & ham = 2027)	[20], [24], [30], [31], [41], [32], [34], [35], [17], [40], [5], [36], [37], [38], [39], [27], [18]
Indian corpus	1	Not specified	[19]

- Image metadata: These attributes contain the depth, width, height, and compression ration of the image files. Mathematically, in an image, the compression ratio (CR) is given as:

$$CR = \frac{height * width * channels}{size\ of\ file} \quad (1)$$

- Text Obfuscation (Noise): Signal to noise ratio (SNR) and entropy of noise are the two attributes of noise. Spam images usually contain less noise than ham images. The percentage of mean to standard deviation of an image is the SNR.

Several researchers as showed in Table IV used image features to identify an image spams [30]. For instance, [30] proposed an image spam classifier using Maximum Entropy, Decision Tree, and Naïve Bayes methods. They focus only on the low level and image metadata features of the image for the classification and achieved an average accuracy of 95% with a computation time of 2.5-4.4ms. They considered a few features set for the training of the algorithm. Features reduction and elimination techniques such as principal component analysis (PCA), recursive features elimination (RFE), and univariate features selection (UFS) are very vital in optimizing or reducing the number of features in an image in order to achieve better feature classification and accuracy. Author in [35] used PCA and SVM to developed a classifier for image spam. They used a few image spam hunter and personally collected datasets to trained their classifier and claimed 70-97% accuracy. They did not take the processing time into consideration. Author in [17] used the same feature reduction and elimination approach in their work. The authors looked at 38 features of the image and used RFE and UFS to reduce the undesirable features. They employed the SVM method to train their classifier using 920 spam and 810 ham of image spam hunter dataset and 1089 spam and 1029 ham of Dredze and personal collected dataset. Accuracy of 54-98% and false-positive of 0.01-0.79 were obtained. The time taken for the classification was not considered.

D. Spam Classification Techniques Analysis and Review

Spam Email classification techniques as depicted in Fig. 4 are categorized into five (5) groups. These are supervised machine learning, unsupervised machine learning, semi-supervised machine learning, content-based learning, and statistical learning [3], [42], [43]. In supervised machine learning, input instances are given for the learning procedure and the output labels do not conveniently recognize a function that approximates this behavior. Supervised machine learning techniques include Decision Tree, Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, Bayesian Network,

TABLE III. LIST OF PUBLICLY AVAILABLE DATASETS AND THEIR CORRESPONDING LINKS.

#	Dataset	Location
1	Spam archive	https://archive.ics.uci.edu/ml/datasets/sms+spam+collection
2	Dredze	http://www.cs.jhu.edu/∼mdredze/data/
3	Enron corpus	http://www.aueb.gr/users/ion/data/enron-spam/
4	SMS spam	https://www.kaggle.com/uciml/sms-spam-collection-dataset
5	Princeton spam corpus	https://www.cs.princeton.edu/cass/spam/
6	Image spam hunter	https://users.cs.northwestern.edu/∼yga751/ML/ISH.htm#dataset
7	Trec07	http://plg.uwaterloo.ca/∼gvcormac/treccorpus07/
8	ICDAR2003	http://algoval.essex.ac.uk/icdar/Datasets.html
9	Char74k	http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/
10	LingSpam	http://www.csmining.org/index.php/ling-spam-datasets.html
11	SpamAssassin	http://spamassassin.apache.org/publiccorpus
12	Personal collection & twitter	Not available
13	Indian corpus	Not available

Random Forest, Fuzzy logic, Multilayer Perceptron, Neural Networks, and deep learning methods such as Convolution Neural Network. In unsupervised machine learning, the learning procedure is equipped with input instances but with output labels. Here, the leaning procedure tries to recognize related patterns through input instances to determine output. An example of unsupervised machine learning is k-means clustering [3]. Semi-supervised machine learning is a combination of supervised and unsupervised machine learning. In semi-supervised machine learning, some of the input datasets are labels and the learning procedure requires large labelled data. Active learning is one of the examples of semi-supervised machine learning. Content-based techniques use keywords in classifying the spam email [3]. Examples are optical character recognition (OCR) and Sobel filters. In statistical learning, each keyword is assigning a probability and the overall probability is used to classify the image spam. Supervised machine learning is the most frequently used techniques in spam classification even though researchers used all the other types of techniques. Table IX presents the distribution of spam classification techniques [3]. Thirty (30) studies adopted supervised machine learning techniques, four (4) used unsupervised techniques, eight (8) and five (5) studies adopted content-based learning and statistical learning respectively.

Fumera et al. [20] developed an algorithm for detecting and classifying text-based spam using optical character recognition (OCR) tool where they used 445 spam and 4852 ham of spam archive dataset and 5608 spam and 9526 ham of personally collected dataset to train their model using support vector machine (SVM). The authors focus only on the true positive and false positive rate and the result obtained are 0.81 and 0.01 respectively. They did not consider the time taken for the classifier to detect and classify a spam email and the method used is inefficient since it cannot handle large datasets conveniently. The proposed classifier cannot detect image spam email. The same OCR tool was used in the work of some researchers [21], [22], [24], [23]. They examined and applied OCR software to filter image spam email. While [22] used KNN, Naïve Bayes and Reverse DBSCAN in his work, [24] used Sobel operators (filters) to process the image as displayed in Table VI.

Image spam classifiers have been proposed using a near-duplicate detection approach but with different distance measurements [39], [38], [37], [36]. They both considered low level and image similarity features of the image spam in training their models. While [39] used Visual and Object Semantics as a distance measure to classified the image spam and achieved an accuracy of 96 %, [38] used Histogram and Euclidean distance measures to obtain a better result of 98% accuracy. The reason for the difference observed in the two results was because of the former used a larger dataset than the later. The computation time was not considered except in the study of [36]. The time taken to detect image spam and classify it as either spam or ham in this research is 50ms. This is displayed in table VII. Table VIII presents the keys of the abbreviations as used in Tables IV, V, VI and VII

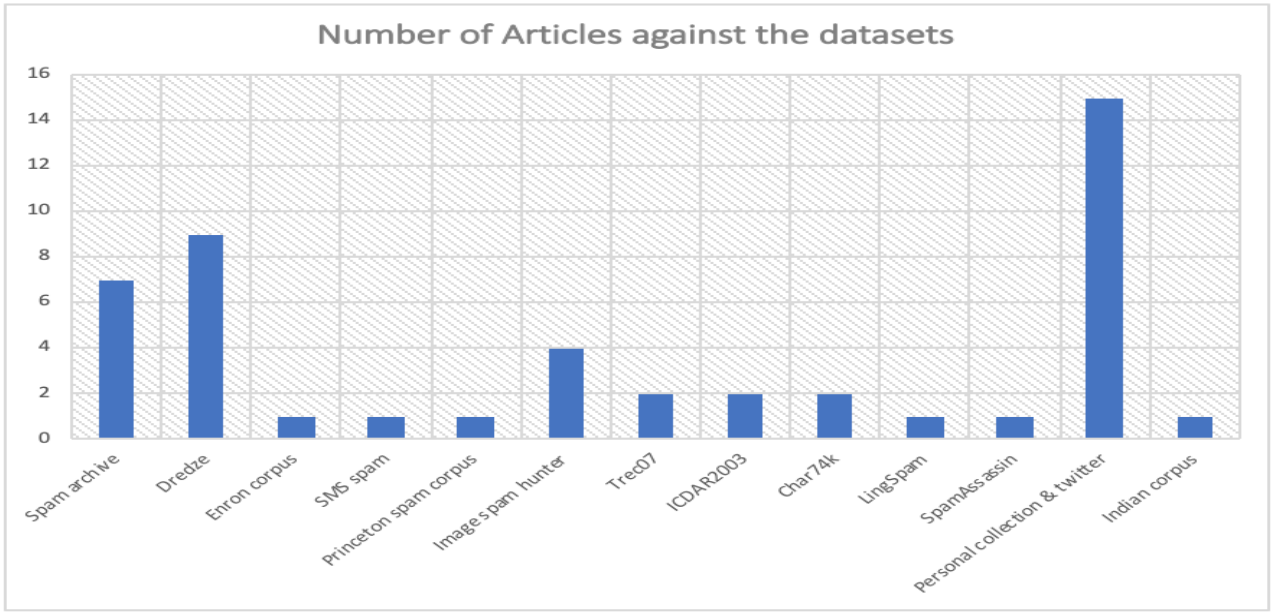


Fig. 5. Number of Articles versus Datasets

TABLE IV. RESEARCHES ON IMAGE-BASED CLASSIFICATION USING IMAGE FEATURES.

Ref	Features						Method	Dataset & Size				Results	Time(ms)	Remark
	TA	LL	IS	IRS	IM	TO		DS	Spam	Ham	Both			
[30]		y			y		Decision Tree, Naïve Bayes	d, sa pc	3239 9503	12742 2550		Acc=90 – 99	2.5 – 4.4	Few features
[31]	y	y				y	SVM	d pc	3239 8549	2550 2006		TP=0.94 - 0.98 FP=0.02 - 0.05	1200	Real-time not achieved
[32]		y			y		SVM & Active Learning	pc	1190	—	—	Acc= 99.0- 99.3	—	Few features
[33]		y				y	SVM	d	—	—		Unspecified	—	Few LL features
[34]		y			y		SVM	d sa pc	3203 9280	1786 1371		Acc=95	—	Only LL features
[35]		y	y		y		PCA, SVM	ish pc	920 1000	810		Acc=0.70 - 0.97 FP=0.04-0.25	—	Method inefficient
[17]		y			y	y	SVM	ish d pc	920 1089	810 1029		Acc=0.54 - 0.98 FP=0.01 - 0.79	—	Method inefficient
[5]		y			y	y	NN, DNN, CNN	d ish pc sa		2681	19920 1000	Acc=95.63 - 98.95	—	RFE,UFS not used

Support Vector Machine (SVM) method is one of the most commonly used classification algorithms in image spam classification [3] and has been adopted by many researchers in their works [44], [35], [17], [33], [32], [31]. SVM is suitable for binary classification problems but difficult to handle large datasets [27]. In the work of [31], in order to identify the image as spam or ham, they considered 3 features domain namely, text area, low-level features (image color), and text obfuscation (noise) of the image. They claimed to have obtained 94-98% accuracy with 1200ms computation time.

Singh [5] proposed an image spam algorithm using deep learning algorithms. They did not consider the time it took to identify and classify image spam and used only a few datasets concentrating on low level, image metadata and image obfuscation (noise) features of the image. They obtained 95.63 to 98.95% accuracy. An approach to object segmentation was not used to detect the segmented spam area. After their advent, deep learning has not been well exploited in classifying image spam. Deep learning has the ability to handle large dataset and can more accurately extract image features than existing image processing techniques [5].

Web content-based approaches can be combined with machine learning techniques to build a system for phishing website and email detection [45]. The author in [45] used this approach to designed a 92% accuracy detection system known as CANTINA+. Web structured-based method using Google PageRank has been

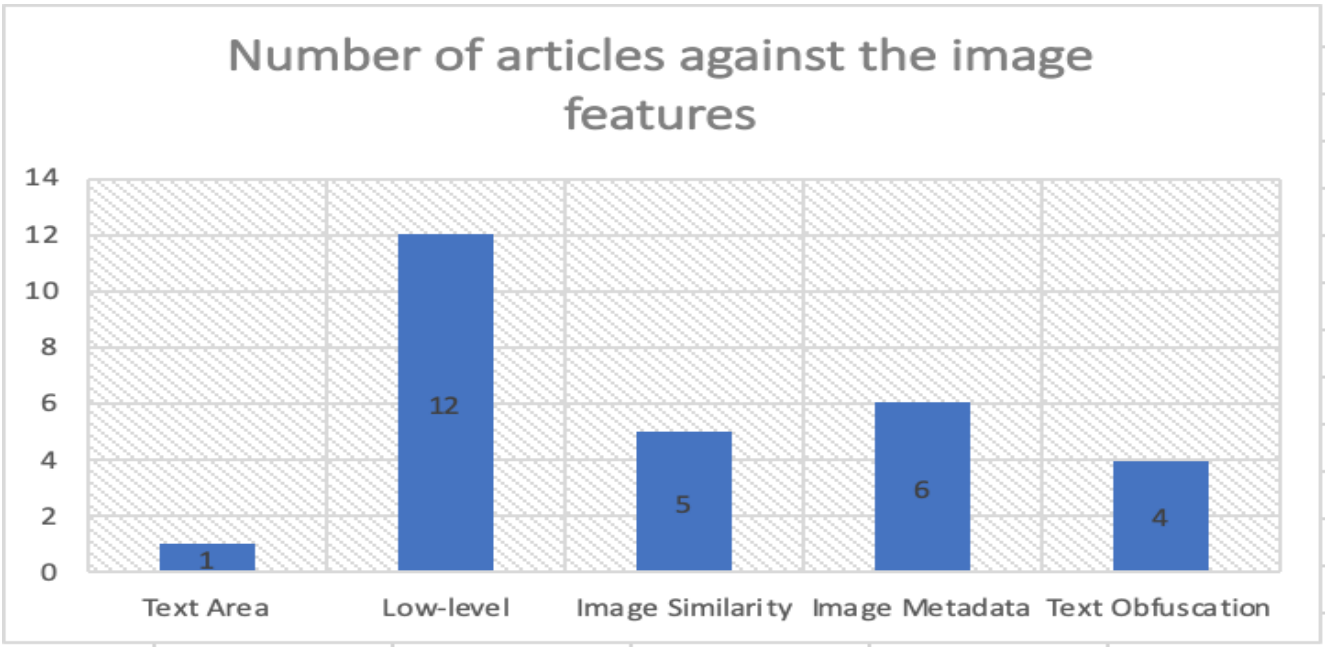


Fig. 6. Number of Articles versus Image Features.

TABLE V. FEATURES USED IN IMAGE SPAM CLASSIFICATION. SEE TABLE VIII FOR THE ABBREVIATIONS OF THE FEATURES WHERE "+" MEANS 'USED' AND "-" MEANS 'NOT USED'.

Image features						Ref
TA	LL	IS	IRS	IM	TO	
-	+	-	-	+	-	[30]
-	+	+	-	-	-	[36]
+	+	-	-	-	+	[31]
-	+	+	-	-	-	[37]
-	+	+	-	-	-	[39]
-	+	+	-	-	-	[38]
-	+	-	-	+	-	[34]
-	+	-	-	+	-	[32]
-	+	+	-	+	-	[33]
-	+	+	-	+	-	[35]
-	+	-	-	+	+	[17]
-	+	-	-	+	+	[5]
-	-	-	+	-	-	[29]
-	-	-	+	-	-	[9]
1	12	5	2	6	4	Total

TABLE VI. RESEARCHES ON TEXT-BASED CLASSIFICATION USING OCR METHOD.

Ref	Method	Dataset and size			Results	Time(ms)	Remark
		DS	Spam	Ham			
[20]	OCR, SVM	sa	445	4852	TP= 0.77-0.81	__	Method not efficient
		pc	5608	9526	FP=0.01		
[21]	OCR	Trec07, d, ish	__	__	Acc=99.83%	__	The OCR not suitable
[22]	OCR,KNN, Naïve Bayes, Reverse DBSCAN	Enron corpus	__	__	Acc= 87%	__	The OCR not reliable, restricted to certain fonts
[23]	OCR	__	__	__	__	__	OCR not suitable
[24]	Sobel filters, OCR	pc	3299	2027	Acc=45.30 - 90.12	2.6	Method not efficient

TABLE VII. RESEARCHES ON IMAGE CLASSIFICATION USING NEAR-DUPLICATE APPROACH.

Ref	Features						Distance measure	Dataset & Size			Results	Time(ms)
	TA	LL	IS	IRS	IM	TO		DS	Spam	Ham	Both	
[36]		y		y			Manhattan	pc	1071	107	TP=0.63 - 0.96 FT=0 - 0.173	50
[37]		y		y			Jensen-Shannon	sa psc, d, pc	1004	__	Acc=95 - 98 TP=0.76 - 0.84	-
[38]		y		y			Histogram, Euclidean	pc, sa	1977	8000	Acc=81-98	-
[39]		y		y			Visual and Object Semantics	sa pc	6459	1473	Acc=96.66 FP=3.34	-

used to achieve 98% accuracy in classification [46]. A Bayesian algorithm and the incremental forgetting weight algorithm were used to create a model that effectively tackled idea drift and data bias in the classification of spam emails [25]. It is possible to Combine statistical analysis of website URLs with machine learning techniques to develop a classification algorithm with a better precision rate [47].

Many researchers work on detecting and classifying email phishing but did not focus on spam emails. [14], for example, used the dataset gathered from twitter and implemented an algorithm using SVM, KNN, Random Forest, and classification features to improve the accuracy of phishing tweets detection. Their findings yield 94.75% classification accuracy with only 11 selected features, which is higher than 94.56% obtained by other researchers who used more than 11 features for the same dataset. To build a phishing detection model and solve the complexities of phishing attacks

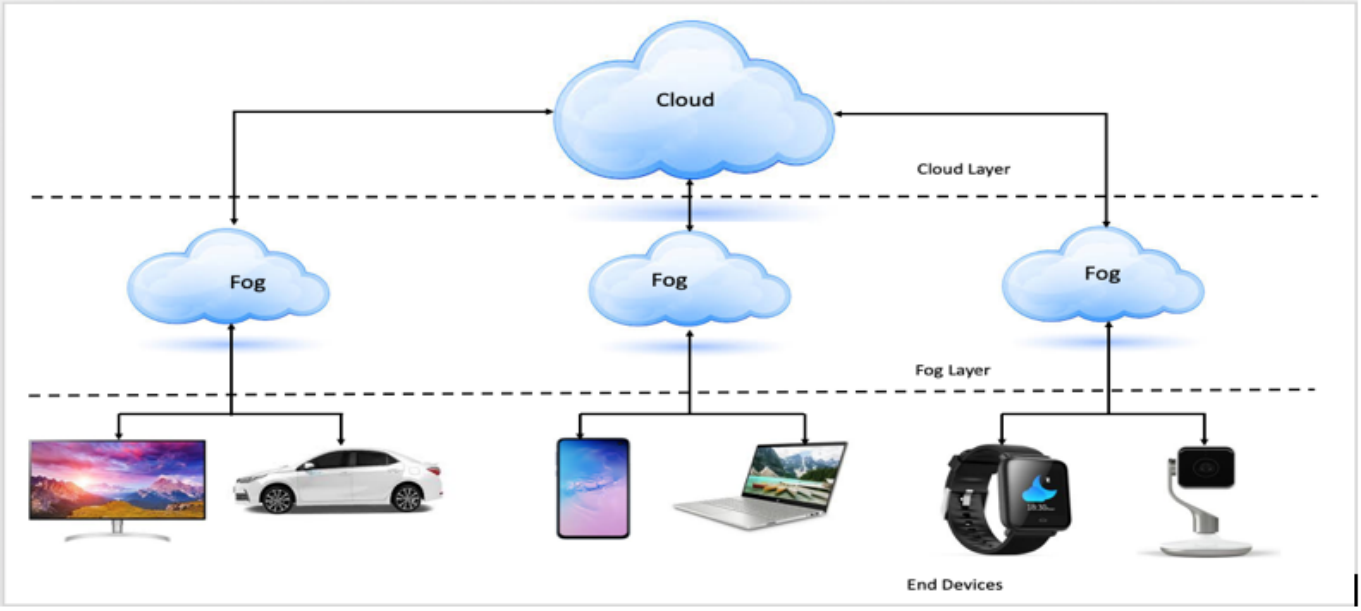


Fig. 7. Fog Computing Architecture [48].

TABLE VIII. KEYS USED IN TABLES IV, V, VI, & VII.

Features	
TA	Text area
LL	Low-level
IS	Image similarity
IRS	Image regions similarity
IM	Image metadata
TO	Text obfuscation

Datasets	
pc	Personal collection(s)
d	Dredze
sa	Spam Archive
psc	Princeton Spam Corpus
ish	Image Spam Hunter

Results	
Acc	Accuracy
TP	True positive rate
FP	False positive rate

in the real world, deep packet inspection, and software-defined networking (SDN) techniques with artificial neural networks (ANN) were applied. They reported a 98.39% accuracy and their model can provide an effective and efficient solution for detecting and minimizing phishing emails [49].

One of the hybridized approaches used in email phishing detection is neuro-fuzzy, which is the combination of fuzzy logic and neural network. [1] used this approach to developed an anti-phishing model and obtained an improved detection accuracy of 98.36%. A better result of 99.29% accuracy was obtained using the same method [50]. While [50] research did not focus on missed detection and false alarm rates, a high rate of missed detection and a false alarm was reported by [1].

In the literature, decision tree data mining techniques such as associative rule mining and classification were well used. A classification algorithm has been proposed using these methods to derive new rules from the phishing data sets [51], [52]. The main challenge with this approach is that the set of rules is not objective and largely depends on the programmer [1]. A classifier that can categorize emails written in Chinese into spam or ham based on a specific feature was created using the same method [26]. Data mining knowledge discovery procedures were used to develop an intelligent classification model that was tested using Random Forest, J48, SVM, MLP, and Bayes Net. Using the Random Forest and J48 algorithm, an accuracy of 99.1% and 98.4% was achieved respectively [53].

Convolutional Neural Network has recently been used to create a text-based spam classifier with the introduction of long short time memory neural network (LSTM NN) and an accuracy of more than 92-98% has been achieved [18]. [28], [44] used KNN and Naïve Bayes to implemented his work with the Dredze image dataset. The authors used a distributed associative memory tree to extract features of the image. This feature extraction method performs best in comparison with other distributed approaches with a relatively small amount of resources for spam detection. A 98% accuracy has been reached [28]. A Random Forest has the best accuracy, precision, recall, and F-measure than SVM and multilayer perceptron when PCA was used to construct a twitter-dataset image spam model. An accuracy of 96.3% has been achieved in this study [27].

Naiemi et al. [9] proposed a new algorithm to recognize characters in image spam by improving the existing feature extraction of HOG using SVM as the classifier. The study improved scale and translation robust HOG (STRHOG) developed with the Chars74K dataset with an accuracy of 72.2% [29]. In STRHOG, the matrices of the oriented gradient for input images of different sizes have a high computation value and a large part of this matrix does not have any effect in recognizing the image. [9] were able to overcome these problems in their work and obtained a detection accuracy of 84.91%. Some of this study's weaknesses are briefly debated. Support Vector Machine (SVM) adopted in the work is good and suitable for problems in binary classification [27]. SVM works perfectly when dealing with 2,3,4 classes but the Char74K dataset used in the work has 62 classes and is therefore a multiclass problem. Additionally, we are trying as much as possible not to lose data in machine and deep learning. In fact, generating data for any missing attribute within a dataset is advisable. In HOG, the image passes through cropping, and in the process, data is loose. Finally, the study did not consider the

time it took to detect and classify the image spam. Because of its complex computation, the canny algorithm used for the edge detection consumes a lot of time and it will be hard to implement to hit the real-time response.

In most of the reviewed articles, the computational time was not considered. Table X shows the reference of the articles that considered time in text-based and image-based spam classification.

TABLE X. LIST OF ARTICLES THAT CONSIDERED COMPUTATIONAL TIME IN SPAM CLASSIFICATION.

#	Ref	Application Domain	Computation Time (ms)
1	[24]	Text	2.6
2	[30]	Image	2.5 - 4.4
3	[31]	Image	1200
4	[36]	Image	50

E. Performance Metrics Review and Analysis

Confusion matrix (CM) as shown in Fig. 8 measure the performance of a classification algorithm in terms of accuracy, recall, precision, and F-measure. These definitions are enumerated below. CM is a matrix between True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). TP is when the image is a spam image and the classifier label it as spam. TN is when the image is a ham image and the classifier label it as a ham. FP is when the image is a ham image and the classifier label it as spam. FN is when the image is a spam image and the classifier label it as a ham [5].

Fig. 8. Confusion Matrix [5].

The often-utilized performance metrics and their formulas as highlighted in the works of [3], [55] are discussed below.

(a) Accuracy: This is the percentage of predictions that are correct. It is used to determine how well a classifier works. It is defined mathematically as:

$$Accuracy = \frac{TP + TN}{P + N} \tag{2}$$

where P = TP + FN and N = TN + FP

(b) Precision: This is the percentage of image spam classified correctly as ham. It is calculated as:

$$Precision = \frac{FPR}{100} = \frac{\frac{FP}{N}}{100} = \frac{\frac{FP}{FP+TN}}{100} \tag{3}$$

(c) Recall: This is the percentage of image spam classified correctly as spam. It is defined as:

$$Recall = \frac{TPR}{100} = \frac{\frac{TP}{N}}{100} = \frac{\frac{TP}{TP+FN}}{100} \tag{4}$$

(d) F-Measure: This is how effectively a classifier identifies positive labels. It is the weighted average of precision and recall. F-Measure is calculated as:

$$F - Measure = \frac{2 * Recall * Precision}{Recall + Precision} \tag{5}$$

(e) Simplicity: This is how effectively a classifier identifies negative labels. It is defined as:

$$Simplicity = \frac{TN}{FP + TN} \tag{6}$$

(f) Area Under Curve (AUC): This is the ability of a classifier to prevent incorrect classification. It is given as:

$$AUC = \frac{1}{2}(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}) \tag{7}$$

The performance of the existing state-of-the-art image spam models using the above metrics is shown in Table XI. The existing works considered one or more of the performance metrics. For instance, [36], [37], [31], [40], [5], [56] considered only the accuracy.

III. FUTURE RESEARCH DIRECTIONS

We discuss some of the challenges and open issues in the existing studies on image spam classification research in this section.

- Dataset: Image spam classification is a binary classification problem (ham or spam). Some of the datasets used in the reviewed articles have four or more classes and these types of datasets are suitable and work perfectly for a multiclass problem and not for binary problems. A more challenging dataset is required for future image spam classification research.
- Optical Character Recognition (OCR) Approach: Most of the existing works used the OCR technique. In the OCR method, data is lost by cropping the image during the pre-processing stage. The images don't have the same dimension and are forced to be of the same size, thereby losing some of the important data. Machine learning algorithms tries as much as possible not to lose data. In fact, it generates data for any missing attribute in a dataset. More suitable techniques are needed for the extraction of the features of an image in future research on image spam classification.
- Deep Learning Technique: The state-of-the-art image spam classifiers developed using machine learning techniques, which work with few datasets have difficulty in extracting the relevant features of the images and this has negative effects on the overall output of the classification. Deep learning models have the capability to handle large datasets and can extract image features more accurately than machine learning techniques [5]. This approach has not been well exploited in image spam classification since its advent [16]. With this in mind, the future image spam classifier can be implemented using deep learning techniques like deep neural networks, and convolutional neural networks to make the classifier more powerful and improve the performance in terms of the accuracy and precision of the classification algorithms.
- Fog Architecture: Fog Computing, also known as fog networking or fogging is a newly introduced concept. It is an internet of thing (IoT) architecture that expands the cloud so that it is closer to end devices. It supplies information, computing asset like storage and application services to the end devices. More also, at the edge of networks, fog bolsters high versatility because it pulls services given at places close to the end-users [61]. Fig. 7 shows the architecture of fog computing where it clearly depicts the three (3) layers namely, end devices (IoT) layer, fog layers, and cloud layers [48]. This concept which was recently used by [1] to detect phishing websites produced high detection accuracy. Also, the authors revealed that fog-based services are faster than cloud-based services and that it is manageable and easy

TABLE IX. SUMMARY OF SPAM CLASSIFICATION TECHNIQUES WHERE "Y" MEANS 'YES' AND "X" MEANS 'NO' IN RESPECT TO TEXT-BASED AND IMAGE-BASED CLASSIFICATION.

Category	Method	No. of Articles	Text-Based	Image-Based	Reference
Supervised Machine Learning	1. Decision Tree	2	y	y	[30], [26]
	2. Naïve Bayes	3	y	y	[22], [30], [28]
	3. Support Vector Machine (SVM)	9	y	y	[20], [31], [32], [33], [34], [35], [17], [27], [9], [44]
	4. K-Nearest Neighbor	3	y	y	[22], [28], [29], [44]
	5. Bayesian Network	1	y	x	[25]
	6. Random Forest	1	x	y	[27]
	7. Fuzzy Logic	0	x	x	-
	8. Multilayer Perceptron	1	x	y	[27]
	9. Neural Networks	2	x	y	[28], [5]
	10. Deep Neural Netwoks	1	x	y	[5]
	11. CNN	3	y	y	[5], [18], [19]
	12. CNN+LSTM	2	y	x	[19], [18]
Unsupervised Machine Learning	13. K-Means Clustering	1	x	y	[27]
	14. Reverse DBSCAN	1	y	x	[22]
	15. Manhattan Distance	1	x	y	[36]
	16. Visual and Object Semantic Distance	1	x	y	[54]
Semi-Supervised Learning	17. SVM + Active Learning	1	x	y	[32]
Content-Based Learning	18. OCR Filter	5	y	x	[20], [21], [22], [23], [24]
	19. HOG	2	x	y	[29], [9]
	20. Sobel Filter	1	y	x	[24]
Statistical Learning	21. PCA	2	x	y	[27], [35]
	22. Jensen-Shannon	1	x	y	[37]
	23. Histogram/ Euclidean Distance	1	x	y	[38]
	24. Distributed Associative Memory Tree	1	x	y	[28]

to implement a machine learning algorithm on fog nodes than on the cloud. In view of this, an algorithm can be implemented on a fog node to increase the detection speed of the image spam classification.

- Computation Time: Image spam detection and classification should be a real-time process in order to minimize response delay. In the reviewed articles, the time taken to classified the image is neglected. The canny algorithm mostly used for edge detection in the histogram of oriented gradients (HOG) method consumes a lot of time due to its complex computation. It is difficult to implement to reach the real-time response. Future research should consider reducing the processing and classification time using recent hardware technology.

IV. CONCLUSION

This study provides a thorough overview of image spam classification studies to help researchers in this field in gaining excellent knowledge and understanding of current image spam classification solutions in the major areas. Journal articles published between 2006 to 2020 on image spam detection and classification were thoroughly studied and grouped into two application domains; text-based and image-based. The selected papers were analyzed from five dimensions of rationality: spam classification application domains, datasets adopted and features sets utilized in the two application domains, the methods used, and the matrices considered for the performance evaluation. More than 50 articles on spam classification were energetically picked and examined. A comprehensive analysis of several techniques, features set, datasets, and performance evaluation metrics used in spam detection and classification were summarized. The survey revealed that Personal Collection, Dredze, and Spam Archives datasets are the most commonly adopted datasets. Similarly, low-level and image metadata features are the most widely used features sets in spam classification research. The various methods of image spam classification as pinpointed in this study are supervised machine learning, unsupervised machine learning, semi-supervised machine learning, content-based and statistical learning. Among these methods, the most commonly used is the supervised machine learning method. Support Vector Machine (SVM) provides the best performance and it is often used in supervised learning. This is followed by Naïve Bayes and K-Nearest Neighbor techniques. The commonly investigated matrices for the performance evaluation are accuracy, recall, precision, f-measure, simplicity, and confusion matrix that depicts the relationship between TP, TN, FP, and FN. Finally, we present promising directions for future research.

FUNDING

This work was kindly supported by Putra Grant Scheme under project no: 9621600

TABLE XI. PERFORMANCE OF EXISTING STATE-OF-THE-ART IMAGE SPAM CLASSIFIERS USING THE ABOVE METRICS

Ref	Performance Metrics				Method	Dataset
	Accuracy	Precision	Recall	F-Measure		
[36]	0.92	–	–	–	Nearest Neighbour (Manhattan distance)	Personal Collection
[30]	0.91 0.80 0.87	– – –	– – –	0.93 0.83 0.89	MaxEntropy Naive Bayes Decision Tree	Dredze, Spam Archive
[37]	0.97	-	-	-	Nearest Neighbour (Jansen Shannon), and SVM	Spam Archive, Dredze, Princeton, and Personal collection
[31]	0.96	–	–	–	SVM	Dredze and Personal collection
[21]	0.99 0.88 0.99 0.88 0.68	1 0.99 1 0.99 1	0.99 0.83 0.99 0.83 0.53	– – – – –	RF KNN DT Naive Bayes SVM	Dredze and Image Spam Hunter
[57]	0.94	0.95	0.99	0.97	CNN	SMS Spam and Twitter
[40]	0.98	–	–	–	SVM	Dredze
[28]	0.98	0.99	–	–	NN	Dredze
[58]	0.95	0.96	0.98	0.97	LSTM	SMS Spam and Twitter
[5]	0.99	–	–	–	CNN	Dredze
[18]	0.95	0.96	0.99	0.97	CNN and LSTM	SMS Spam and Twitter
[56]	0.92	–	–	–	SVM and Particle Swarm Optimization	Spam Archive
[59]	0.97 0.97 0.99	0.98 0.98 0.99	0.96 0.95 1	0.97 0.96 0.99	CNN	Image Spam Hunter, Dredze, Personal collection
[60]	0.79 0.96 0.99	– – –	– – –	– – –	SVM Multilayer Perceptrons CNN	Image Spam Hunter, Personal collection

ACKNOWLEDGMENT

The authors would like to acknowledge the Universiti Putra Malaysia for supporting this research.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] C. Pham, L. A. Nguyen, N. H. Tran, E.-N. Huh, and C. S. Hong, "Phishing-aware: A neuro-fuzzy approach for anti-phishing on fog networks," *IEEE Transactions on Network and Service Management*, vol. 15, no. 3, pp. 1076–1089, 2018.

[2] A.-P. W. Group, "Phish activity trends report," APWG, Tech. Rep., 3rd Quarter 2018.

[3] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed, and M. A. Al-Garadi, "Email classification research trends: Review and open issues," *IEEE Access*, vol. 5, pp. 9044–9064, 2017.

[4] A.-P. W. G. (APWG), "Phishing activity trends report 2nd quarter 2019," Tech. Rep., 2019.

[5] A. P. Singh, "Image spam classification using deep learning," Master's thesis, San Jose State University, 2018.

[6] S. Dhanaraj and V. Karthikeyani, "A study on e-mail image spam filtering techniques," in *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*. IEEE, 2013, pp. 49–55.

[7] A. Bhowmick and S. M. Hazarika, "Machine learning for e-mail spam filtering: review, techniques and trends," *arXiv preprint arXiv:1606.01042 (2016)*, 2016.

[8] J. Constine, "Facebook climbs to 1.59 billion users and crushes q4 estimates with $5.8 b revenue," *TechCrunch, (2016)*, vol. 27, 2016.

[9] F. Naiemi, V. Ghods, and H. Khalesi, "An efficient character recognition method using enhanced hog for spam image detection," *Soft Computing*, pp. 1–16, 2019.

[10] S. Gao, C. Zhang, and W.-B. Chen, "Identifying image spam authorship with variable bin-width histogram-based projective clustering," in *2011 IEEE International Conference on Multimedia and Expo*. IEEE, 2011, pp. 1–6.

[11] A. Attar, R. M. Rad, and R. E. Atani, "A survey of image spamming and filtering techniques," *Artificial Intelligence Review*, vol. 40, no. 1, pp. 71–105, 2013.

[12] A. kapoor. (2019, May) deep-learning-vs-machine-learning-a-simple-explanation. [Online]. Available: https://mc.ai/deep-learning-vs-machine-learning-a-simple-explanation/

[13] T. Guo, J. Dong, H. Li, and Y. Gao, "Simple convolutional neural network on image classification," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(*. IEEE, 2017, pp. 721–724.

[14] S. W. Liew, N. F. M. Sani, M. T. Abdullah, R. Yaakob, and M. Y. Sharum, "An effective security alert mechanism for real-time phishing tweet detection on twitter," *Computers & Security*, vol. 83, pp. 201–207, 2019.

[15] ——, "Improvement of classification features to increase phishing tweets detection accuracy." *Journal of Theoretical & Applied Information Technology*, vol. 96, no. 10, 2018.

[16] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[17] F. D. T. Aneri Chavda, Katerina Potika and M. Stamp, "Support vector machines for image spam analysis," in *Support Vector Machines for Image Spam Analysis*, vol. 1, no. 978-989-758-319-3, In Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (ICETE 2018). Copyright © 2018 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved, 2018, pp. 431–441.

[18] G. Jain, M. Sharma, and B. Agarwal, "Spam detection in social media using convolutional and long short term memory neural network," *Annals of Mathematics and Artificial Intelligence*, vol. 85, no. 1, pp. 21–44, 2019.

[19] M. S. C. S. Dongre, "E-mail spam classification using long short-

term memory method," *International Journal of Scientific Research & Engineering Trends, (2019)*, vol. 5, no. ISSN (online) 2395-566X, 2019.

[20] G. Fumera, I. Pillai, and F. Roli, "Spam filtering based on the analysis of text information embedded into images," *Journal of Machine Learning Research*, vol. 7, no. Dec, pp. 2699–2720, 2006.

[21] A. S. Manek, D. Shamini, V. H. Bhat, P. D. Shenoy, M. C. Mohan, K. Venugopal, and L. Patnaik, "Rep-etd: A repetitive preprocessing technique for embedded text detection from images in spam emails," in *2014 IEEE International Advance Computing Conference (IACC)*. IEEE, 2014, pp. 568–573.

[22] A. Harisinghaney, A. Dixit, S. Gupta, and A. Arora, "Text and image based spam email classification using knn, naïve bayes and reverse dbscan algorithm," in *2014 International Conference on Reliability Optimization and Information Technology (ICROIT)*. IEEE, 2014, pp. 153–155.

[23] D. Yamakawa and N. Yoshiura, "Applying tesseract-ocr to detection of image spam mails," in *2012 14th Asia-Pacific Network Operations and Management Symposium (APNOMS)*. IEEE, 2012, pp. 1–4.

[24] P. Wan and M. Uehara, "Spam detection using sobel operators and ocr," in *2012 26th International Conference on Advanced Information Networking and Applications Workshops*. IEEE, 2012, pp. 1017–1022.

[25] C. Jou, "Spam e-mail classification based on the ifwb algorithm," in *Asian Conference on Intelligent Information and Database Systems*. Springer, 2013, pp. 314–324.

[26] H. Chen, Y. Zhan, and Y. Li, "The application of decision tree in chinese email classification," in *2010 International Conference on Machine Learning and Cybernetics*, vol. 1. IEEE, 2010, pp. 305–308.

[27] K. S. Adewole, T. Han, W. Wu, H. Song, and A. K. Sangaiah, "Twitter spam account detection based on clustering and classification methods," *The Journal of Supercomputing*, pp. 1–36, 2018.

[28] A. Amir, B. Srinivasan, and A. I. Khan, "Distributed classification for image spam detection," *Multimedia Tools and Applications*, vol. 77, no. 11, pp. 13 249–13 278, 2018.

[29] J. Chen, H. Zhao, J. Yang, J. Zhang, T. Li, and K. Wang, "An intelligent character recognition method to filter spam images on cloud," *Soft Computing*, vol. 21, no. 3, pp. 753–763, 2017.

[30] M. Dredze, R. Gevaryahu, and A. Elias-Bachrach, "Learning fast classifiers for image spam." in *CEAS*, 2007, pp. 2007–487.

[31] B. Biggio, G. Fumera, I. Pillai, and F. Roli, "Improving image spam filtering using image text features," in *Proc of the fifth conf on email and anti-spam*, 2008.

[32] Y. Gao, A. Choudhary, and G. Hua, "A comprehensive approach to image spam detection: from server to client solution," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 826–836, 2010.

[33] M. Das and V. Prasad, "Analysis of an image spam in email based on content analysis," in *Proc. Int. Conf. On Natural Language Processing And Cognitive Computing*, vol. 201, no. 4, 2014.

[34] C. Wang, F. Zhang, F. Li, and Q. Liu, "Image spam classification based on low-level image features," in *2010 International Conference on Communications, Circuits and Systems (ICCCAS)*. IEEE, 2010, pp. 290–293.

[35] A. Annadatha and M. Stamp, "Image spam analysis and detection," *Journal of Computer Virology and Hacking Techniques*, vol. 14, no. 1, pp. 39–52, 2018.

[36] Z. Wang, W. K. Josephson, Q. Lv, M. Charikar, and K. Li, "Filtering image spam with near-duplicate detection." in *CEAS*, 2007.

[37] B. Mehta, S. Nangia, M. Gupta, and W. Nejdl, "Detecting image spam using visual features and near duplicate detection," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 497–506.

[38] P. He, X. Wen, and W. Zheng, "A simple method for filtering image spam," in *2009 Eighth IEEE/ACIS International Conference on Computer and Information Science*. IEEE, 2009, pp. 910–913.

[39] Z. Qu and Y. Zhang, "Filtering image spam using image semantics and near-duplicate detection," in *2009 Second International Conference on Intelligent Computation Technology and Automation*, vol. 1. IEEE, 2009, pp. 600–603.

[40] A. Chavda, "Image spam detection," Master's thesis, San Jose State University, 2017.

[41] Y. Gao, A. Choudhary, and G. Hua, "A nonnegative sparsity induced similarity measure with application to cluster analysis of spam images," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 5594–5597.

[42] R. Mitchell, J. Michalski, and T. Carbonell, *An artificial intelligence approach*. Springer, 2013.

[43] M. Balakumar and V. Vaidehi, "Ontology based classification and categorization of email," in *2008 International Conference on Signal Processing, Communications and Networking*. IEEE, 2008, pp. 199–202.

[44] Y. K. Zamil, S. A. Ali, and M. A. Naser, "Spam image email filtering using k-nn and svm." *International Journal of Electrical & Computer Engineering (2019), 2088-8708,*, vol. 9, no. 1, 2019.

[45] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "Cantina+: A feature-rich machine learning framework for detecting phishing web sites," *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 2, p. 21, 2011.

[46] A. N. V. Sunil and A. Sardana, "A pagerank based detection technique for phishing web sites," in *2012 IEEE Symposium on Computers & Informatics (ISCI)*. IEEE, 2012, pp. 58–63.

[47] R. Verma and K. Dyer, "On the character of phishing urls: Accurate and robust statistical learning classifiers," in *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*. ACM, 2015, pp. 111–122.

[48] H. Atlam, R. Walters, and G. Wills, "Fog computing and the internet of things: a review," *Big Data and Cognitive Computing*, vol. 2, no. 2, p. 10, 2018.

[49] T. Chin, K. Xiong, and C. Hu, "Phishlimiter: A phishing detection and mitigation approach using software-defined networking," *IEEE Access*, vol. 6, pp. 42 516–42 531, 2018.

[50] L. A. T. Nguyen and H. K. Nguyen, "Phishing identification using a novel non-rule neuro-fuzzy model," *International Journal of Computer Science and Information Security*, vol. 14, no. 4, p. 8, 2016.

[51] W. Hadi, F. Aburub, and S. Alhawari, "A new fast associative classification algorithm for detecting phishing websites," *Applied Soft Computing*, vol. 48, pp. 729–734, 2016.

[52] N. Abdelhamid, "Multi-label rules for phishing classification," *Applied Computing and Informatics*, vol. 11, no. 1, pp. 29–46, 2015.

[53] A. Yasin and A. Abuhasan, "An intelligent classification model for phishing email detection," *arXiv preprint arXiv:1608.02196 (2016)*, 2016.

[54] Z. Qu and Y. Zhang, "A new near-duplicate detection system using object semantics for filtering image spam," in *2009 International Conference on Information Management, Innovation Management and Industrial Engineering*, vol. 3. IEEE, 2009, pp. 607–610.

[55] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management*, vol. 45, no. 4, pp. 427–437, 2009.

[56] T. Kumaresan, P. Subramanian, D. S. Alex, M. T. Hussan, and B. Stalin, "Email image spam detection using fast support vector machine and fast convergence particle swarm optimization," *International Journal of Recent Technology and Engineering (IJRTE) (2019)*, vol. 8, May 2019.

[57] G. Jain, M. Sharma, and B. Agarwal, "Spam detection on social media using semantic convolutional neural network," *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)*, vol. 8, no. 1, pp. 12–26, 2018.

[58] ——, "Optimizing semantic lstm for spam detection," *International Journal of Information Technology*, vol. 11, no. 2, pp. 239–250, 2019.

[59] S. Sriram, R. Vinayakumar, V. Sowmya, M. Krichen, D. B. Noureddine, A. Shashank, and K. Soman, "Deep convolutional neural networks for image spam classification," *(2020)*, vol. hal-02510594, 2020.

[60] T. Sharmin, F. Di Troia, K. Potika, and M. Stamp, "Convolutional neural networks for image spam detection," *Information Security Journal: A Global Perspective*, vol. 29, no. 3, pp. 103–117, 2020.

[61] K. Gandhimathi and M. Vijaya, "Identifying similar web pages based on automated and user preference value using scoring methods," *International Journal of Data Mining & Knowledge Management Process*, vol. 3, no. 6, p. 41, 2013.