Ascertaining Speech Emotion using Attention-based Convolutional Neural Network Framework

Ashima Arya[1], Vaishali Arya[2], Neha Kohli[3], Namrata Sukhija[4], Ashraf Osman Ibrahim[5]*, Salil Bharany[6]*, Faisal Binzagr[7], Farkhana Binti Muchtar[8], Mohamed Mamoun[9]

Department of Computer Science and Information Technology, KIET Group of Institutions, Delhi-NCR, Ghaziabad, India[1]
Department of Computer Science and Engineering, GD Goenka University, Sohna (Gurgaon), India[2, 3]
Department of Computer Science and Engineering, SRM University, Delhi-NCR, Sonipat 131029, India[4]
Creative Advanced Machine Intelligence Research Centre, Faculty of Computing and Informatics, Universiti Malaysia Sabah[5]
Department of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India[6]
Department of Computer Science, King Abdulaziz University, P.O. Box 344, Rabigh 21911, Saudi Arabia[7]
Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia[8]
Faculty of Computer Science and Information Technology, Alzaiem Alazhari University, Khartoum North 13311, Sudan[9]

***Abstract*—Conversation among people is a profuse form of interaction that also carries emotional information. Speech input has been the subject of numerous studies over the last ten years, and it is now crucial for human-computer connection, as well as for medical care, privacy, and stimulation. This research aims to evaluate if the suggested framework can aid in speech emotion recognition (SER) activities and determine if Convolutional Neural Network (CNN) systems are efficient for SER activities using transfer learning models on spectrogram. In this investigation, the authors present a brand-new attention-based CNN framework and evaluate its efficacy against several well-known CNN architectures from earlier research. The effectiveness of the suggested system is assessed using the SAVEE dataset, an open-access resource for emotive speech, compared to famous CNN models like VGG16, InceptionV3, ResNet50, InceptionResNetV2, and Xception. The authors used stacked 10-fold cross-validation on SAVEE for all of our trials. Amongst these CNN structures, the suggested model had the greatest accuracy (87.14%), followed by VGG16 (83.19%) and InceptionResNetV2 (82.22%). Compared to contemporary techniques, the test results and evaluation show our proposed approach to have steady and impressive results.**

***Keywords*—Convolutional neural network; emotions; speech; transfer learning models; spectrogram**

I. INTRODUCTION

Automatic recognition and identification of emotions from speech signals in speech emotion recognition (SER) using machine learning is challenging [1]. SER is a quick and usual method of communication and exchanging information among humans and computers and has many real-world applications in the domain of Human-computer interaction (HCI). Feelings expressed via speech must be accurately identified and appropriately handled to provide more natural and HCI. However, building an effective SER is a complex and arduous effort due to utterance levels and abstract emotions [2]. Moreover, determining a methodologically felicitous algorithm is crucial in realizing and achieving a performance superior to the established benchmarks. The preparation and entry of sound data, obtaining features, and identifying emotions are only a few of the basic steps covered by standard SER approaches. In the quickest-growing study area currently, emotion detection in speech signals, scientists have created techniques to identify sentiments in speech signals inherently [3]. As soon as they are suggested, the concept of SER will be widely applied in the domains of schooling and medical care [4].

Although multimodal techniques can more effectively accomplish algorithmic methods for emotion identification [5], audio has been a useful medium for this job owing to the variety of data given by a person's voice [6]. Choosing a reliable approach for obtaining prominent and distinguishing characteristics from spoken words to describe the feelings of an individual speaking based on their auditory elements has become a key problem for feature mining experts. For SER, low-level handmade characteristics, including vitality, zero-crossing, length, linear classifier factor, Mel-frequency MFCC, and non-linear attributes like tiger power activator, were extensively studied during the previous ten years. Most investigations now use as a Mel-scale filtering of the financial institution voice spectrogram source characteristic when employing ML approaches for SER. CNNs often utilize spectrograms, a 2-D model representing speech sounds, to gather notable and distinct characteristics for implementation in SER [7] alongside other computational purposes [3, 8, 9]. Most two-dimensional CNNs are created specifically for optical evaluations [10, 11], and investigators have been motivated to investigate two-dimensional CNNs in the discipline of SER by their effectiveness. The CNN model may obtain excellent, important data to identify feelings in messages using spectrograms as appropriate descriptions of verbal data.

The potential of ML techniques in SER includes the systematic retrieval of emotional qualities from the unprocessed utterance and comprehending the correlations among those features. It has proven to be more effective than traditional methods. For instance, the researchers in [12] presented a combination of models constructed using long short-term memory (LSTM) and CNNs to implement temporal participation. The length of the talk hadn't been planned or

resolved, regardless of the speaker's depiction. Since feelings can fluctuate throughout a prolonged conversation, missing crucial details if spoken data has been split could influence the outcome. Leveraging continuous SER, [13] created a CNN framework with two downsampling/upsampling architectures and variable stratum compression coefficients. A system's efficiency might be impacted by fluctuating variables, leading to an overfitting issue. Consequently, the degree of complexity and feature count must be reduced to address and mitigate the issues. Recent developments in AI are also demonstrated by SER modelling with the attention system [2], transfer learning [7], as well as deep neural systems [14, 15]. They mainly referred to exceptional models for speech characteristics. The core components of SER include characteristics extraction and choosing, notwithstanding the improvement of the CNN mentioned above models. Speech sentiment may be inferred from several speech cues.

Current research develops the model using a complete methodology in light of its significance and application. As a result, it lacks a second predictor to do the categorization. Additionally, extract the emotions in communication using the attention component as a CNN layer. Additionally, because our representation uses the fourth pooling stage of the VGG-16 approach, it needs fewer components. In particular, this pooling stage collects priceless intriguing speech data, facilitating quick emotion identification. The following are the primary benefits provided by our suggested approach:

- One of the best strategies for SER that our findings suggest is a unique CNN model that combines the VGG-16 with the attention unit.
- By combining the attention and convolution modules on VGG-16, the recommended approach may identify areas of utterance that are more prone to degrade at each level.
- Since the suggested CNN approach may be taught completely, an additional filter for development and evaluation is not necessary.
- The benchmark SAVEE datasets are used to assess our model.
- The proposed technique has been assessed both qualitatively and quantitatively. The assessment's findings show that our approach beats cutting-edge techniques.

The remainder of the research paper is structured in the following fashion: Several comparable investigations are included in Section II. The suggested model is presented in Section III. The results of the study are reviewed and examined in Section IV. Section V presents our conclusion.

II. Related Work

In the modern day, the study of computational signal processing is still in its infancy. Many academics have established a variety of approaches in this field for SER during the last ten years. The two primary parts of the SER work are often separated into choosing characteristics and grouping. It is difficult to find a prejudiced choice of characteristics and grouping approach that accurately detects the speaker's feelings in this area [16]. ML algorithms are being quickly employed for SER because of the rise in information and expense processing [17, 18, 19, 20], and numerous investigators are using these techniques for reliable depiction of features in various domains [21]. Huang et al. [8] introduced a CNN-inspired strategy for SER due to its outstanding success in detecting images. In a comparable vein, [22] employed CNN to acquire excellent prejudiced characteristics based on spoken wave spectrograms and identify individuals' emotions. The Gaussian mixture method has been implemented by certain investigators [23] to determine the feelings of the speaker using reliable information.

Today, the majority of scholars derive excellent differentiation characteristics from speech recordings using two-dimensional CNNs. To discover concealed data, SER researchers are now capturing spectrograms, graphing messages concerning duration, and sending the results to CNNs [7, 24]. Additionally, researchers may use transfer learning procedures for SER by sending audio spectrograms across already trained CNN networks such as VGG [25]. To identify the feelings of the individual speaking through the SER framework, the CNNs approach can deduce excellent prejudiced characteristics from communication signals using spectrograms [26]. Likewise, to how LSTM-RNNs are continually exploited in the SER structure, hidden time-related data in messages is primarily learned using these networks [27]. ML methodologies now significantly contribute to the growth of SER curiosity.

The latest research by [28] revealed a, throughout its entirety, LSTM-DNN driven framework for SER that automatically extracts expression using unprocessed information instead of creating features manually. The combined strategy of CNN-LSTM is described in [29] for obtaining the most prominent characteristics derived from unprocessed spoken word data employing CNN and provided to the LSTM system for collecting the orderly data reminiscent of [30]. Ma et al. [31] established a framework for the model to handle varying audio lengths for SER. In this technique, CNN represented the verbal spectrogram characteristics, while RNNs processed the varying length phrases. Zhang et al. [32] introduced a method for SER using the already trained Alex-Net system for characteristics encoding and the conventional support vector machine (SVM) for feelings categorization.

To increase the identification efficiency of spoken messages, several techniques in the discipline of SER use CNN simulations with various forms of data [33]. Corresponding to this, other investigators developed an independent predictor [34] for identification. They employed the previously trained algorithm for obtaining fundamental characteristics using speech spectrograms that improve costs associated with the system data processing. In the current investigation, the authors present a brand-new attention-based CNN model that integrates the attention component alongside VGG-16 and evaluates its efficacy against a number of well-known CNN approaches from previous investigations. The following section provides a thorough description of the suggested framework with illustrations.

III. PROPOSED MODEL

The attention component and the well-known already trained CNN framework (VGG-16) are the foundation for our suggested approach. Considering two distinct explanations, the authors recommend the VGG-16 architecture. Firstly, it uses its reduced kernel shape, making it ideal for SER having fewer layers than its alternative equivalent VGG-19 approach, to gather the characteristics at the lowest level. Furthermore, it offers improved feature mining capabilities for SER identification. Among the transfer learning strategies is the tweaking strategy that authors employ. Authors employ the already trained size of ImageNet [35] in conjunction alongside the VGG-16 framework to perform the fine-tuning procedure. Since there aren't enough speeches available for learning, avoiding the overfitting issue is possible. Fig. 1 displays the suggested model's comprehensive component layout.

A. Pre-processing

An abundance of accurately tagged information is of utmost importance because SER is an identification challenge. It is crucial to select a collection of data that members of the identical group already tag. In contrast, the collection's producers, since multiple individuals, may interpret the identical words as expressing different feelings. To replicate the desired consistency, authors experimented with simulated records wherein specific performers or individuals deliver aloud a series of lines. In this study, researchers employ the SAVEE [36] dataset, which comprises precisely classified and noise-free audio specimens. There are 480 English speeches in the SAVEE database. The seven distinct emotions—neutral, angry, sad, happy, fear, disgust, and surprise—are represented by 15 phrases. Except for neutral, which contains 120 speeches, every emotion has 60 instances. Among the data collection, 60 illustrations for every feeling had been captured. Additionally, 420 speeches were recorded for the experiment's assessment. 30% of the data collection had been employed for testing, while 70% had been leveraged for training.

1) Augmentation: For each CNN framework, the dataset dimensions are a key determining element. A lack of input hampers the ability of CNN algorithms to effectively project inputs to concrete labels. A significant issue that insufficient data might cause is high variation in the assessment data forecasts. Authors employ augmentation to generate numerous training examples using the sparse audio recordings in the SAVEE dataset to get around this issue. The method implemented in this time-shifting as in Eq. (1).

$$\delta_{t_{new}} = \delta[t_{old} \pm \mu] \quad (1)$$

where, μ is the quantity of instances shifted, and δt_{old} is the audio stream. Originally recorded sound is captured at 44100 Hz in the current composition, and shifting is accomplished by s instances moving to the opposite direction.

2) Feature extraction: Selecting powerful characteristics within the voice input is crucial to achieving outstanding functionality for the proposed framework. The Mel spectrogram is a perfect feature because authors leverage CNN techniques for learning the data we provide. Fourier Transforms on recordings are used to create spectrograms, done independently to each of the smaller time portions of the audio input. Consequently, authors are presented with a frequency vs. time chart where the hue of the spectrogram represents the harmonic's intensity. Individuals, on the other hand, interpret frequencies exponentially instead of linear. This issue is resolved by the Mel magnitude, which converts a tone's perceptual pitch to its actual pitch.

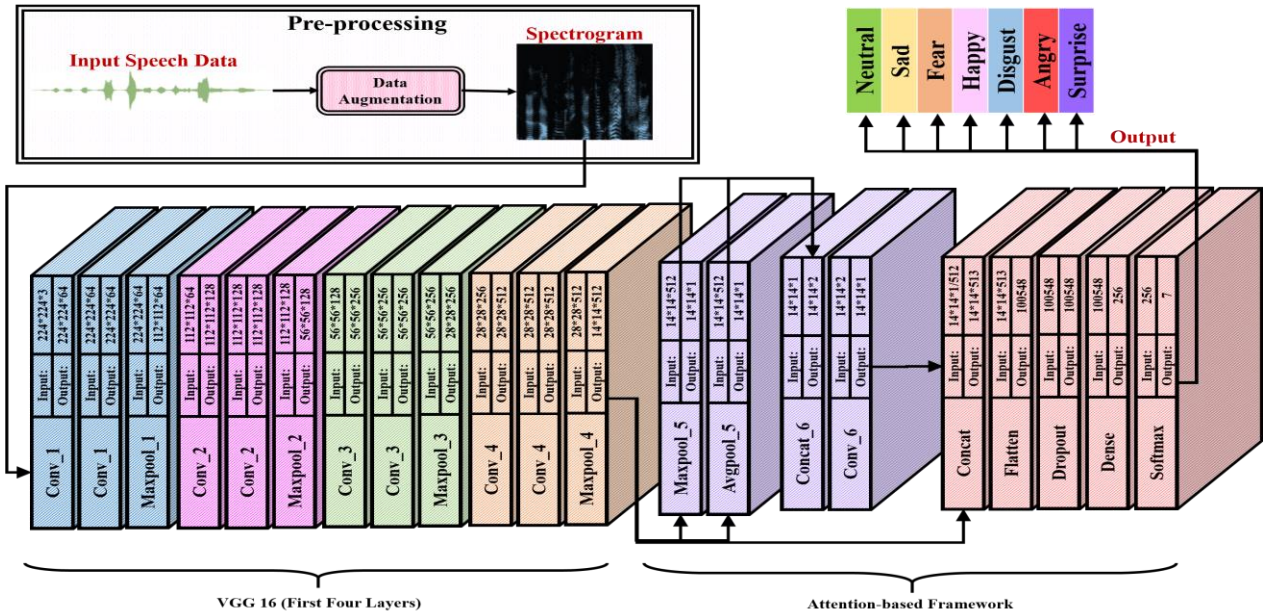

Fig. 1. The proposed architecture.

B. Attention-based Framework

Authors adopt the attention idea put forward by Woo et al. [37] for the proposed framework. The source tensor, which corresponds to the fourth pooling section of the VGG-16 framework utilized by the proposed technique, is subjected to maximum and average pooling operations. Then, employing the Sigmoid operation, these two resulting tensors are fused to conduct a convolution with a filtration value (ρ) of seven*seven. Fig. 1 displays the attention unit's layout. Eq. (2) defines the conjugated consequent tensor $(\partial_\mu(\beta))$.

$$\partial_\mu(\beta) = \theta(\rho[\beta_\mu^{avg}, \beta_\mu^{max}]) \quad (2)$$

where, the two-dimensional tensors obtained by average pooling and maximum pooling operations on the source tensor β are denoted by β_μ^{avg} and β_μ^{max}, respectively.

The fourth pooling section of the VGG-16 approach is employed in our approach. The scale-invariant component captures the intriguing hints of the picture. The midlevel area, or fourth pooling, better suitable for spectrogram, is where the intriguing hints are recovered. However, since spectrogram pictures are neither broader nor particular, the characteristics from other levels are inappropriate for spectrogram. As a result, the authors start by giving the attention component from the output of the fourth pooling layer. The production of the corresponding module is combined with the actual fourth pooling layer. Authors employ completely linked layers for expressing the concatenated characteristics obtained from the attention and convolution phase as a one-dimensional characteristic. In the proposed approach, the dense layer is set at 256, and the dropout is fixed at 0.5. The softmax layer groups the characteristics taken from the previous layers. The amount of groups in the softmax layer determines the measurement of the amount. The softmax layer produces the multinomial variation in likelihood ratings depending on the accomplished grouping. Eq. (3) defines the outcome of this distribution.

$$\varphi(x = z|y) = {e^{y_i}}/{\sum_n e^{y_n}} \quad (3)$$

where, y and z indicate the likelihoods that the softmax layer has been calculated and, correspondingly, one of the data categories employed by the suggested technique.

IV. Experiment and Results

A. State-of-the-art CNN Models

Five distinct CNN frameworks, including VGG16, InceptionV3, ResNet50, InceptionResNetV2, and Xception, have been employed in the present investigation. Among those most renowned and well-liked CNN designs is VGGnet. The distinctive VGGnet architecture consists of 138–144 million variables, approximately nineteen convolutional layers, Three*three convolutional filtering, five max–pooling stages, three completely linked layers, and a classification level as the final level [38]. By increasing its depth and breadth, Inceptionv3 is used to improve the processing capacity [39]. There are forty-eight layers in the design. The recommended framework is iterated with max-pooling to decrease the number of variables. ResNet is a standard feed-forward system with a residual link, in addition. The $(\gamma - 1)$ th results of the preceding level, also known as $(\gamma_t - 1)$, would be used to generate the residual layer outcome. The result of various procedures, including the convolution with various filtering widths and batch normalization accompanied by an activation operation on $(\gamma_t - 1)$, is referenced as $\sigma(\gamma_t - 1)$. Eq. (4) [38] can be employed to determine the residual section's ultimate result, γ_t.

$$\gamma_t = \sigma(\gamma_t - 1) + \gamma_t - 1$$

Each of the various fundamental residual blocks makes up the residual system. However, the tasks performed in the residual block fluctuate due to the various topologies of residual systems. A residual system with fifty levels is referred to as ResNet50. The InceptionResNetV2 is an amalgamation of 164-layer inception architectures featuring a residual link. To avoid any associated deterioration issues, the system uses multiple-sized CNNs that undergo training on various pictures [38]. "Extreme inception" is the abbreviation for the CNN structure known as Xception. The Xception design is a linear pile of residually connected separable by depth levels. Its 36 convolutional levels serve as the system's extraction of characteristics foundation [40].

B. K-Fold Cross-Validation

A popular method for determining the genuine forecasting errors of networks and fine-tuning the system's characteristics [41] to avoid generalization mistakes is cross-validation. Given the inconsistency in data collection, numerous models commonly encounter the overfitting problem. This outstanding approach is widely employed to address this issue [42]. The training information needs to be divided into K sections, each including an n/k specimen, wherein n is the initial sample amount, to begin the K-fold cross-validation operation. As a result, k-1 portions are employed in learning, whereas the residual portions are exploited in validation [43, 44]. The grid-search technique in our suggested method incorporates this significant strategy. The present research additionally employed the holdout approach, particularly a 3-split holdout, which is a way of partitioning the samples into several parts and engaging one part to learn the mathematical framework alongside additional parts for validating and testing the predictions. Additionally, the 10-fold cross-validation is used for several CNN networks; the study will go into more depth about the outcomes later [45-48].

C. Results

The performance of the proposed model and various CNN models is presented in Table I.

TABLE I. Performance of Various CNN Models

	Accuracy	***Precision***	***Recall***	***F1-score***
VGG16	83.19	90.68	90.97	90.82
InceptionV3	81.34	90.63	88.8	89.71
ResNet50	80.74	88.58	90.13	89.34
InceptionResNetV2	82.22	90.31	90.17	90.24
Xception	80.99	89.22	89.78	89.5
Proposed Model	87.14	92.85	93.42	93.13

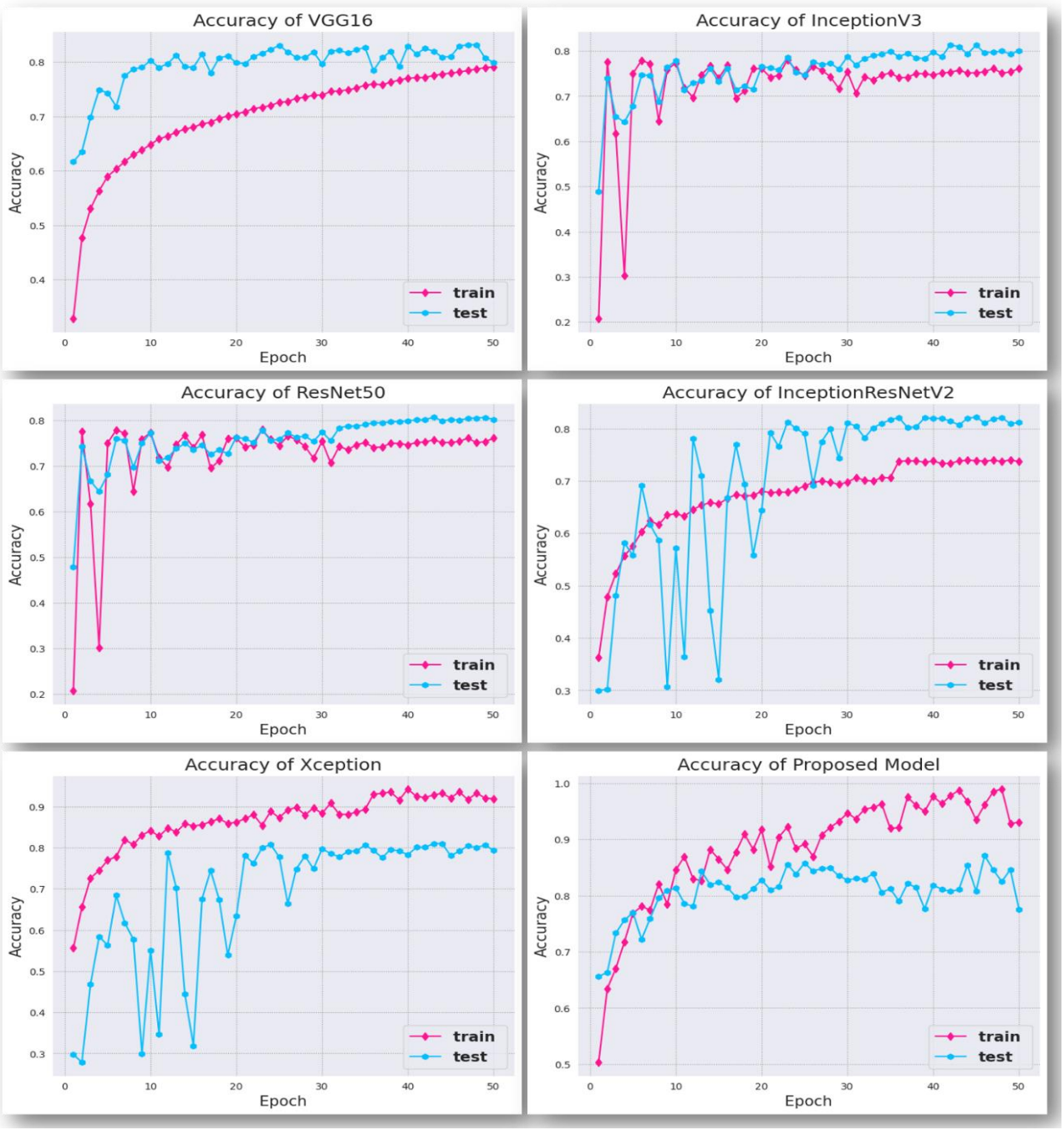

Fig. 2. Accuracy graph of various CNN models.

The proposed model achieved the highest performance with 87.14% accuracy, 92.85% precision, 93.42% recall and 93.13% f1-score. The second-best performance is demonstrated by VGG16 (83.19% accuracy, 90.68% precision, 90.97% recall and 90.24% f1-score) followed by InceptionResNetV2 (82.22% accuracy, 90.31% precision, 90.17% recall and 93.13% f1-score) and then InceptionV3 (81.34% accuracy, 90.63% precision, 88.8% recall and 89.71% f1-score).

ResNet50 observes the last performance among the underlying models with 80.74% accuracy, 88.58% precision, 90.13% recall and 89.34% f1-score. The accuracy graph, loss graph and confusion matrix for the various models are depicted in Fig. 2, Fig. 3 and Fig. 4.

The accuracy grows significantly in the initial few epochs, as seen in the Fig. 2, showing that the system is acquiring knowledge quickly. Following that, the trajectory becomes flatter, implying that there aren't sufficient epochs necessary for refining the simulation anymore. Overfitting occurs when the initial data precision improves, but the test accuracy deteriorates. It means that the framework has begun to remember the data. Inception V3 and ResNet 50 have been observed to have the minimum overfitting issue [49, 50].

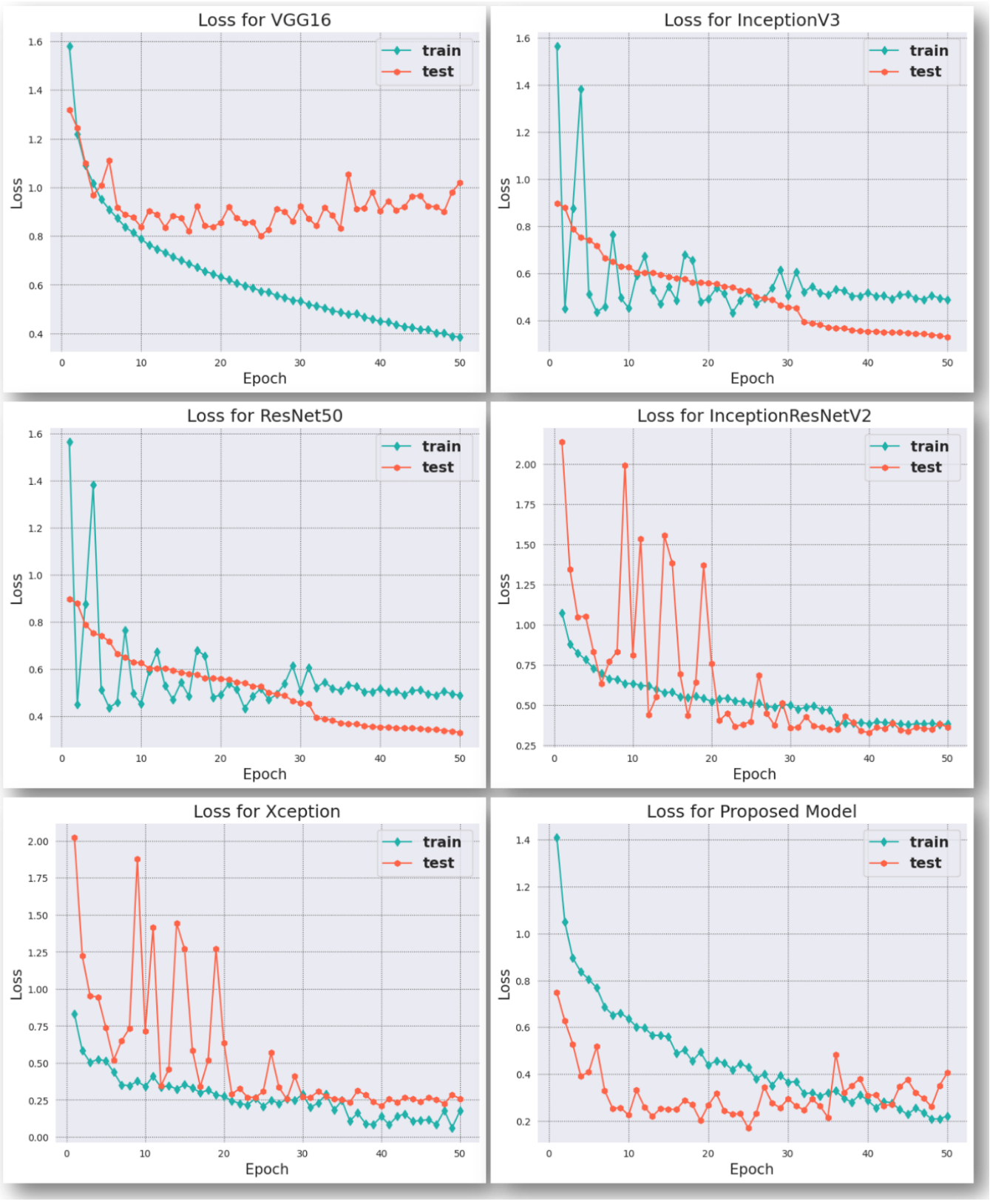

Fig. 3. Loss graph of various CNN models.

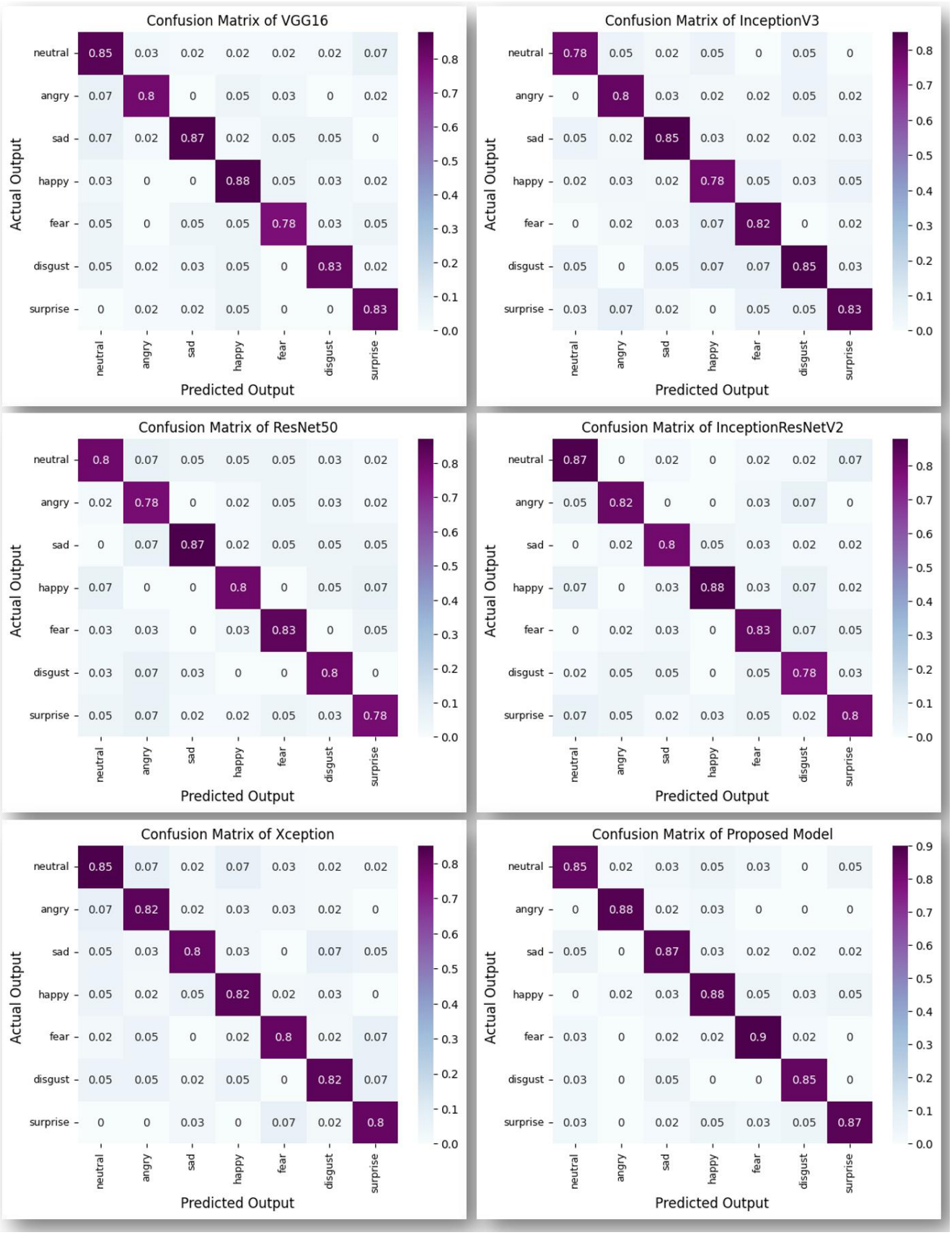

Fig. 4. Confusion matrix of different CNN models.

The loss on the learning set falls significantly during the initial few epochs, as seen in the Fig. 3. The loss in the test dataset is not declining at a comparable pace as in the preliminary set, but maintains nearly constant throughout numerous epochs. This indicates that the proposed framework generalizes effectively to new inputs.

The confusion matrix shows if the framework is "confused" in distinguishing among the various classes. It resembles a two-dimensional matrix, illustrated in the Fig. 4. The proposed model is observed to provide best results in comparison to the other prevailing models in the existing literature.

In future, authors aim to extend this research to incorporate the ensemble of various models to be fed to the attention module and compare it with existing research.

V. CONCLUSION

A key challenge to improving the model's accuracy compared to industry standards is creating a rigorous approach to gather relevant speech characteristics. Acquiring and evaluating voice elements for emotion recognition in spoken language may be challenging. The key challenges in designing a SER framework are extracting valuable characteristics and properly classifying those traits. However, developing contemporary ML techniques reduces the difficulty associated with these complicated activities. As a result, authors developed a novel SER model using attention-based CNN units, which acquire significant characteristics simultaneously with those required to group feelings. The SAVEE dataset, an open-access resource for emotional speech, is used to compare the performance of the proposed system against well-known CNN models such as VGG16, InceptionV3, ResNet50, InceptionResNetV2, and Xception. For all our experiments, the authors employed stacked 10-fold cross-validation on SAVEE. The recommended model exhibited the highest accuracy (87.14%) among these CNN architectures, followed by VGG16 (83.19%) and InceptionResNetV2 (82.22%). The test findings and assessment reveal that our suggested strategy has consistent and excellent outcomes compared to modern methods.

REFERENCES

[1] B. Liu, H. Qin, Y. Gong, W. Ge, M. Xia and L. Shi, "EERA-ASR: An energy-efficient reconfigurable architecture for automatic speech recognition with hybrid DNN and approximate computing," *IEEE Access*, vol. 6, pp. 52227-52237, 2018.

[2] Y. Zhang, J. Du, Z. Wang, J. Zhang and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Honolulu, HI, USA, 2018.

[3] Mustaqeem and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, p. 183, 2019.

[4] M. Swain, A. Routray and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, pp. 93-120, 2018.

[5] J. Han, Z. Zhang, Z. Ren and B. Schuller, "EmoBed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 553-564, 2019.

[6] B. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90-99, 2018.

[7] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl and B. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proceedings of the 25th ACM international conference on Multimedia*, California, USA , 2017.

[8] J. Huang, B. Chen, B. Yao and W. He, "ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network," *IEEE access*, vol. 7, pp. 92871-92880, 2019.

[9] S. Kumar, P. Goswami and S. Batra, "Fuzzy Rank-Based Ensemble Model for Accurate Diagnosis of Osteoporosis in Knee Radiographs," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 4, 2023.

[10] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. Baik and V. de Albuquerque, "Cloud-assisted multiview video summarization using CNN and bidirectional LSTM," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 77-86, 2019.

[11] S. Khan, I. Haq, S. Rho, S. Baik and M. Lee, "Cover the violence: A novel Deep-Learning-Based approach towards violence-detection in movies," *Applied Sciences*, vol. 9, no. 22, p. 4963, 2019.

[12] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Piscataway, NJ, USA, 2016.

[13] S. Khorram, Z. Aldeneh, D. Dimitriadis, M. McInnis and E. Provost, "Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition," *arXiv preprint arXiv:1708.07050*, 2017.

[14] M. Lech, M. Stolar, C. Best and R. Bolia, "Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding," *Frontiers in Computer Science*, vol. 2, p. 14, 2020.

[15] J. Zhao, X. Mao and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical signal processing and control*, vol. 47, pp. 312-323, 2019.

[16] S. Jiang, Z. Li, P. Zhou and M. Li, "Memento: An emotion-driven lifelogging system with wearables," *ACM Transactions on Sensor Networks (TOSN)*, vol. 15, no. 1, pp. 1-23, 2019.

[17] H. Wang, Q. Zhang, J. Wu, S. Pan and Y. Chen, "Time series feature learning with labeled and unlabeled data," *Pattern Recognition*, vol. 89, pp. 55-66, 2019.

[18] S. Batra, R. Khurana, M. Z. Khan, W. Boulila, A. Koubaa and P. Srivastava, "A Pragmatic Ensemble Strategy for Missing Values Imputation in Health Records," *Entropy*, vol. 24, no. 4, p. 533, 2022.

[19] S. Batra and S. Sachdeva, "Organizing standardized electronic healthcare records data for mining," *Health Policy and Technology*, vol. 5, no. 3, pp. 226-242, 2016.

[20] A. Pathak, S. Batra and V. Sharma, "An Assessment of the Missing Data Imputation Techniques for COVID-19 Data," in *Proceedings of 3rd International Conference on Machine Learning, Advances in Computing, Renewable Energy and Communication: MARC 2021*, Singapore, 2022.

[21] R. Khalil, E. Jones, M. Babar, T. Jan, M. Zafar and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327-117345, 2019.

[22] A. Khamparia, D. Gupta, N. Nguyen, A. Khanna, B. Pandey and P. Tiwari, "Sound classification using convolutional neural network and tensor deep stacking network," *IEEE Access*, vol. 7, pp. 7717-7727, 2019.

[23] M. Navyasri, R. RajeswarRao, A. DaveeduRaju and M. Ramakrishnamurthy, "Robust features for emotion recognition from speech by using Gaussian mixture model classification," in *Information and Communication Technology for Intelligent Systems (ICTIS 2017)-Volume 2 2*, Cham, Switzerland, 2018.

[24] S. Batra, H. Sharma, W. Boulila, V. Arya, P. Srivastava, M. Z. Khan and M. Krichen, "An Intelligent Sensor Based Decision Support System for Diagnosing Pulmonary Ailment through Standardized Chest X-ray Scans," *Sensors*, vol. 22, no. 19, p. Sensors, 2022.

[25] A. Krizhevsky, I. Sutskever and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, p. 1097–1105, 2012.

[26] E. Ocquaye, Q. Mao, H. Song, G. Xu and Y. Xue, "Dual exclusive attentive transfer for unsupervised deep convolutional domain adaptation in speech emotion recognition," *IEEE Access*, vol. 7, pp. 93847-93857, 2019.

[27] M. Zeng and N. Xiao, "Effective combination of DenseNet and BiLSTM for keyword spotting," *IEEE Access*, vol. 7, pp. 10767-10775, 2019.

[28] T. Sainath, O. Vinyals, A. Senior and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, South Brisbane, QLD, Australia, 2015.

[29] P. Tzirakis, G. Trigeorgis, M. Nicolaou, B. Schuller and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of selected topics in signal processing*, vol. 11, no. 8, pp. 1301-1309, 2017.

[30] X. Ma, H. Yang, Q. Chen, D. Huang and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, Amsterdam, The Netherlands, 2016.

[31] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng and L. Cai, "Emotion recognition from variable-length speech segments using deep learning on spectrograms," *Interspeech*, pp. 3683-3687, 2018.

[32] S. Zhang, S. Zhang, T. Huang and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576-1590, 2017.

[33] Q. Mao, M. Dong, Z. Huang and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE transactions on multimedia*, vol. 16, no. 8, pp. 2203-2213, 2014.

[34] P. Liu, K. Choo, L. Wang and F. Huang, "SVM or deep learning? A comparative study on remote sensing image classification," *Soft Computing*, vol. 21, pp. 7053-7065, 2017.

[35] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009.

[36] "Surrey Audio-Visual Expressed Emotion (SAVEE) Database," [Online]. Available: http://kahlan.eps.surrey.ac.uk/savee/. [Accessed 19 June 2023].

[37] S. Woo, J. Park, J. Lee and I. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, Munich, Germany, 2018.

[38] D. Theckedath and R. Sedamkar, "Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks," *SN Computer Science*, vol. 1, pp. 1-7, 2020.

[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[40] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[41] L. Yates, Z. Aandahl, S. Richards and B. Brook, "Cross validation for model selection: a review with examples from ecology," *Ecological Monographs*, vol. 93, no. 1, p. e1557, 2023.

[42] M. Kaariainen, "Semi-supervised model selection based on cross-validation," in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, Montreal, QC, Canada, 2006.

[43] D. Anguita, A. Ghio, S. Ridella and D. Sterpi, "K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines," in *DMIN*, Shenzhen, China, 2009.

[44] Ibrahim, A.O., Shamsuddin, S.M., Abraham, A. and Qasem, S.N., 2019. Adaptive memetic method of multi-objective genetic evolutionary algorithm for backpropagation neural network. Neural Computing and Applications, 31, pp.4945-4962.

[45] Bharany, S.; Sharma, S. "Intelligent Green Internet of Things: An Investigation." In Machine Learning, Blockchain, and Cyber Security in Smart Environments, Chapman and Hall: London, UK; CRC: Boca Raton, FL, USA, 2022; pp. 1–15.

[46] S. Alam et al., "Blockchain-Based Solutions Supporting Reliable Healthcare for Fog Computing and Internet of Medical Things (IoMT) Integration," Sustainability, vol. 14, no. 22. MDPI AG, p. 15312, Nov. 18, 2022. doi: 10.3390/su142215312.

[47] S. Bharany, S. Sharma, N. Alsharabi, E. Tag Eldin, and N. A. Ghamry, "Energy-efficient clustering protocol for underwater wireless sensor networks using optimized glowworm swarm optimization," Frontiers in Marine Science, vol. 10. Frontiers Media SA, Feb. 02, 2023. doi: 10.3389/fmars.2023.1117787.

[48] E. M. Onyema et al., "A Security Policy Protocol for Detection and Prevention of Internet Control Message Protocol Attacks in Software Defined Networks," Sustainability, vol. 14, no. 19. MDPI AG, p. 11950, Sep. 22, 2022. doi: 10.3390/su141911950.

[49] E. A. Adeniyi, P. B. Falola, M. S. Maashi, M. Aljebreen, and S. Bharany, "Secure Sensitive Data Sharing Using RSA and ElGamal Cryptographic Algorithms with Hash Functions," Information, vol. 13, no. 10. MDPI AG, p. 442, Sep. 20, 2022. doi: 10.3390/info13100442.

[50] A. Sundas, S. Badotra, S. Bharany, A. Almogren, E. M. Tag-ElDin, and A. U. Rehman, "HealthGuard: An Intelligent Healthcare System Security Framework Based on Machine Learning," Sustainability, vol. 14, no. 19. MDPI AG, p. 11934, Sep. 22, 2022. doi: 10.3390/su141911934.