WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Improving Speech Emotion Recognition and Classification Accuracy Using Hybrid CNN-LSTM-KNN Model

Md. Imran Hossain [1, a], Md. Mojahidul Islam[2, b,*], Tania Nahrin[1, c], Md. Rashed[1,d], Md. Atiqur Rahman[2,e]

[1]Department of ICE, Pabna University of Science and Technology, Rajapur, Pabna-6600, Bangladesh.
[2]Department of CSE, Islamic University, Kushtia-7003, Bangladesh.
Email: [a]imran05ice@pust.ac.bd, [b]mojahid.cse@gmail.com*, [c]tanianahrin9@gmail.com, [d]rashedulislam.200613@s.pust.ac.bd, [e]atique@cse.iu.ac.bd
Corresponding Author Email : [b]mojahid.cse@gmail.com*,

A B S T R A C T

Emotion recognition from speech is a critical component of human-computer interaction and has numerous applications in various domains, including healthcare, customer service, and entertainment. Speech Emotion Recognition (SER) is complex and challenging because different speakers exhibit distinct emotional nuances. The functionality of SER systems heavily relies on the features extracted from speech signals. Creating effective feature extraction and classification models remains a difficult task. This study proposes a hybrid model to extract important information precisely and improve predictions with better probability. Initially, Convolutional Neural Networks (CNN) is employed to extract relevant acoustic features from the speech spectrogram, capturing both local and global patterns. These features are then processed by Long Short-Term Memory (LSTM) networks, which effectively capture temporal dependencies in the speech signal. The k-Nearest Neighbors (KNN) algorithm is used to classify emotional states. It is applied to the extracted features to make final emotion predictions. Samples are first preprocessed using dataset balance and data improvement techniques to enhance the speech. Our model is trained and tested in this study using the TESS dataset. Experimental results demonstrate the effectiveness of the proposed CNN-LSTM-KNN model in achieving high accuracy and robust performance in SER. The model outperforms traditional methods and shows promise in real-world applications, including human-computer interaction systems, sentiment analysis, and virtual assistants. The findings of this study contribute to the advancement of speech-emotion recognition techniques and pave the way for enhanced human-computer interaction experiences.

Keywords: Speech Emotion Recognition, Convolutional Neural Networks, Long short-term memory, k-nearest Neighbors.

1. INTRODUCTION

Voice signals, the most natural and useful form of human communication, are composed of a multitude of non-linguistic cues such as facial expressions, speech emotions, and linguistic information, including language type and semantics. The speech-emotion recognition (SER) field has gained prominence recently as artificial intelligence has advanced [1]. It focuses on developing algorithms and systems that automatically detect and understand emotional cues within speech signals. As human communication is based on emotional information, speech-emotion recognition has grown in importance in the field of human-computer interaction [2]. Emotional cues in speech play a critical role in everyday conversations, enabling individuals to express joy, sadness, anger, fear, and a wide spectrum of feelings. These cues not only provide additional layers of meaning but also facilitate empathy, engagement, and effective communication [3].

Machines employ SER to identify human emotions from speech. Many practical uses of SER exist when implemented effectively, including sentiment analysis, human-computer interface (HCI), mental health monitoring, and suicide prevention [4]. Research in this area focused on acoustic features and basic machine-learning techniques to detect emotional states. Researchers explored features such as pitch, speech rate, and spectral properties to identify emotional cues. Early systems were limited in their scope and faced challenges in handling the complexity and subjectivity of emotional expression.

The field benefited from integrating deep learning techniques, particularly neural networks, as technology advanced. These models have shown remarkable capabilities in capturing intricate patterns and nuances in emotional expression from speech signals. As a result, the accuracy and robustness of SER systems have improved, making them more suitable for real-world applications.

We propose a hybrid approach using a CNN-LSTM-KNN model. This hybrid model leverages the strengths of deep learning, sequence modelling, and traditional machine learning techniques to enhance the accuracy and robustness of emotion recognition. The CNN component is well-suited for extracting features from audio spectrograms, enabling it to capture temporal patterns and spectral features that are relevant to emotional expression in

speech signals. LSTM networks excel in modelling sequential information, effectively capturing temporal dependencies within speech data. This capability is crucial for recognizing emotional patterns. The KNN algorithm is used to classify emotional states. It is applied to the extracted features to make predictions. The utilization of TESS data enables comprehensive testing and evaluation of the proposed CNN-LSTM-KNN model, considering various emotional expressions and the challenges associated with variability in emotional expression.

2. LITERATURE REVIEW

Speech is one of the most efficient methods to transmit emotions. A person's emotional state affects how they interact and behave, which significantly impacts verbal and visual communication, including facial expression and voice characteristics. In order to better explain the manifestation of emotions that are frequently experienced in enhanced contact with machines, SER is being incorporated into increasingly complicated natural man-machine interfaces. Three crucial phases are usually involved in SER systems: preprocessing, feature extraction, and classification. The foundation of the overall recognition process consists of these steps. Mel spectrogram is an efficient feature extraction approach that may be used to capture subjective information and solve SER. There are several ways that spectrogram can be applied to speech analysis, such as speaker identification [2], speech enhancement [5], speech recognition [6], voice and speech separation [7, 8] and speech emotion detection [3] are a few of these uses.

Deep learning-based techniques are also beginning to be used instead of statistics and other machine-learning techniques to complete the classification task. The effective setup makes use of a deep neural network (DNN), a quickly developing method for figuring out the statistical properties of conventional discourse-level analysis. Therefore, the support vector machine (SVM) and DNN were merged to increase the recognition accuracy of the classical classifier. For instance, the authors [9] and [10] employ recursive neural networks (RNNs) and deep feedforward networks in the fully connected technique to learn short-term auditory properties. The network is first constructed at the framework level, and then, with the aid of extreme learning machines, conventional mapping approaches are applied to sentence-level representations. One of the most widely used methods for supervised learning is SVM [11], which is widely used for binary and multiclass classification of emotions in speech. The core principle of SVM is centred on linear data classification.

In the SER field, various deep learning methods are applied, including DNN, CNN, Recurrent Neural Networks (RNN), and the recently introduced Transformer model [12]. In a study conducted by [13], a novel model employing a layer-based 2D Convolutional Neural Network (CNN) was proposed for speech recognition. The authors utilized a combination of Mel Frequency Cepstral Coefficient (MFCC) and Log-Mel Spectrogram (LMS) to represent speech signals. The classification of these features into different emotional states was accomplished through CNN. The model's performance was evaluated using two widely recognized datasets, RAVDSESS and TESS. Notably, incorporating a joint feature of MFCC and LMS resulted in a significant performance improvement.

In their work, [14] proposed an approach incorporating a Bi-directional Long Short-Term Memory (BiLSTM) and CNN model for the recognition and classification of emotions. The authors systematically considered various acoustic features to augment the accuracy of the classification process, employing the RAVDESS and CREMAD datasets for both training and testing. Notably, the model demonstrated a testing accuracy of 83.06% over the RAVDESS dataset; however, this accuracy declined to 63.76% when applied to the CREMA-D dataset. In their research, [15] developed a 1D-CNN model utilizing the primary features of MFCC. To address the potential overfitting issue, the authors employed data augmentation techniques. The proposed model achieved a recognition accuracy of 83%.

Additionally, [16] presented a deep neural network based on speech spectrum emotion recognition. The model combines CNN for feature extraction and LSTM for temporal information. The University of Southern California's Interactive Emotion Capture (IEMOCAP) dataset was chosen for data collection. The final weighted accuracy of the model was reported as 61%, with an unweighted accuracy of 56%. According to [17], the authors present a systematic and robust approach employing a 1D CNN to implement an emotion recognition system tailored for the low-resource language Persian. The Sharif Emotional Speech Database (ShEMO) was utilized in this study. The data underwent initial processing through the MFCC feature extraction method, with the extracted MFCC serving as input features for the neural network. The proposed method demonstrated a classification accuracy of approximately 74%.

In their work, [18] introduced a hybrid model that amalgamates two widely employed deep learning methods. This model combines the Long-Short Term Memory (LSTM) and Transformer architectures to capture long-term dependencies effectively using the extracted MFCC features. The RAVDESS dataset was employed for experimentation, and the model exhibited a weighted accuracy (WA) of 75.33% and an unweighted accuracy (UA) of 73.12% over the RAVDESS dataset. Due to their proficiency in processing and making predictions on time series data, LSTM networks and their feedback connection capabilities are highly effective for tasks involving sequence recognition, such as speech and emotion recognition [19]. A specific data transmission approach becomes essential, especially when working with long data sequences and when the model is required to understand the relationships between past and future words. In response to this challenge, the bidirectional network was introduced, enabling the use of LSTM in a bidirectional fashion [20].

Bidirectional Long Short-Term Memory (Bi-LSTM) represents an evolution of traditional LSTMs. In Bi-LSTMs, each signal is trained in both forward and backward directions, effectively splitting the RNN into two separate components [21]. While LSTM trains a single model, Bi-LSTM introduces the concept of two models, enhancing the understanding of temporal dependencies in both directions [22]. However, little research has been done to learn the long-term dependencies in speech audio to classify emotions. It has been applied extensively because the LSTM design can represent long temporal sequences [23]. Longer-term dependencies, however, might not be sufficiently modelled [24]. Conversely, the use of positional encoding in the

Transformer layer is thought to result in the disadvantage of high computational costs [25], as each time step requires the Transformer to recompute its whole history. Still, the Transformer performs well when managing long-term dependencies.

Therefore, this paper offers a hybrid model combining CNN, LSTM, and KNN to improve SER performance by learning the long-term dependencies in audio speech files. To mitigate the challenge of low accuracy, we introduced a technique that leverages a blend of machine learning and deep learning algorithms, resulting in improved performance. The model is specifically trained using Mel spectrograms derived from audio data. We have preprocessed audio so that the abnormal condition could easily be removed before giving input to the model. The proposed model is evaluated on the TESS dataset and compared with previous research to inspect its performance.

3. PROPOSED METHODOLOGY

This paper suggests a hybrid SER model that takes advantage of networks' temporal and spatial characteristics by combining a CNN, LSTM, and KNN. End-to-end feature extraction is done via CNN. A convolutional layer and a pooling layer are features of a standard CNN model. Throughout the training phase, the convolutional layer controls the input features and different filters. Following the convolutional layer, the pooling layer eliminates redundant data from the feature map in order to gather concentrated activation features from the convolutional feature map. We decided to use the LSTM model to preserve more valuable features further. In essence, the LSTM network is a recurrent neural network that was created to solve the traditional RNN's vanishing gradient issue. It is ideal for learning from time-varying sequences because of its long-term memory retention. Once a new feature is extracted from the CNN-LSTM model, KNN uses the majority class or mean value of its k closest data points in the feature space to classify it. That gives KNN a resilient and intuitive feature that is well-suited for SER. Fig. 2 shows the overall architecture of the proposed model.

3.1. DATASET DESCRIPTION

We use the Toronto Emotional Speech Set (TESS) dataset, which comprises audio recordings of actresses uttering 200 target words within the carrier phrase. Two actresses, aged 26 and 64, performed these recordings, each portraying seven distinct emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutrality. These target words are the words that the actresses were asked to pronounce in their recordings. Each target word is used as part of the carrier phrase "Say the word _." For example, if one of the target words is "happy," the actress would say, "Say the word happy," and this utterance would be recorded.

Now, for each of these 200 target words, the actresses recorded their speech while expressing each of the seven different emotions. That means seven different recordings for each of the 200 target words, each representing another emotion. These seven emotions are anger, disgust, fear, happiness, pleasant surprise, sadness, and neutrality. This process applies to all 200 target words and all seven emotions for both actresses. As a result, the dataset contains a total of 2,800 audio files (2 actresses * 200 target words * 7 emotions = 2,800 audio files), with each audio file representing a unique combination of actress, target word, and emotion. Each audio recording is stored as a separate audio file in WAV format. Fig. 1 shows the TESS dataset visualization.

Fig. 1- Dataset Visualization.

3.2. DATASET DESCRIPTION

Spectrogram and segmentation are essential components of SER. Spectrograms provide a time-frequency representation of audio data, and segmentation allows us to divide speech into smaller, manageable units for analysis. Before the feature extraction procedure, speeches are loaded and preprocessed. To avoid the speech beginning with a pause, the first 0.5 seconds are removed. The speech is then extracted for three seconds in order to extract features. Shorter signals are padded with zeros, whereas longer ones are terminated after three seconds.

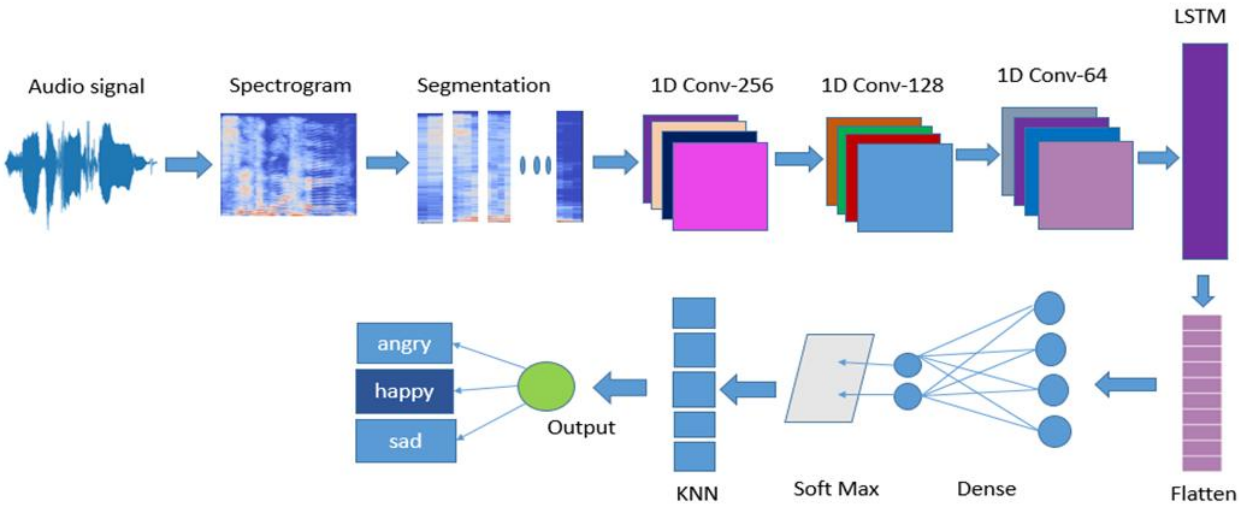

Fig. 2- Proposed model of speech emotion recognition.

3.3. FEATURE EXTRACTION

Here, the MFCC feature is applied to recognize the emotion. MFCCs are helpful for simulating emotional content in audio because they are proficient at capturing the spectral properties of speech. A complete diagram for MFCC extraction from the audio signal is depicted in Fig. 3. Pre-emphasis is the beginning phase of MFCC feature extraction, which is done in order to amplify the speech signal's high-frequency components. The speech signal is divided into small, overlapping frames in the frame-blocking phase.

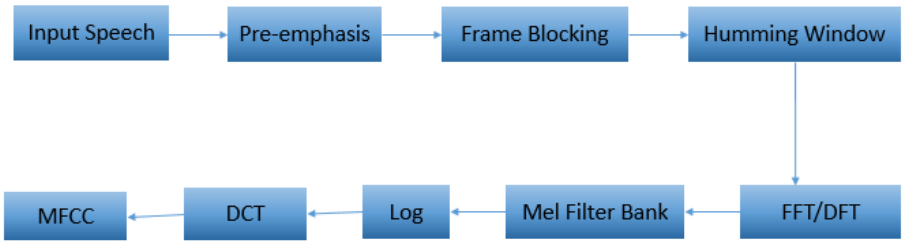

Fig. 3- Framing blocking MFCC feature extracting.

The time duration of each frame is typically 20–30 milliseconds and is often overlapped by 50% with the previous frame. Then, each frame is multiplied by a window function such as Hamming to diminish spectral leakage at the edges of the frame. After windowing, a Fast Fourier Transform (FFT) is applied to each windowed frame to convert the signal from the time domain to the frequency domain. The power spectrum obtained from the FFT is then passed through a series of triangular filters. These filters are spaced on a Mel frequency scale, which is a perceptually motivated scale that models how humans hear the pitch. The filter bank coefficients represent the energy in different frequency bands, emphasizing lower frequencies, as they are more relevant for speech. In the log phase, it mimics the human auditory system's response to sound, which is approximately logarithmic in nature. The obtained log-filter bank energies are subjected to a discrete cosine transform. The filter bank coefficients are de-correlated in this phase, which lowers their dimensionality and improves their analytical suitability. The DCT coefficients obtained in the preceding phase are known as cepstral coefficients. These coefficients are the final MFCC.

3.4. CLASSIFICATION

CNNs were primarily used for pattern recognition, image processing, and video processing. However, a customized 1D-CNN model can also be employed for SER, as shown in the proposed model in Figure 1. The model starts with two Convolutional layers (Conv1D), each equipped with 256 filters, a kernel size of 5, and 'same' padding to preserve the input size. Following each convolution, a ReLU activation function is applied, allowing the model to learn hierarchical features from the input sequence. After the second Convolutional layer, a Batch Normalization layer is introduced to stabilize and expedite the training process. A dropout layer with a rate of 0.25 is incorporated to counter overfitting. Dropout is a regularization technique that guards against overfitting by randomly deactivating a fraction of neurons during training. The model is further enriched with a Max Pooling layer, featuring a pool size of two, which downsamples the spatial dimensions of the data. Subsequently, the model consists of four Conv1D layers, each with 128 filters, a kernel size of 5, and 'same' padding to maintain the input dimensions. Following each convolutional layer, a ReLU activation function is applied to introduce non-linearity. After these four convolutional layers, a BatchNormalization layer is added to normalize

activations and enhance training stability. A dropout layer with a rate of 0.25 is included to prevent overfitting. Finally, a MaxPooling1D layer with a pool size of 8 is used to downsample the feature maps.

In addition, two more Conv1D layers with 64 filters and ReLU activation are included to extract further features from the data. A Long Short-Term Memory (LSTM) layer with 128 units was introduced. The LSTM output is subsequently flattened to a 1D vector, preparing it for the final classification layer. A Dense layer is then introduced, featuring seven units and a softmax activation function. This layer is responsible for carrying out the classification task across seven distinct output classes. The softmax activation function converts the model's output into class probabilities. The model now needs to be compiled. The optimizer argument specifies the optimization algorithm to be used during training. In this case, the Adam optimizer is employed. The Adam optimizer is renowned for its adaptive learning rates and is widely favoured for training deep neural networks. The loss argument specifies the loss function that the model will optimize throughout training, and 'categorical_crossentropy' is selected as the loss function. The metrics argument is a list of evaluation metrics to be monitored during training. In this instance, 'accuracy' is used as the metric, which is a commonly employed metric for classification tasks. Now, the model is ready for training. A batch size of 64 is utilized, and the model will undergo 30 training epochs. Additionally, 20% of the training data is reserved for validation. Following the training process, the test and training accuracy of the model are printed. Subsequently, features are once again extracted from the trained CNN-LSTM model. Each row in the array corresponds to a different sample or input, and each column represents a specific feature. These features are the high-level representations of the input speech data the model has learned to extract during training. While the feature values don't directly convey information about specific emotions, they represent high-level characteristics of the input speech data the model has learned to extract. The CNN-LSTM model has been trained to automatically extract relevant features that are indicative of different emotions from the raw speech data. These features are not directly interpretable by humans but are crucial for the model's ability to make emotional predictions. To understand the output in terms of emotions, interpret the emotion category corresponding to the highest probability in the final classification. For example, if the emotion category with the highest probability is "happy," the model predicts that the input speech is associated with the emotion of happiness.

KNN represents a straightforward yet efficient algorithm applied for both classification and regression assignments. It derives predictions by considering the majority class or mean value of its k closest data points within the feature space, endowing it with resilience and an intuitive nature suited to various applications. KNN's capacity to adapt to various data distributions, straightforward implementation, and minimal training requirements render it invaluable for endeavors, including recommendation systems, anomaly detection, and pattern recognition.

However, this dataset will serve as our training dataset for the KNN classifier. Creating a KNN classifier with n_neighbors=5 means we want to consider the five nearest neighbors when making predictions. To classify a new feature extracted from the CNN-LSTM model, the KNN classifier performs the following steps:

Euclidean Distance Calculation: The KNN algorithm calculates the Euclidean distance (or other specified distance metric) between the new feature and all the features in the training dataset. The Euclidean distance measures the similarity between two feature vectors in the feature space.

Selecting the Nearest Neighbors: It identifies the five features from the training dataset with the shortest Euclidean distances to the new feature. These are the five nearest neighbours to the new feature.

Majority Voting: To predict the new feature, the KNN algorithm looks at the class labels of the five nearest neighbours. It counts the number of neighbours in each class and selects the class that appears most frequently among the neighbours.

Assigning a Class: The KNN classifier assigns the class label with the highest count (the majority class) as the predicted class for the new feature.

Prediction Result: The KNN classifier returns the predicted class label for the new feature.

To assess the KNN model's performance, we utilize the accuracy score function to calculate the accuracy of both the training and testing data.

4. RESULT AND DISCUSSION

This section presents the results and discusses the study's findings on SER. The primary objectives of this research were to explore the effectiveness of combining CNN, LSTM, and KNN in the context of SER. This section explores into the experiments' outcomes and the chosen techniques' implications.

4.1. WAVEFORM AND SPECTROGRAMS OF EMOTION SPEECH

Waveform analysis in SER involves the examination of the raw audio signal's amplitude over time. The waveform encapsulates fundamental speech characteristics, including pitch, intensity, and prosody, which are crucial for emotion detection. By analyzing waveform features, such as pitch contours and speech rate, it becomes possible to uncover distinctive patterns associated with different emotional states. The waveforms for seven emotions are shown in Fig. 4. The variations in their waveforms indicate that emotions can be effectively conveyed through speech, as different emotions are associated with distinguishable waveform patterns.

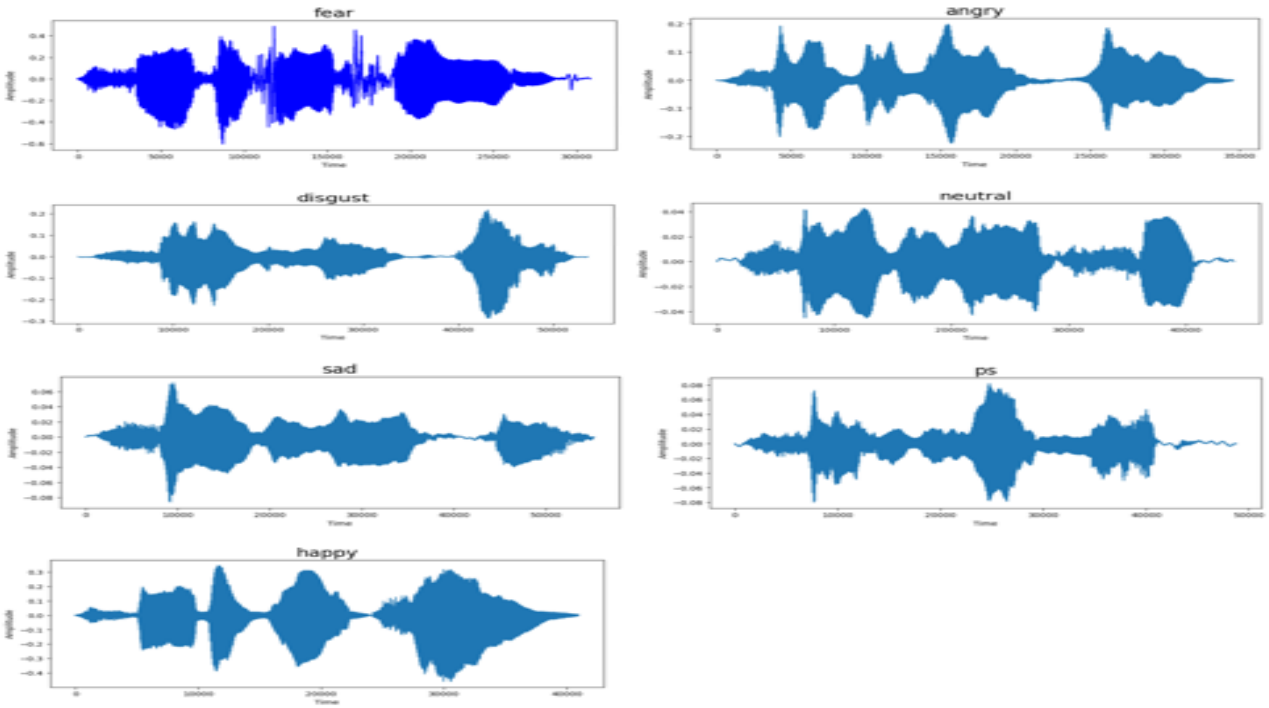

Fig. 4-Visual representations of waveforms of seven different emotions

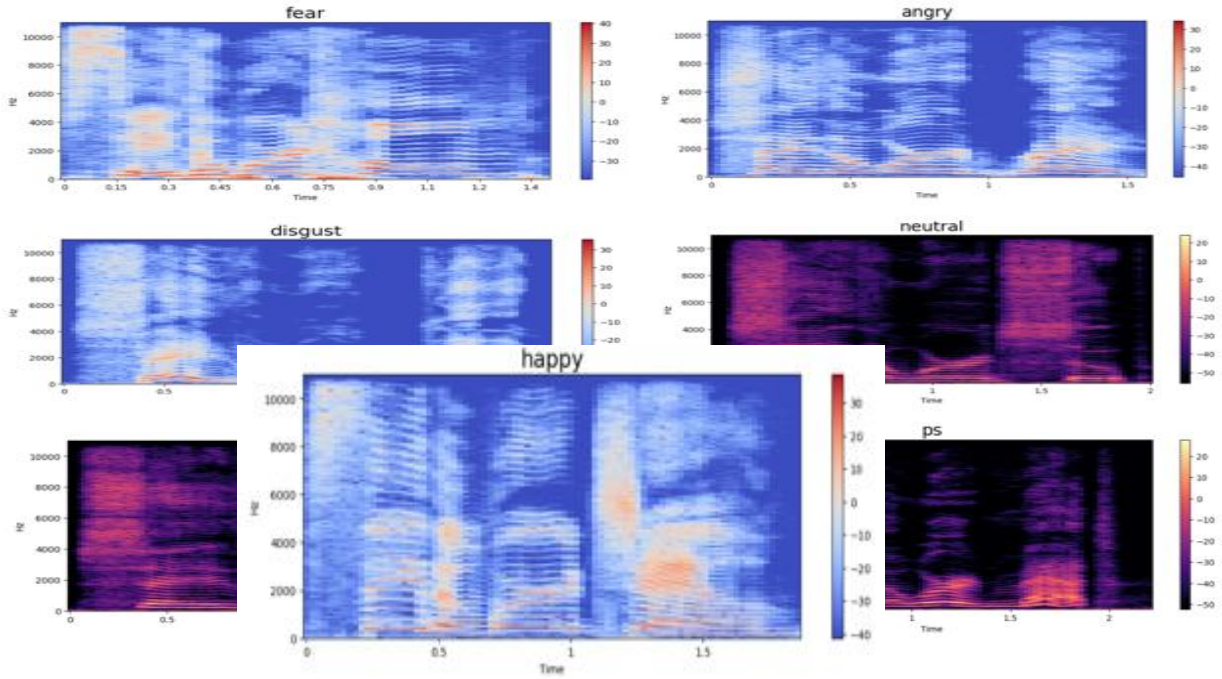

Fig. 5- Visual representations of speech signals in 2D spectrograms of 7 different emotions

Whereas a spectrogram is a visual representation of the time-varying frequencies of a signal, often used to analyze audio data. It provides a 2D representation where time is plotted on the x-axis, frequency on the y-axis, and the colour intensity represents the magnitude of the frequencies at different points in time. Fig. 5 shows the spectrogram of seven emotions, and it can be seen that there is a difference for the seven emotions. Distinct spectrogram patterns for each emotion serve as unique acoustic signatures.

4.2. MODEL PERFORMANCE EVALUATION

By implementing a CNN with LSTM and applying specific optimizations, we achieved training and testing accuracies of 97.92% and 96.79% on the TESS dataset. Fig. 6 illustrates the accuracy curve for the CNN_LSTM model.

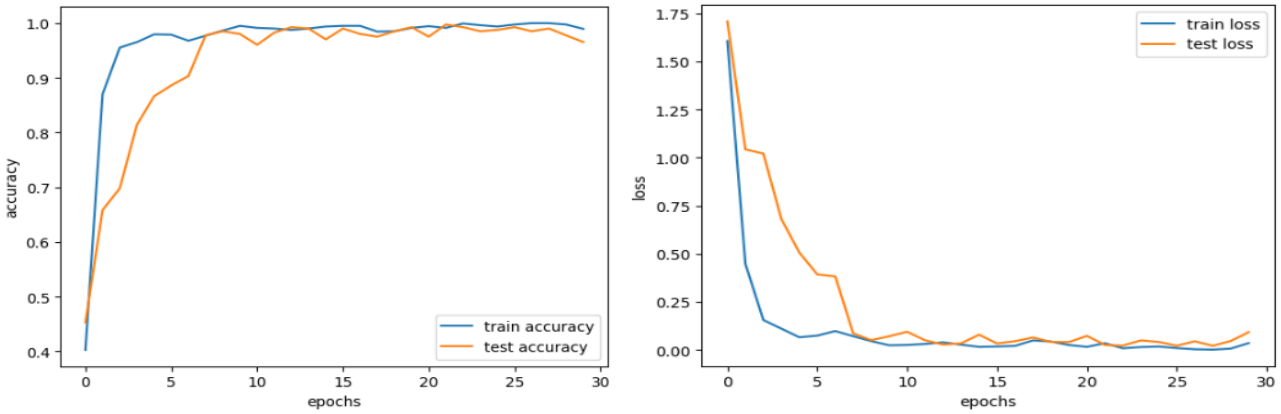

Fig. 6-Training verification model output accuracy.

Furthermore, after extracting features using KNN, we were able to attain a training accuracy of 99.60% and a testing accuracy of 98.21%. These results are presented in Figure 6, which displays the accuracy of the CNN_LSTM_KNN model. The experimental results for each emotion are presented in the form of a confusion matrix. A confusion matrix also referred to as an error matrix, is a square matrix used to summarize the performance of a machine learning or deep learning model in a multiclass classification task. In this matrix, each row corresponds to the actual (ground truth) class labels, while each column represents the predicted class labels generated by the model. The values within the matrix indicate the count of samples falling into each potential category based on the model's predictions. The confusion matrix serves as a valuable tool for evaluating the model's proficiency in accurately identifying emotions from speech samples. It plays a crucial role in helping researchers and practitioners gain insights into the model's performance in the classification of these emotions. The confusion matrix is divided into four key components:

True Positives (TP): These are samples where both the actual and predicted emotions match correctly. For example, if a speech sample is genuinely happy, and the model predicts it as happy, it's a true positive for the "happy" class.

True Negatives (TN): These represent samples that are correctly classified as not belonging to the class in question. For other emotion classes, they are neither falsely classified as the class of interest nor truly belonging to it.

False Positives (FP): When the model predicts a sample to be the class of interest, it is not. In SER, this means the model incorrectly identifies an emotion when it is not present.

False Negatives (FN): These are samples where the model fails to predict the class of interest when it is genuinely present in the data. In SER, this means the model misses or misclassifies an emotion in a speech sample.

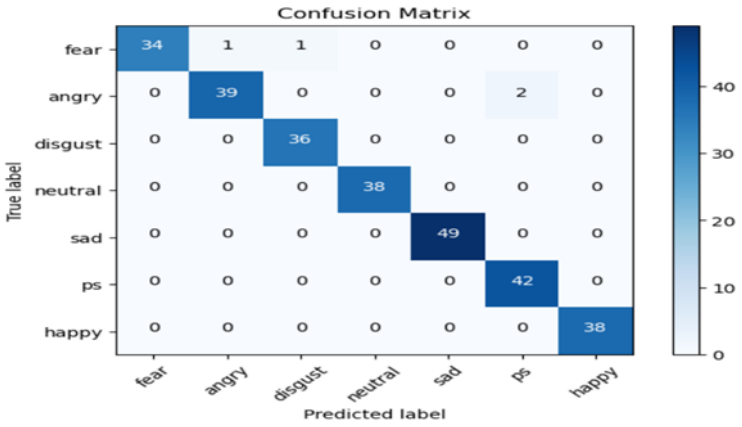

Fig. 7- Confusion Matrix of the Model.

In Fig. 7, the confusion matrix depicts the classification of seven distinct emotions. The actual and predicted emotions are displayed along the vertical and horizontal axes, respectively. The matrix analysis reveals that the model exhibited accurate predictions for disgust, neutrality, sadness, and happiness speech emotions compared to the emotions of fear, anger, and PS. Overall, the model's performance is outstanding, with only minimal misclassifications of speech emotions, which can be considered negligible. Classification matrices, Accuracy, Recall, and F1-Score of the model

obtained from the confusion matrix are mentioned in Table 1. Here, Precision measures how many predicted positive samples were positive. In SER, this indicates the ability of the model to identify a specific emotion correctly. Recall measures how many of the actual positive samples were correctly predicted. It indicates the model's ability to capture all instances of a particular emotion. The F1-score is the harmonic mean of Precision and recall, providing a balance between the two. It's a valuable metric when you want to consider both Precision and recall together.

Table 1- Performance Matrices

Class of Emotion	Precision	Recall	F1-Score
Fear	100.00%	97.56%	98.77%
Angry	100%	95.12%	96.30%
Disgust	100%	100.00%	98.63%
Neutral	100.00%	100.00%	100.00%
Sad	100.00%	100.00%	100.00%
PS	96.45%	100.00%	97.67%
Happy	100.00%	100.00%	100.00%

4.3. COMPARATIVE ANALYSIS

From Figure 8, proposed model is clearly the best performer in the comparison table. The model achieves enhanced accuracy and robustness in recognizing a wide range of emotions in speech signals. Fig. 8 shows the comparative analysis of the developed trained models. Passricha, Vishal, Aggarwal, and Rajesh Kumar [21] proposed a method for emotion recognition in human speech using a CNN with two layers of Bi-LSTM. They achieved a testing accuracy of 86.43%. In a separate study, SemeSarker, Khadija Akter, and Nursadul Mamun [13] introduced a model that employed 2D CNN architecture. Their model achieved 88.93% accuracy using LMS features and an impressive 93% accuracy using a combination of MFCC and LMS features.

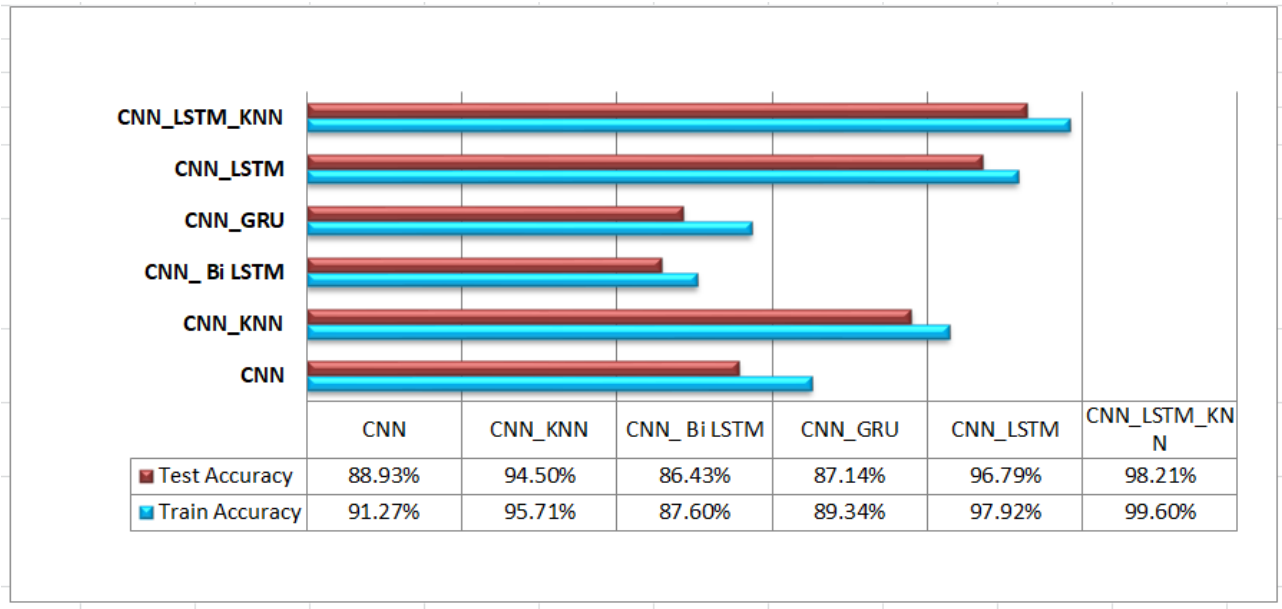

Fig. 8- Train and test accuracy comparison among six models

D. Issa, M. F. Demirci, and A. Yazici [10] introduced a novel approach to emotion recognition in human speech employing a 1D CNN, which yielded an accuracy of 87.14%. The ensemble model, CNN_LSTM_KNN, achieved the highest test accuracy (98.21%) among the models evaluated, indicating its effectiveness in recognizing speech emotions. The combination of CNN, LSTM, and KNN allowed for utilizing spectral and temporal information, improving accuracy. The model demonstrated robustness in recognizing emotions across diverse speech samples, showing adaptability to different contexts.

5. CONCLUSION

SER, a critical component of affective computing, has seen significant advancements with the utilization of a hybrid model combining CNNs, LSTM networks, and KNN. This hybrid approach capitalizes on the strengths of each component to achieve highly accurate emotion recognition in speech

data. The CNNs effectively extract discriminative features from spectrograms, providing valuable representations of the underlying emotional content. LSTM networks, with their capacity to model temporal dependencies, enhance the contextual understanding of emotions to improve recognition accuracy. Finally, KNN serves as a robust classifier that efficiently assigns emotional labels to the extracted features. This combination of models has shown great promise, surpassing the capabilities of individual techniques in capturing both spectral and sequential patterns, enabling fine-grained SER. In future research, integrating multiple data sources, such as text, audio, and visual cues, creates a multimodal recognition system capable of providing a more comprehensive and accurate understanding of emotional states. In practical applications, this may result in more accurate emotion identification. We are adapting the hybrid model for real-time applications, particularly in fields like virtual assistants, call centres, and mental health monitoring, where timely and accurate emotion assessment is critical. We apply the hybrid model to human-AI interaction scenarios, where machines can adapt their responses and behaviours based on detected emotional states.

References

1. Speech Emotion Recognition Using Deep Learning Hybrid Models; JamsherBhanbhro, Shahnawaz Talpur, Asif Aziz Memon2022 International Conference on Emerging Technologies in Electronics, Computing and Communication (ICETECC) Year: 2022 | Conference Paper | Publisher: IEEE
2. Design of Speech Emotion Recognition Algorithm Based on Deep Learning; Xie Ying, Zhang Yizhe,2021 IEEE 4th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)
3. Emotion recognition from speech using MFCC and DWT for security system; Sonali T. Saste, S. M. Jagdale2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)
4. Speech Emotion Recognition Using ANN on MFCC Features; Harshit Dolka. Arul Xavier V M, Sujitha Juliet2021 3rd International Conference on Signal Processing and Communication (ICPSC) Year: 2021 | Conference Paper | Publisher: IEEE
5. M. S. Islam, Y. Zhu, M. I. Hossain, R. Ullah, Z. Ye, "Supervised single channel dual domains speech enhancement using sparse non-negative matrix factorization," Digital Signal Processing, 100, 2020 102697.
6. Y. Zeng, H. Mao, D. Peng, and Z. Yi, "Spectrogram based multi-task audio classification," Multimedia Tools and Applications, vol. 78, no. 3, pp. 3705-3722, February, 2019, doi: 10.1007/s11042-017-5539-3.
7. M. I. Hossain, T. H. A. Mahmud, M. S. Islam, M. B. Hossen, R. Khan and Z. Ye, "Dual transform based joint learning single channel speech separation using generative joint dictionary learning," Multimedia tools and applications, 2022. https://doi.org/10.1007/s11042-022-12816-0.
8. M. I. Hossain, M. S. Islam, M. T. Khatun, R. Ullah, and A. Masood and Z. Ye, "Dual Transform Source Separation Using Sparse Non-Negative Matrix Factorization," Circuits, Systems, and Signal Processing," 40, 1868–1891, 2021
9. R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.
10. D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," Biomedical Signal Processing and Control, vol. 59, p. 101894, 2020, doi: 10.1016/j.bspc.2020.101894.
11. K. Ashok Kumar, J. L. MazherIqba, Machine learning technique-based emotion classification using speech signals, Soft Computing (2023) 27:8331–8343 https://doi.org/10.1007/s00500-023-08185-x
12. Recognition of Emotion in Speech-related Audio Files with LSTM-TransformerFelicia Andayani, Lau Bee Theng, Mark TeeKitTsun, Caslon Chua2022 5th International Conference on Computing and Informatics (ICCI) Year: 2022 | Conference Paper | Publisher: IEEE
13. A Text Independent Speech Emotion Recognition Based on Convolutional Neural Network; SemeSarker, Khadija Akter,Nursadul Mamun2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)
14. Sentiment Analysis of Human Speech using Deep Learning; Manan Savla;DhruviGopani,MayuriGhuge,Sheetal Chaudhari, Pooja Raundale2023 3rd International Conference on Intelligent Technologies (CONIT)
15. S. Lee, D. K. Han, and H. Ko, "Fusion-ConvBERT: Parallel Convolution and BERT Fusion for Speech Emotion Recognition," in Sensors, vol. 20, no. 22, pp. 6688, 2020, doi: 10.3390/s20226688.
16. Research on Speech Emotion Recognition Based on Deep Neural Network Heng Li*, XueZhang,Ming-JiangWang 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP) Year: 2021 | Conference Paper | Publisher: IEEE
17. Emotion recognition from Persian speech with 1D Convolution neural networkSeyedMiladRanaeiSiadat;Ilia M. Voronkov;Alexander A. Kharlamov2022 Fourth International Conference Neurotechnologies and Neurointerfaces (CNN) Year: 2022 | Conference Paper | Publisher: IEEE

18. Recognition of Emotion in Speech-related Audio Files with LSTM-TransformerFelicia Andayani, Lau Bee Theng, Mark TeeKitTsun, Caslon Chua2022 5th International Conference on Computing and Informatics (ICCI) Year: 2022 | Conference Paper | Publisher: IEEE

19. D. Jing, T. Manting, and Z. Li, "Transformer-like model with linear attention for speech emotion recognition," in Journal of Southeast University (English Ed.), vol. 37, no. 2, pp. 164-170, 2021, doi: 10.3969/j.issn.1003-7985.2021.02.005.

20. Ryan T. J. J., LSTMs Explained: A Complete, Technically Accurate, Conceptual Guide with Keras, Published in Analytics Vidhya, Sep 2, 2020.

21. Passricha, Vishal and Aggarwal, Rajesh Kumar, "A Hybrid of Deep CNN and Bidirectional LSTM for Automatic Speech Recognition," Journal of Intelligent Systems, vol. 29, no. 1, 2020, pp. 1261-1274.

22. Bengali Speech Emotion Recognition: A hybrid approach using B-LSTMMd Jahid Hasan;MohammadSakibHossain;S.M. Nazmul Hassan; Mohammad Al-Amin;Md. NakibRahaman;MashukArefinPranjol2022 4th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)

23. O. V. Verkholyak, H. Kaya, and A. A. Karpov, "Modeling Short-Term and Long-Term Dependencies of The Speech Signal for Paralinguistic Emotion Classification," SPIIRAS Proceedings, 2019, pp. 30-56, doi: 10.15622/sp.18.1.30-56.

24. Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context," Proc. of the 57th ACL, Florence, Italy, 2019, pp. 2978–2988. Available: https://aclanthology.org/P19-1285.pdf, [Online].

25. J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visualrecognition and description,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 677–691, 2017.