SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I7P104

Volume 12 Issue 7, 35-46, July 2025
© 2025 Seventh Sense Research Group®

*Original Article*

# Enhancing Speech Emotion Recognition with Multi-Modal Hybrid Features and CNN

Rashmi Rani[1], Manoj Kumar Ramaiya[2]

[1,2]*Department of CSE, Sage University, Indore, Madhya Pradesh, India.*

[1]*Corresponding Author : rashmi.sagecs@gmail.com*

Received: 02 May 2025 Revised: 04 June 2025 Accepted: 03 July 2025 Published: 31 July 2025

***Abstract*** *- Speech emotion recognition is pivotal in human-computer interaction, enabling machines to understand and respond to human emotions. While traditional methods often rely on low-level feature extraction, this paper proposes a language-independent, deep learning-based framework for accurate speech emotion classification. By combining the strengths of hybrid feature extraction techniques and 3D Convolution Neural Networks (CNN), the proposed framework effectively captures both local and global information from speech signals. The model is evaluated on RAVDESS, CREMA-D, SAVEE, and TESS datasets, achieving impressive accuracy rates of 98.48%. In order to verify the proposed work, we have compared the model with the LSTM and Bi-LSTM models too. Experimental results demonstrate the superiority of the proposed framework over state-of-the-art methods, paving the way for more sophisticated and empathetic human-computer interactions.*

***Keywords*** *- Emotion recognition , Speech , Convolution Neural Networks , ZCR, RMSE, MFCC, LSTM, Bi-LSTM.*

## 1. Introduction

Emotion recognition in speech is vital for human-computer interaction, enabling more natural communication. Accurate emotion recognition allows automated systems to respond more effectively, enhancing user experiences. One of the most intuitive ways to communicate is speech, and despite several advancements in technology, truly natural interaction between humans and machines remains a challenge.

This is primarily because current machines lack the ability to comprehend human emotions and respond appropriately. However, the complexity and dynamic nature are challenging, too. The Emotion recognition in speech is difficult due to the subtle and diverse nature of human emotions, which are often conveyed through complex acoustic features like Pitch, tone, volume, and rhythm. Individual speech patterns and cultural differences further complicate accurate emotion detection.

Previous research on emotion recognition mainly focused on identifying acoustic features from emotion, and various mathematical models, like SVM, KNN, etc, were approached [1]. Traditional machine learning models often struggle to capture these nuances, and the effective features remain unclear, limiting their performance and reliability. These limitations can be addressed through multi-modal techniques, where hybrid deep learning approaches with diverse data sources enhance data richness and reduce uncertainties [2].

Speech Emotion Recognition (SER) serves as an efficient and widely used approach for facilitating communication between humans and computers, with numerous practical applications in the Human-Computer Interaction (HCI) discipline [3]. Researchers are currently grappling with the challenge of selecting an effective method for extracting meaningful and distinctive features from speech signals to capture a speaker's emotional state [4] accurately. Over the last decade, many studies have focused on analyzing low-level handcrafted features for Speech Emotion Recognition (SER).

This research focuses on multi-modal speech emotion recognition by employing a hybrid feature extraction approach combined with a Convolutional Neural Network (CNN). By integrating various feature extraction techniques, the model aims to capture a wide range of emotional cues from speech signals, thereby improving recognition accuracy. The proposed method takes advantage of both handcrafted and deep learning-based features, resulting in a more robust and generalized emotion detection system. Additionally, the CNN is used for efficient feature learning and classification, making detecting small differences in emotions easy [5]. Mel-scale filter bank speech spectrograms are commonly used as input features for CNN, as spectrograms provide a two-dimensional representation of speech signals, which are extensively applied in Convolutional Neural Networks (CNNs) to extract key and distinctive features for SER and various other signal processing tasks [6].

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

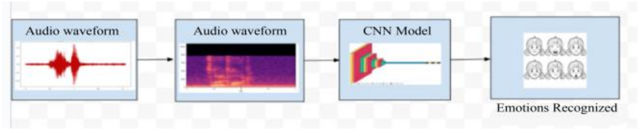


**Fig. 1 Proposed framework for emotion classification**

Despite several advancements in machine learning, there are n numbers of challenges to deal with. The major challenge with machine learning models is that their performance suffers while working with a multi-modal SER model, and they also face issues in understanding the accents of hybrid datasets. Another issue with existing models is that they struggle to differentiate background noise, vocal sounds and variations in speech, making them unreliable in dealing with real-life situations. To improve the accuracy of the SER model, it is very necessary to build a hybrid model that can efficiently work on multiple datasets simultaneously.

A hybrid multi-model is proposed to deal with all these drawbacks, providing a complete, comprehensive investigation by integrating spatial and temporal features. Here, 4 different datasets, Ravadess, Savee, Tess, and Crema-D-D, are taken as raw data, which is then stacked up to create a new dataset comprising data of variance fields, for feature extraction. Hybrid feature extraction is used by combining MFCC, ZCR, and RMSE, resulting in the best features out of the box. The proposed SER model is tested on various classifiers, such as LSTM, Bi-LSTM, and CNN, and it was found that CNN provides the best results for emotion detection. The present model achieved an accuracy of 98.48%, outperforming other methods such as Probabilistic Neural Networks (95.56%), Long Short Term Memory networks (48.1%), and Bi-LSTM( 77.64%). The rest of this manuscript is structured as follows: Chapter II highlights a review of existing SER techniques, Chapter III elaborates on the proposed SER framework, Chapter IV presents the results of the suggested approach, and Chapter V includes a detailed discussion of the experiments. At last, Section VI concludes the manuscript and outlines future research directions.

## 2. Literature Survey

Developing a reliable and adaptable speech emotion recognition model is the main challenge that affects the performance of the system. Goncalves et al. Presented the research paper where the emotions are classified on the MSP-Podcast datasets. The Author classified categorial emotions and attribute-based regression of audio variance, dominance, in which 24.7% Macro F1-score accuracy and CCC scores were obtained as 0.7465, 0.6467, and 0.6712 for arousal, valence, and dominance. Despite these advancements, paper has some difficulties, including managing imbalanced classes, a lack of using multi-model configuration, and making the system less robust. There is a need for multimodal fusion for the improvement of the generalizability of SER systems in real-world scenarios [1].

Multimodal systems have gained vast popularity these days. Costa et. al proposed an integrated system combining text and speech data features for emotion detection. The Author has used a variety of deep learning models like Wav2vec2.0, HuBERT and BERT to fuse audio features and text features. The model achieved a 34.14% macro F1 score accuracy in emotion detection. Although this project performed well in audio and text fusion, more efforts are required to achieve better accuracy in managing noisy transcripts and learning tiny emotional differences [2].

Bhattacharya et. al (2024) proposed a Multilingual emotion detection model where validation is performed on RAVADESS, EmoDB and EmoVO datasets using MFCC and CNN. The system achieved accuracy (97.89%) for multilingual emotion detection and for RAVADESS and Emo-Db, 96.3%, 96.22% respectively. Although the accuracy achieved is good, the model needs to work with a diversity of datasets. Also, the input given is cleaned, so the Author should implement it on real-life datasets [15].

Byun et. al (2024) proposed the SER model and detected emotions from Korean datasets. Emotions are classified using RNN, with the involvement of CNN datasets, and attributes are identified. LSTM, Bi-LSTM and MFCC are approached for the model training and evaluation. Although this project achieved short intervals emotion recognition accuracy as 83.81% and speech emotion using one long LSTM model 75.51% but main challenge with this model is that it is only tested on only Korean datasets making the model language dependent creating unknown performance for other languages, also alternative fusion based feature extraction is suggested in the the future [18].

I. I. Rizhinashvili et al. presented a paper on SER where labeled emotions are converted into continuous values relying on the Valence-Arousal-Dominance (VAD) model, which is later averaged to find VAD scores. RAVADESS datasets are experimented with using a combination of Wave2Vec 2.0 and LSTM, consisting of 256 neurons in each layer. The model achieved 93.98 %. Despite the project receiving good accuracy, future involvement may include multi-modal emotion recognition [22].

Deshmukh et al. demonstrated classification of emotions from a speech data set. The Author investigated EMO-DB, RAVADESS datasets and concluded happiness, sadness, anger, disgust, fear, and melancholy emotions. Probabilistic Neural Network (PNN) classifies the emotions and achieves 95.75% accuracy for the EMO-DB dataset and 84.64% accuracy for the RAVADESS dataset [23]. Although the project performed well, it did not evaluate noise resistance or speaker variability, which are crucial for practical applications. Choudhary et al. developed a SER model that is based on deep learning based systems combining CNNs and LSTMs, including MFCC features. The Author evaluated the

CNN+LSTM model on the RAVDESS and TESS datasets, attaining 97.1% accuracy. The model has only performed its operation on 2 datasets, restricting its ability to work on the diversity of emotions. Also, an 11% drop in overall performance can be seen in the research on the RAVADESS dataset [24].

Bhangale et al. put forward an emotion detection model using 1 D DCNN. The Author utilized EMO-DB and RAVADESS datasets to recogonize long-term emotional patterns. A variety of emotional features are extracted by applying LPCC, MFCC, wavelet packet transform, RMS, Pitch, and Jitter, and they are sent to DCNN for emotion classification. The model obtained 93.13% and 94.18% accuracy for EMO-DB and RAVADESS datasets. The model produced favourable results but failed in working with multidisciplinary datasets, which limits its general real-life usage. Also, speaker assents and demographic differences are not evaluated [25].

Feature extraction plays an important role in emotion classification. For SER, capturing prosodic emotion is very important. Koti et al. classified emotion in the Berlin database of emotional speech using GMM and EML. The model received 73% accuracy, but there was a scalability issue. The use of EML makes the model less efficient when working with complex and large datasets [26].

**Table 1. Related work studies**

| S.no | Reference | Work Done | Techniques Used | Results |
|---|---|---|---|---|
| 1 | Gismelbari et. al (2024) [3] | Presented a Speech emotion recognition model that successfully provided a research analysis on the usage of CNN and HuBERT for emotion classification. | Convolutional Neural Networks (CNNs) - HuBERT model | The study gives importance to the usage of HuBERT, which outperformed CNNs, and increased accuracy and efficiency. |
| 2 | Lian, Z et. al (2024) [4] | Proposed Multimodal Emotion Recognition (MER) | NLP, /neural Networks, AI, MER | Presented a review paper where key challenges in MER are identified, such as complex environments and incorrect annotations. |
| 3 | Ismaiel et. al (2024) [5] | Outlined Speech emotion recognition in Arabic speech and identified emotions from Arabic speech data with 844 features. | - Deep learning (SERDNN model) - Traditional machine learning models (Xgboost, Adaboost, KNN, DT, SOM) | The Author used the SERDNN model, which achieved the maximum accuracy (97.40%) and loss (0.1457). |
| 4 | Geetha et. al (2024) [6] | Multimodal emotion recognition challenges and advancements are discussed. | CNN, RNN, MER, DL techniques | Recent and up-to-date comprehensive reviews on the advancement of deep learning technologies are discussed. |
| 5 | Vaidehi et. al (2024) [7] | Presented a Speech Emotion Recognition model and tested it on 4200 audio samples from the RAVADESS and TESS datasets. | CNN, Random Forest, SVM | CNN achieved the highest accuracy, especially for neutral emotion |
| 6 | Duret et. al (2024) [8] | Presented a Speech emotion recognition model, which details the 2024 edition of MSP- Podcast SER is discussed. | Self-Supervised Learning model, SVM, MFCC | The model achieved an F1-macro score of 0.35% |
| 7 | Jayakumar et. al (2024) [9] | Emotions are detected from the Facial expression embedding using the tensorflow framework in a deep learning model, providing spoken responses from the voice assistant. | Deep learning, TensorFlow, CNN | The model achieved 90 % accuracy in recognizing emotion from the spoken feedback. |
| 8 | Madhura et. al (2024) [10] | The Author presented a review paper on the classification of emotion from text, audio, and video datasets. | CNN, NLP, MFCC, SVM, KNN Classifier | The Author Highlighted the benefits of combining modalities (text, visual, audio) for emotion recognition. |

| | | | | |
|---|---|---|---|---|
| 9 | Diatlova et. al (2024) [11] | The Author investigated the WavLM large model for speech emotion detection and performed fine-tuning.<br>MSP Podcast Corpus datasets are experimented with multiple tasks including Sex, information from spoken utterances,<br>and different pooling methods are also employed. | Speech self-supervised learning (SSL), WavLM CNN, AdamW Optimizer | WavLM Large model is used for investigation of fine-tuning and achieved 0.35 and 0.02 F1 -macro score. |
| 10 | Tomar et. al (2024) [12] | Introduced fusion-based techniques for SER with Multi-modal emotion recognition | Feature-level fusion (CSTNF) with Speech Temporal NeuMF framework | Present a SER model demonstrating the effectiveness of feature-level fusion for improved performance with 97 % accuracy in emotion classification. |
| 11 | Haldorai et. al (2024) [13] | With audio and visual data integration, the Author proposed a bi-model Emotion Recognition system. Feature-level fusion architecture gives promising results. | Bi-LSTM, Parallel convolution model (Pconv) | The Author presented a study that uses both Audio and Visual inputs and successfully recognized emotions using AI, but the standard benchmark on a real-world dataset was not disclosed. |
| 12 | Kapileswar et. al (2024) [14] | The Author proposed a Speech emotion recognition model in which they introduced an Artificial Intelligence Assisted Learning Scheme with fusion-based LSTM and CNN. | CNN, LSTM, ALIAS, MFCC | The proposed model achieved 98% accuracy on the CREMA-D dataset. |
| 13 | Härm et. al (2024) [16] | proposed a Wav2Vec2-BERT-based Speech Recognition model and LLaMA2-7 B-based text recognition model. Multi-class<br>regression model is proposed by combining both models on 68,119 speaking turns.<br>The model successfully recognized emotions from the text and voice data. | Wav2Vec2-BERT for audio ,LLaMA2-7B for text, LLM, | With the involvement of fusion strategy, the paper shows promising results with an average F1 score of 0.354 for speech emotion recognition, but inconsistency can be seen in the emotion attributes, which may create a challenge in capturing subtle emotional differences. |
| 14 | Khan et. al (2024) [19] | Proposed a hybrid Speech emotion recognition model using CNN and Bi-LSTM. MFCC, ZCR, and Chroma are employed for the feature extraction, which are further fed to CNN and Bi-LSTM to recognize emotions<br>from both vocal and<br>spoken words. | CNN, Bi-LSTM. MFCC, ZCR, Chroma | The proposed model achieved 0.95 accuracy on the RAVDESS dataset, but the model is only tested on RAVDESS.<br>The Author should also test the model on multiple datasets to achieve robustness in real-world challenges. |
| 15 | Mahmoudi et. al (2023) [21] | Proposed Speech emotion recognition in<br>Arabic dialect. The Author classified and<br>identified emotions from Arabic datasets. | LSTM, GRU, RNN, CNN, SVM models | The proposed model attempted to improve accuracy for<br>Arabic speech emotion recognition with 77.14%+ accuracy<br>and<br>compared the<br>DNN accuracy with SVM. |

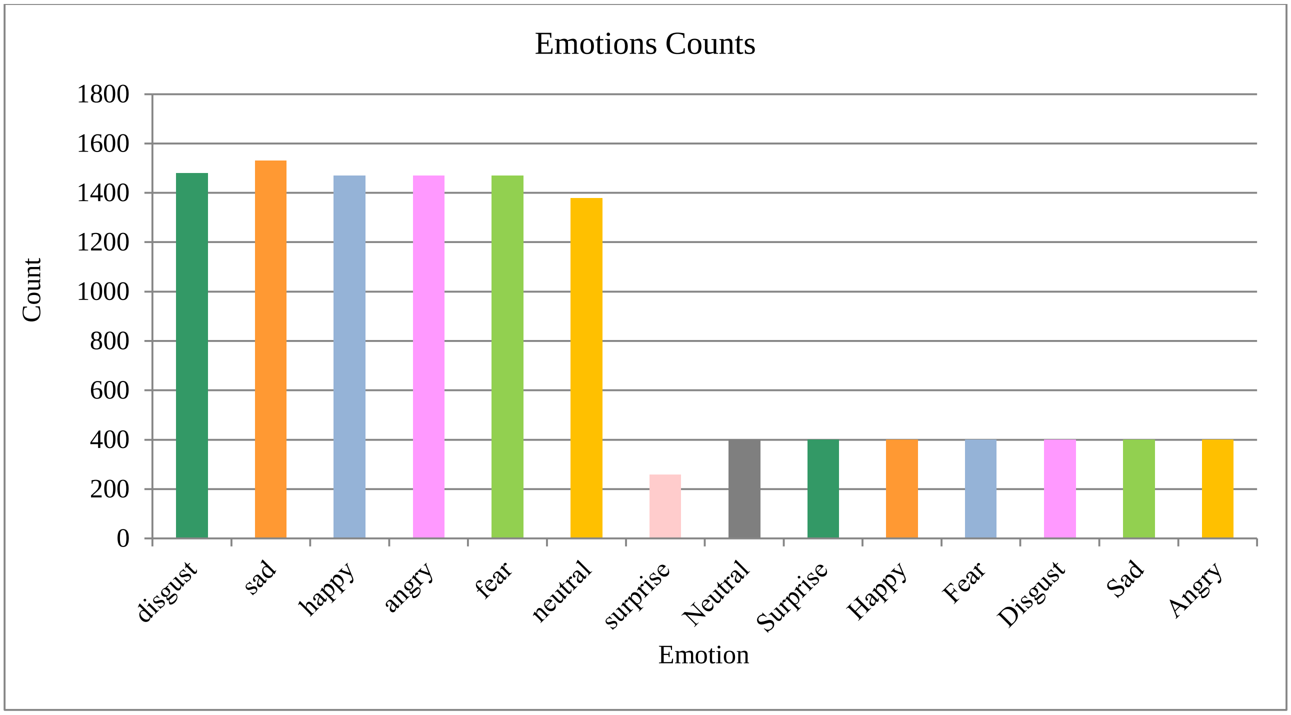


**Fig. 2 Multi-dataset bar-chart emotions count**

## 3. Proposed Methodology

This chapter provides a comprehensive explanation of the specific methods used for implementation. This project utilizes four datasets: RAVDESS, CREMA-D, TESS, and SAVEE. Upon importing these datasets, all the datasets are integrated to develop a robust multi-dataset model. The primary objective of this research is to enhance the system's capability to process diverse datasets concurrently, enabling the model to effectively learn from multiple data sources simultaneously.

Figure 2 is the bar chart that displays how many times each emotion appears in the dataset. These audio signals are further converted into a spectrum so that various deep learning models, which work efficiently with images, can process speech data for analysis and classification. After data analysis, the final processed data is prepared for data augmentation. Data augmentation expands the dataset's size and enhances the system's robustness, ensuring its effectiveness in various critical scenarios [7].

Following data augmentation, feature extraction is performed where multiple feature extraction techniques are integrated in a pipeline to ensure the selection of high-quality and relevant features. The hybrid feature extraction approach is applied by combining several techniques to enhance the model's performance.

Specifically, MFCC (Mel-Frequency Cepstral Coefficients) [8], RMSE (Root Mean Square Energy) [9], and ZCR (Zero-Crossing Rate) [10] are utilized to extract meaningful features, enabling the model to improve emotion classification accuracy.

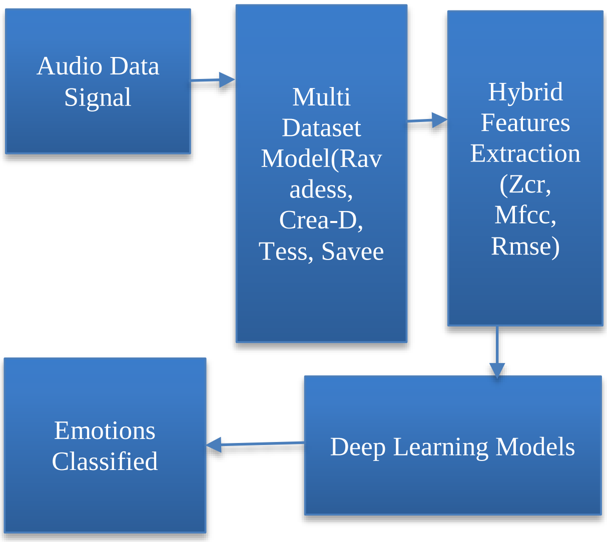

Audio Data Signal
Multi Dataset Model(Ravadess, Crea-D, Tess, Savee
Hybrid Features Extraction (Zcr, Mfcc, Rmse)
Deep Learning Models
Emotions Classified

**Fig. 3 Overall architecture of the proposed system**

The extracted features are then fed into a Convolution Neural Network (CNN), Bi-Directional Long Short Term Memory (Bi-LSTM), and Long Short Term Memory(LSTM) separately for the classification. Specifically, utilized 3D CNN is utilised for emotion recognition, allowing the model to learn spatial and temporal patterns within the features effectively.

The datasets are typically split into training and testing sets. A common split is 70% for training and 30% for testing. This division ensures that the model is evaluated on unseen data, providing a reliable estimate of its performance.

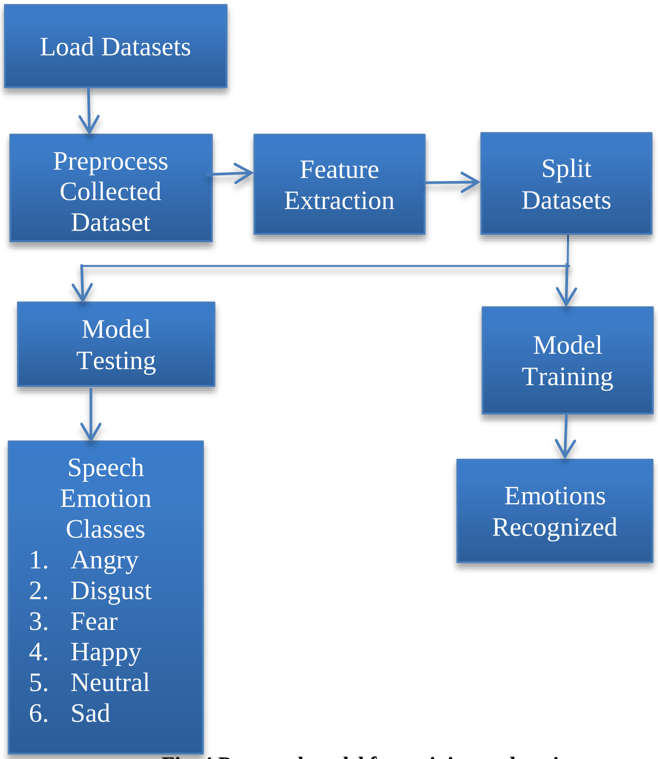


**Fig. 4 Proposed model for training and testing**

The detailed explanation of each component within the framework is provided in the following sections.

### *3.1. Data Augmentation*

Audio-based machine learning models are sensitive to variations in the input data [11]. To improve robustness, data augmentation techniques introduce controlled distortions that help the model generalize better [12]. This paper presents four augmentation methods—noise addition, Time shifting, Pitch shifting, and Time stretching—to enhance the variability of input signals and reduce model overfitting.

#### *3.1.1. Noise Addition*

Noise addition is an essential technique that simulates real-world conditions by introducing Gaussian noise into an audio signal. In this project, we added a noise of 0.035. The noise rate is randomly selected within a threshold of 0.075. The mathematical formulation of the noise augmentation can be expressed as:

$$\boldsymbol{X'} = \boldsymbol{X} + \boldsymbol{\eta}.\boldsymbol{N}(0,1) \qquad [13]$$

Where X is the original audio data , $\boldsymbol{\eta}$ = rate×max(X),N(0,1) represents a standard normal distribution.

#### *3.1.2. Pitch Shifting*

Pitch shifting is another augmentation method that shifts the audio temporally, altering the starting points of the speech signal [14]. During Pitch shifting, the input audio's Pitch is altered by a 0.7 factor using Librosa's pitch_shift function. When random=True, the pitch shift factor is selected randomly within the range [0,0.7].

#### *3.1.3. Time Shifting*

The shifting function shifts the audio signal along its temporal axis by rolling the array [15]. The shift amount is randomly selected within the range of [-5000,5000] samples, which is determined by multiplying a random integer between −5 and 5 by a rate of 1000.

#### *3.1.4. Time Stretching*

The stretching function modifies the temporal characteristics of the audio signal by applying Time stretching with a fixed rate of 0.8. This transformation alters the speed of the audio playback while maintaining the Pitch. Librosa's time_stretch function is used to achieve this effect [16].

### ***3.2. Feature Extraction***

In speech processing, extracting relevant acoustic features is essential for various applications such as speech recognition, speaker identification, and emotion detection. The Zero Crossing Rate (ZCR), Root Mean Square Energy (RMSE), and Mel-Frequency Cepstral Coefficients (MFCCs) are widely used features that capture different aspects of the speech signal [10].

The Zero Crossing Rate (ZCR) is computed to analyze the rate at which a signal changes its sign. It is particularly useful for distinguishing between voiced and unvoiced speech segments.

$$\text{ZCR} \quad =\frac{\mathbf{1}}{\boldsymbol{N-1}}\sum_{\boldsymbol{n=1}}^{\boldsymbol{n-1}}\mathbf{1}[(\boldsymbol{x_n n}.\boldsymbol{x_{n-1}}) < \mathbf{0}] \qquad [17]$$

Where: N is the total number of samples

$\boldsymbol{x_n}$ is a signal at sample n 1 is an indicator function that counts the zero-crossing rate

Another crucial feature is the Root Mean Square Energy (RMSE), which measures the short-term energy of the signal. RMSE is vital for analyzing speech intensity variations [11]. The proposed function for RMSE computation applies a frame-based analysis, allowing for detailed energy profiling of the speech waveform.

$$\text{RMSE} = \sqrt{1/\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\mathrm{y_i} \text{ - } y_i\hat{} \qquad [18]$$

The MFCC extraction process involves computing a set number of coefficients (e.g., 13) from the Mel spectrogram of the input signal. The implementation allows flattening the extracted features, which is beneficial for machine learning applications where a fixed-size input vector is required [19].

$$MFCC_k=\sum_{n-1}^{N} X_m(n)\ cos\ [\ k(n\text{-}1/2)\pi/2], k = 0,1,2,3\ldots\ldots, K-1$$

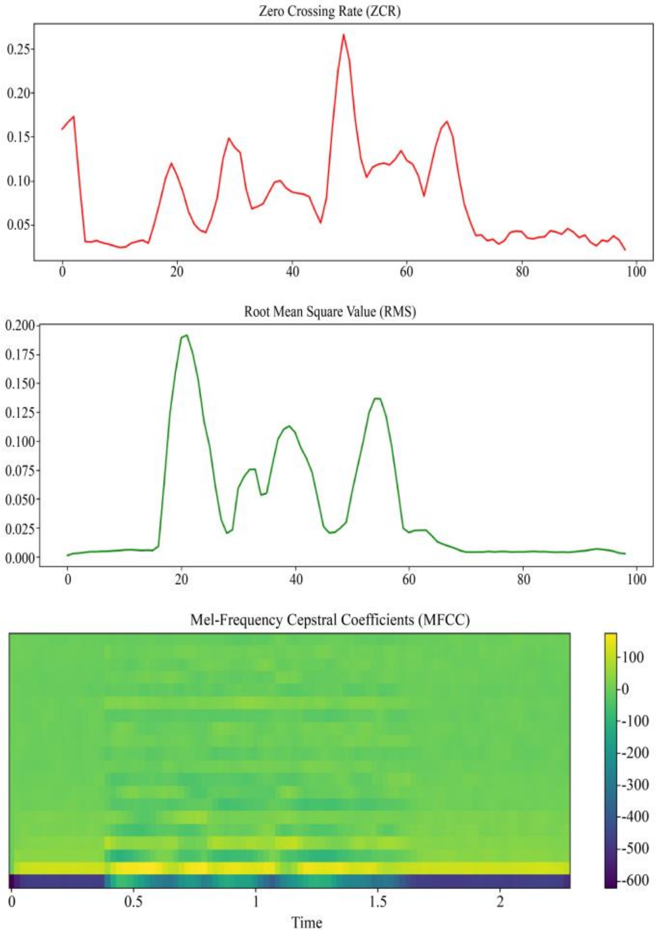


**Fig. 5 Parallel feature extraction network**

To integrate these features into a comprehensive analysis framework, a feature extraction function is designed to concatenate ZCR, RMSE, and MFCCs into a single feature vector, as presented in Figure 5.

This approach ensures that the extracted features capture the speech signal’s time-domain (ZCR and RMSE) and frequency-domain (MFCCs) characteristics. The proposed method maintains parameter consistency across all feature extraction functions using standardised frame and hop length values.

Additionally, the flexibility to toggle between flattened and non-flattened MFCC representations allows adaptability for different processing and classification tasks [20]. These features will be identified using the lybrosa library function of Python, and then the signal will be sampled 16,000 times, which will help to identify unique features for each emotion phase.

### *3.3. Model Architecture*

The experiments were conducted on CNN, LSTM, and LSTM. The best emotion recognition is done by the CNN. All the hybrid data features for the audio processing are converted into the spectrogram and given as input to CNN, LSTM, and Bi-LSTM. In this project, 3D CNN is employed where each layer includes convolution, normalization, pooling, dropout, and dense layers. At the first layer, 64 filters are applied to the convolution layer. The ReLU activation function is applied to preserve the spatial dimension. A 0.3 dropout will prevent the overfitting problem. In the second layer, the filter size is increased to 128, allowing the model to capture more complex spectrogram values, with the same batch normalization and dropout values. In the third convolution layer, the filter size is 256, so the model will work deeper in identifying patterns, decreasing computation count and making the model more efficient.

To evaluate the network performance, the dataset is divided into a 70:30 ratio, where 70% of the data is used for training purposes and 30% for testing. Cross-entropy is also used to calculate the validation error rate. To optimize the network and reduce the cost function, the Adam optimizer is applied with a fixed learning rate of 0.001, facilitating effective and stable training.

## 4. Results and Discussions

This section focuses on assessing the performance of the proposed system for Speech Emotion Recognition (SER) and comparing it against baseline approaches using standard benchmark datasets. To make our model more robust, 4 standard datasets (RAVADESS, SAVEE, TESS, CREMA-D) are integrated to form new hybrid datasets consisting of 4000 audio files, which are further sampled at a 16000 sampling rate. The frame length chosen is 2048, ensuring good balance between Pitch and formants.

The hop length taken for the audio signal is 512, which helps to prevent excessive overlap while maintaining temporal precision. Model performance was evaluated on LSTM, Bi-LSTM, and CNN. The comparison chart of all these models is shown below with the respective confusion matrix. Results show that CNN achieves 98% accuracy during training and 96% accuracy during testing, while LSTM gained 63.75 % for training and 48.43% accuracy in validation, whereas Bi-LSTM achieved 83% accuracy during training and validation accuracy achieved is 61%. Early stopping was applied in the model to deal with the excessive training problem, causing the model to stop working when achieving consistent results. CNN model stopped at 25 epochs, only showing outstanding performance, while LSTM and Bi-LSTM worked till 50 epochs.

Figures 6, 7, 8, 9, 10, 11 show LSTM, Bi-LSTM, CNN model accuracy and loss with respective confusion matrix, while Figure 12 summarize the performance of various models for emotion recognition based on Training Accuracy and Validation Accuracy.

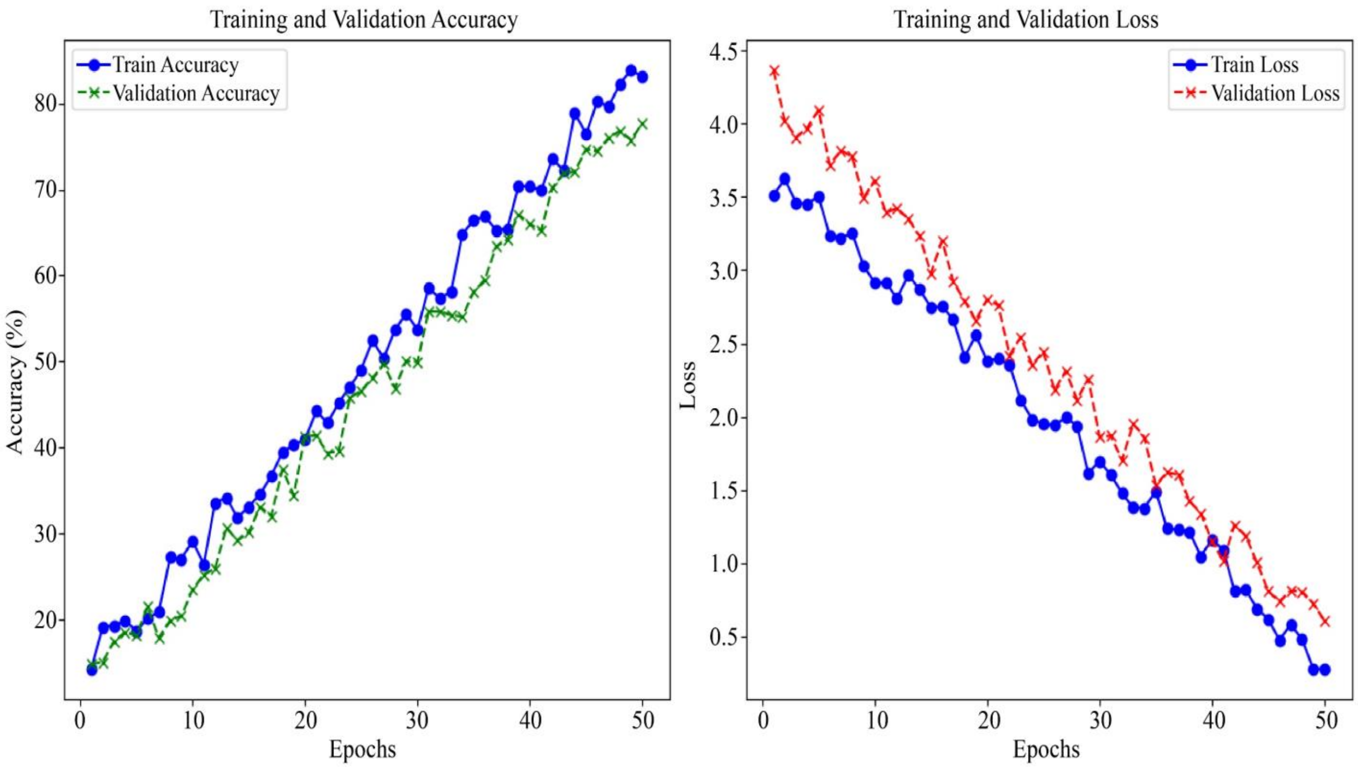
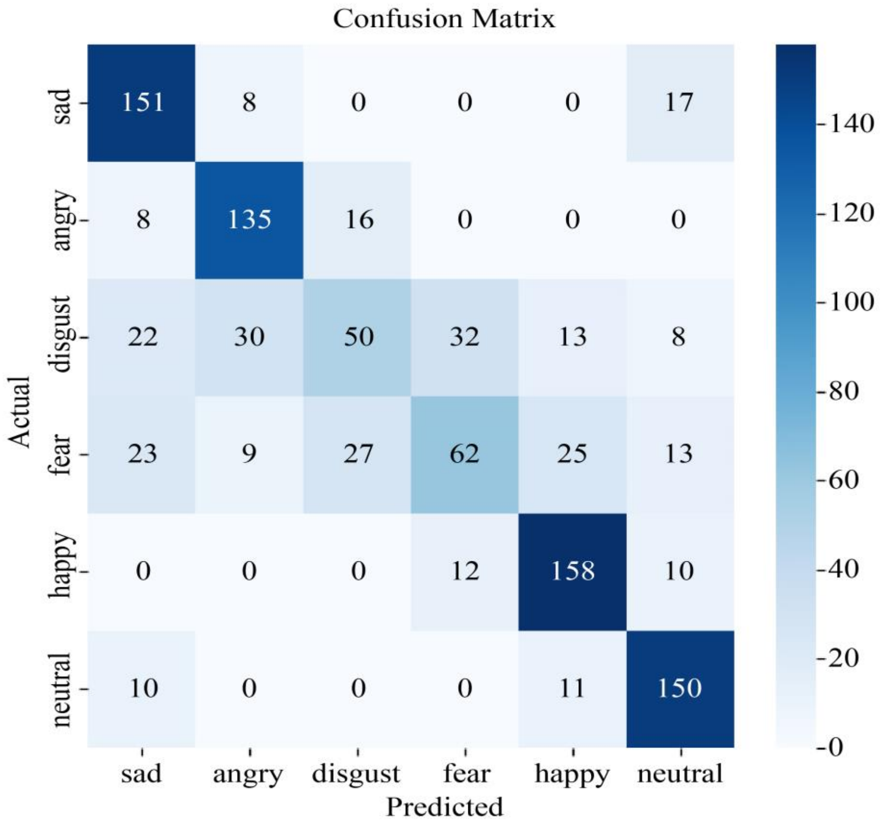


**Fig. 6 LSTM accuracy and loss**



| | sad | angry | disgust | fear | happy | neutral |
|---|---|---|---|---|---|---|
| sad | 151 | 8 | 0 | 0 | 0 | 17 |
| angry | 8 | 135 | 16 | 0 | 0 | 0 |
| disgust | 22 | 30 | 50 | 32 | 13 | 8 |
| fear | 23 | 9 | 27 | 62 | 25 | 13 |
| happy | 0 | 0 | 0 | 12 | 158 | 10 |
| neutral | 10 | 0 | 0 | 0 | 11 | 150 |

**Fig. 7 LSTM confusion matrix**

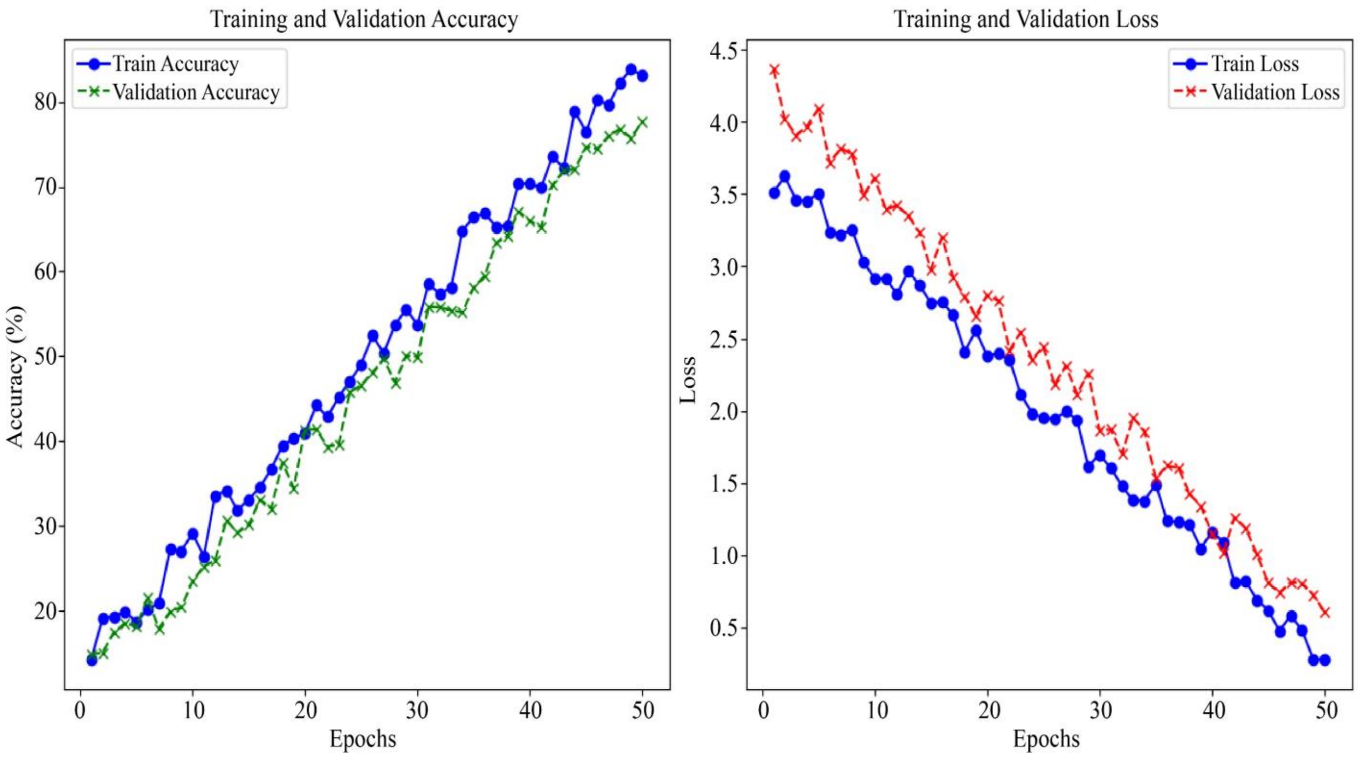


**Fig. 8 Bi-LSTM accuracy and loss**

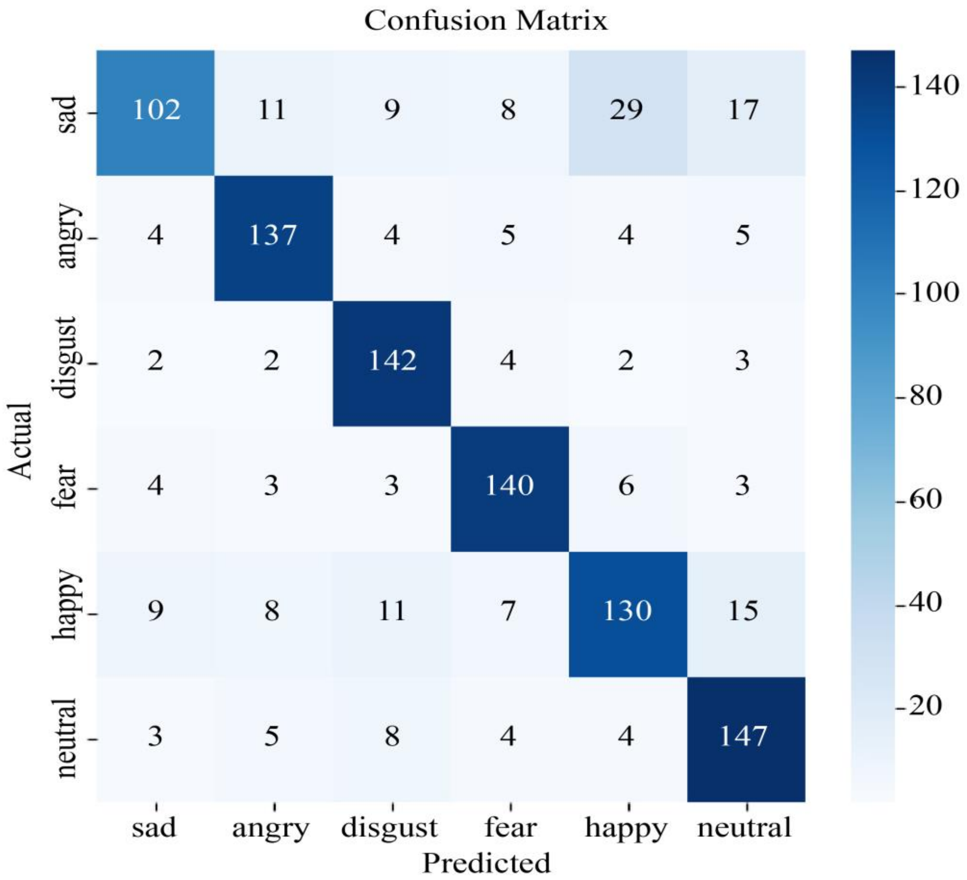


**Fig. 9 Bi-LSTM accuracy and loss**

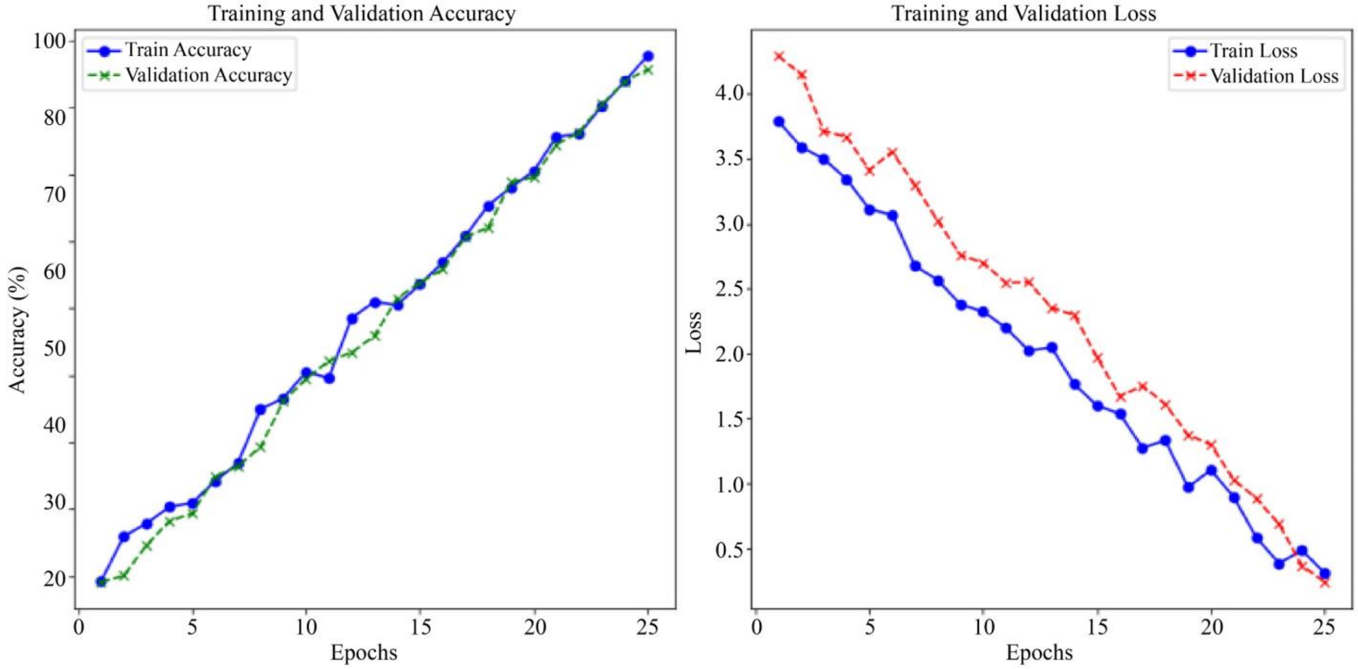


**Fig. 10 CNN accuracy and loss**

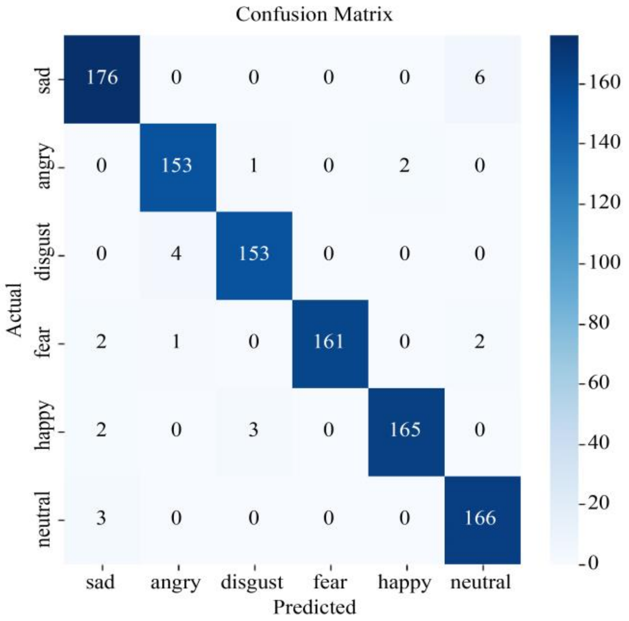


**Fig. 11 CNN confusion matrix**

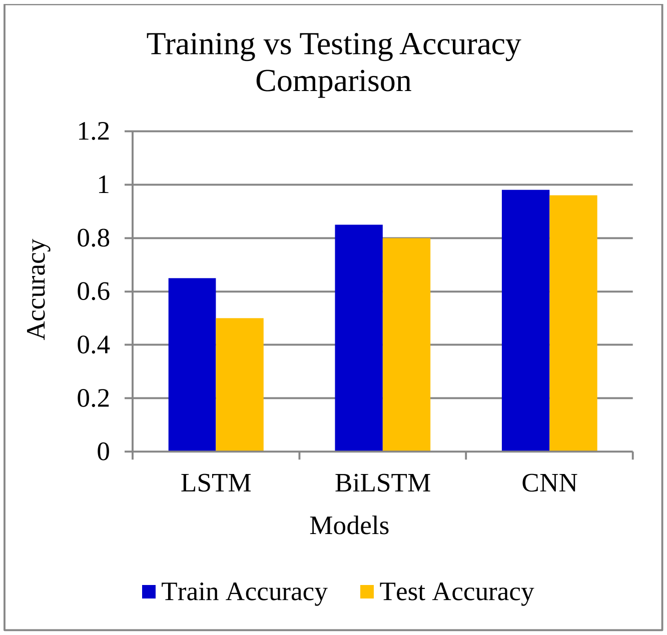


**Fig. 12 Comparison results of LSTM, Bi-LSTM, CNN**

## 5. Conclusion

The stream of Deep learning is expanding every moment, but there is still vast work to be done in the field of emotion detection. This paper presented an emotion detection model employing hybrid feature extraction techniques on the hybrid datasets. In this paper, various deep learning models are tested to check model accuracy. Results highlight that the highest accuracy achieved is 98.48% with CNN, 83% for Bi-LSTM and 63% in the LSTM model. This model also offers a novel approach to classifying emotion on hybrid datasets. Transfer models also underperform compared to the hybrid approach, emphasizing the effectiveness of our model in capturing emotional cues from speech data. Our end-to-end evaluation highlights the effectiveness in enhancing emotion recognition accuracy and reliability, making it a robust solution for real-world applications in speech analysis.

## References

[1] Lucas Goncalves et al., "Odyssey 2024-Speech Emotion Recognition Challenge: Dataset, Baseline Framework, and Results," *The Speaker and Language Recognition Workshop (Odyssey 2024)*, Quebec City, Canada, pp. 247-254, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Federico Costa, Miquel India, and Javier Hernando, "Double Multi-Head Attention Multimodal System for Odyssey 2024 Speech Emotion Recognition Challenge," *arXiv preprint*, pp. 1-8, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Mohamed A. Gismelbari et al., "Speech Emotion Recognition Using Deep Learning," *2024 XXVII International Conference on Soft Computing and Measurements (SCM)*, Saint Petersburg, Russian Federation, pp. 380-384, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Zheng Lian et al., "MER 2024: Semi-Supervised Learning, Noise Robustness, and Open-Vocabulary Multimodal Emotion Recognition," *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, Melbourne, VIC Australia, pp. 41-48, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Wahiba Ismaiel et al., "Deep Learning, Ensemble and Supervised Machine Learning for Arabic Speech Emotion Recognition," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13757-13764, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[6] A.V. Geetha et al., "Multimodal Emotion Recognition with Deep Learning: Advancements, Challenges, and Future Directions," *Information Fusion*, vol. 105, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Sumera, K. Vaidehi, and Qamar Nisha, "A Machine Learning and Deep Learning based Approach to Generate a Speech Emotion Recognition System," *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, India, pp. 573-577, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Jarod Duret, Mickael Rouvier, and Yannick Esteve, "MSP-Podcast SER Challenge 2024: L'antenne du Ventoux Multimodal Self-Supervised Learning for Speech Emotion Recognition," *arXiv preprint*, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] S. Ashmad Ahemed, and D. Jayakumar, "Voice Assisted Facial Emotion Recognition System for Blind Peoples with Tensorflow Model," *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, Bhopal, India, pp. 1-4, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] M. Madhura et al., "Neural Networks and Emotions: A Deep Learning Perspective," *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, Pune, India, pp. 1-7, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Daria Diatlova et al., "Adapting WavLM for Speech Emotion Recognition," *arXiv preprint*, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Pragya Singh Tomar, Kirti Mathur, and Ugrasen Suman, "Fusing Facial and Speech Cues for Enhanced Multimodal Emotion Recognition," *International Journal of Information Technology*, vol. 16, pp. 1397-1405, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Anandakumar Haldorai et al., "Bi-Model Emotional AI for Audio-Visual Human Emotion Detection Using Hybrid Deep Learning Model," *Artificial Intelligence for Sustainable Development*, pp. 293-315, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] N. Kapileswar et al., "An Intelligent Emotion Recognition System based on Speech Terminologies using Artificial Intelligence Assisted Learning Scheme," *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, Chennai, India, pp. 1-7, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Sudipta Bhattacharya et al., "Emotion Detection from Multilingual Audio using Deep Analysis," *Multimedia Tools and Applications*, vol. 81, pp. 41309-41338, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[16] Henry Harm, and Tanel Alum, "TalTech Systems for the Odyssey 2024 Emotion Recognition Challenge," *The Speaker and Language Recognition Workshop*, Quebec City, Canada, pp. 1-5, 2024. [Google Scholar] [Publisher Link]

[17] Jain Joseph, R.P. Aneesh, and Joseph Zacharias, "Deep Learning based Emotion Recognition in Human-Robot Interaction with Multi-Modal Data," *Fourth International Conference on Advances in Physical Sciences and Materials*, Coimbatore, India, vol. 3122, no. 1, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Sung-Woo Byun, and Seok-Pil Lee, "A Study on a Speech Emotion Recognition System with Effective Acoustic Features using Deep Learning Algorithms," *Applied Sciences*, vol. 11, no. 4, pp. 1-15, 1890. [CrossRef] [Google Scholar] [Publisher Link]

[19] Waleed Akram Khan, Hamad ul Qudous, and Asma Ahmad Farhan, "Speech Emotion Recognition using Feature Fusion: A Hybrid Approach to Deep Learning," *Multimedia Tools and Applications*, vol. 83, pp. 75557-75584, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Rudra Tiwari et al., "Emotion Detection through Human Verbal Expression Using Deep Learning Techniques," *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, Bhopal, India, pp. 1-7, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] Omayma Mahmoudi, and Mouncef Filali Bouami, “Arabic Speech Emotion Recognition using Deep Neural Network,” *International Conference on Digital Technologies and Applications*, Fez, Morocco, pp. 124-133, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[22] Davit Rizhinashvili, Abdallah Hussein Sham, and Gholamreza Anbarjafari, “Enhanced Speech Emotion Recognition using Averaged Valence Arousal Dominance Mapping and Deep Neural Networks,” *Signal, Image and Video Processing*, vol. 18, pp. 7445-7454, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[23] Shrikala Deshmukh, and Preeti Gupta, “Application of Probabilistic Neural Network for Speech Emotion Recognition,” *International Journal of Speech Technology*, vol. 27, pp. 19-28, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Ravi Raj Choudhary, Gaurav Meena, and Krishna Kumar Mohbey, “Speech Emotion based Sentiment Recognition using Deep Neural Networks,” *Journal of Physics: Conference Series*, vol. 2236, pp. 1-11, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[25] Kishor Bhangale, and Mohanaprasad Kothandaraman, “Speech Emotion Recognition based on Multiple Acoustic Features and Deep Convolutional Neural Network,” *Electronics*, vol. 12, no. 4, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[26] Valli Madhavi Koti et al., “Speech Emotion Recognition using Extreme Machine Learning,” *EAI Endorsed Transactions on Internet of Things*, vol. 10, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[27] Tae-Wan Kim, and Keun-Chang Kwak, “Speech Emotion Recognition Using Deep Learning Transfer Models and Explainable Techniques,” *Applied Sciences*, vol. 14, no. 4, pp. 306-311, 2024. [CrossRef] [Google Scholar] [Publisher Link]