Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Hybrid Feature Extraction Technique with Data Augmentation for Speech Emotion Recognition Using Deep Learning

Priyanka Joshi[1], Sonika Kandari[2,*]
[1]Research Scholar, School of Engineering & Technology, Shri Guru Ram Rai, University, Dehradun, Uttarakhand, India
[2]Professor, School of Engineering & Technology, Shri Guru Ram Rai University, Dehradun, Uttarakhand, India
[1]5feb.priyanka@gmail.com
Corrosponding author: *dean.set@sgrru.ac.in

ARTICLE INFO

Received: 18 Dec 2024

Revised: 14 Feb 2025

Accepted: 24 Feb 2025

ABSTRACT

Prediction of emotions from human speech by machines is termed as speech emotion recognition. Speech is one of the most common and fastest methods of communication between humans. Speech emotion recognition (SER) by machines is a challenging task. Various deep learning algorithms are trying to make machines having such learning capabilities to achieve this task. Several researches are being conducted toward this area but identifying correct emotions from human speech is still challenging. The process of speech emotion recognition consists of three main stages – the feature extraction, feature selection and classification. The feature extraction is considered as the most significant among them. Several researches work has been conducted in the past years and most of them used one technique of feature extraction for training the model. In this paper we have presented hybrid feature extraction technique with data augmentation for an effective emotion recognition model. The simulations are performed by using TESS dataset. From the speech data two sets of features are extracted by using two techniques, with first technique we use mel frequency cepstral coefficient (MFCC)as feature sets and in the next approach the mel spectrogram images have been extracted The data augmentation technique has been proposed by noise addition to increase the number of samples. Then a Convolutional Neural network (CNN) is implemented for training and testing the model to achieve better accuracy. Our proposed hybrid feature set with data augmentation technique achieved the accuracy of 95.21%.

Keywords: Speech emotion recognition (SER); deep learning; machine learning, convolutional neural network.

INTRODUCTION

Speech is one of the common methods of communication between human beings. Humans are able to detect emotions naturally while communicating each others. Therefore in speech emotion recognition by machines, we train machines in such a way to predict emotions from human speech. A speech emotion recognition system takes the file containing speech data as input and classifies the speech data into various emotions such as happy, sad, neutral etc. With the increase of technology, the interaction of computer with human beings has also increased .Therefore over the past two decades the researchers are working on, to study several methods to increase the efficiency of human computer interaction[1].

An effective Speech emotion recognition can increase the efficiency of human computer interaction by improving accuracy and intimacy of interaction. We can utilize an effective SER in various fields such as medical emergencies, automotive engineering and also in decision making tasks[2].An effective SER can also applied as an intelligent household robot[3].

SER has been a challenging task due to several reasons[4], one of the most prominent reason is ,that the emotions are subjective and expressions may vary across individual to individual, and the other one is the speech tone pattern ,vocal cords cues are different to exhibit the same emotion by different people. Therefore selection of desired features is necessary by comparing feature sets from speech signals[5].

The need of high quality datasets are very useful for training the model, sometime same utterance of speech exhibit the different emotions because the emotions depend on context
also, therefore situational context shows the emotions. Due to these several reasons it is not easy to predict the emotions correctly and hence research are going on to find better accuracy in evaluating SER. therefore there is a great impact of feature selection for predicting emotions[6]. Data augmentation is also used for increasing the number of samples to increase the possibility of improved efficiency[7]. The motivation of this study is also to obtain better SER. This research proposes the MFCC feature of audio signal as input and Mel Spectrogram images as input for evaluating SER with the use of CNN model.

LITERATURE REVIEW

There is a vast research interest in speech emotion recognition. Several research works are
performed to attempt for prediction of emotions from human speech. There are many research studies which are focusing on how to get improvement in achieving accurate emotion prediction from speech. The research work[8] used the prosodic and linguistic features from speech for identifying emotions.

In the work[9] reflect the deep learning approaches for speech emotion recognition. The research presents a comparative study between neural network approaches used for emotion prediction from speech signal.

In the study [10] the author has described a fusion algorithm by fusion of spatial and temporal features with using parallel CNNs.

In one of the research work[11], sparse auto encoder and gated recurrent unit has been utilized for speech feature extraction .The deeper text features are extracted .the Bi-LSTM model and loss function has been utilized to find accuracy. In one of the research work[12] traditional Hidden Markov Model(HMM)has been introduced for SER.Two methods have been used ,one by deriving features of speech signals using Gussain Mixture Model(GMM) and other by introducing HMM which derive the low level in stantaneous features of audio data the accuracy given by this method is comparatively not as good as modern deep learning techniques are performing.

The work[13] suggested the algorithm which introduce the impact of modified feature extraction in speech emotion recognition and also explored the various benefits of deep convolutional neural network. Pattern recognition method is suggested by researchers [14] for emotion recognition from speech signal.

In the study[15] proposed a joint model architecture that joins discrete and dimensional emotions for SER. It has used self attention and co-attention mechanism to improve the accuracy of emotion prediction. One of the research methodologies proposed as safety-security system by detecting negative emotions from speech. The research used CNN and LSTM model for detecting some keywords based on negative emotions[16]. In the study[17]) proposed the graph theory for classifying emotions extracted from speech., with the help of graphical tools structural and statistical information of time series has been captured, which proved beneficial for identifying emotions. Hand crafted feature set was utilized in the research work with ensemble deep neural network using different speech datasets to conduct the research[18].Machine learning techniques are being best utilized by number of researchers for finding the accuracy in emotion detection[19].

In one of the research emotion prediction accuracy has been improved with the help of data augmentation technique and shifting of spectrogram using different speech datasets. Deep learning algorithms like CNN and multilayer perception with Bi-dimensional LSTM has been deployed[20]. The work[21]combination of features such as MFCC, Spectral,Harmonis etc. with Recurrent neural Network(RNN) are used for better accuracy in SER.

METHODOLOGY

The main steps in Speech emotion recognition is to derive the unique and suitable features from the speech data, and then select the suitable classifier to accurately predict the emotions from the audio speech signals.

In this section we have explained the proposed methodology to predict emotions from audio file, here we have used two approaches of feature extractions and derived two features sets from audio signals, first we have extracted MFCC from speech signals and other the spectrogram features of input audio signals. We have also done data augmentation technique by noise addition to increase the dataset samples[22].

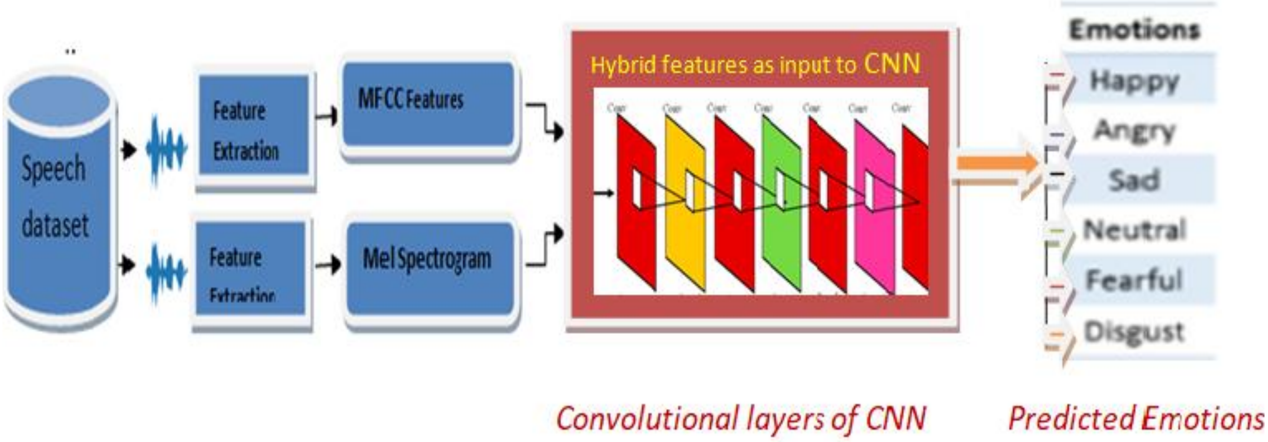


Figure 1: Architecture of proposed hybrid feature extraction using CNN model

The proposed work suggests a method in which this hybrid set of features is extracted from audio file. Then we have proposed data augmentation to increase the data samples. These feature set are further processed using CNN model.

Data augmentation technique is used to make our dataset more robust by increasing the number of samples for better accuracy in real time environment.

Data augmentation has been utilized here by adding noise and extends the range of speech signal pattern and hence increases the number of samples.

The framework TensorFlow and libraries like NumPy help in data augmentation process while training the model.

In the first approach we have used audio file from database to export Mel frequency Cepstral Coefficient (MFCCs) features from the audio speech signal, speech signals are pre emphasized, after that frame blocking is done, it is then goes under windowing to reduce discontinuities of signal In the next approach we extract the a Mel spectrogram features from the audio datasets.

These hybrid features are then processed under CNN model for training and testing to evaluate emotion prediction. There are different datasets are available for research[23],we have utilized TESS dataset.

THE TESS Dataset:

The Toronto Emotional Speech Set is one of the dataset used in emotion prediction from audio speech. It consists of two female speakers (actresses) of age 24 and 64 years old. It consists of total seven emotions happy, fear, sad, neutral, surprise, anger, disgust. This dataset has 200 target words which speaker has to speak with the phrase 'say the word……." The dataset contains 2800 audio files. The file contains WAV format. This dataset has high quality audio file.

Figure 2.(a) Sample of Waveform

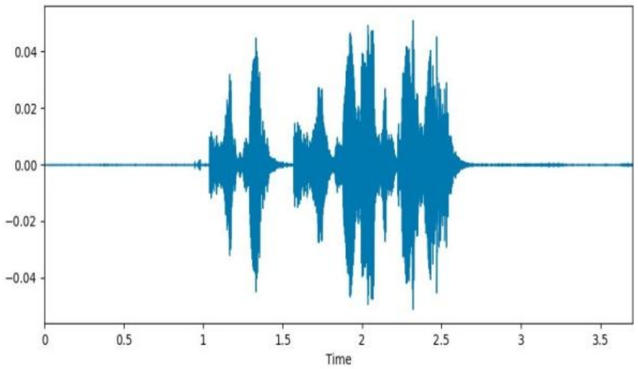
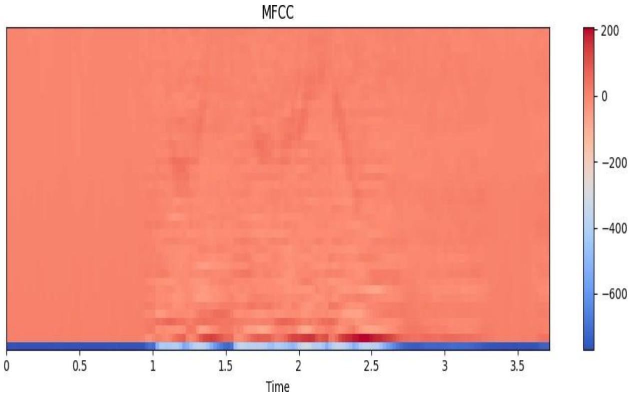


Figure 2.(b) Sample MFCC



Here we have utilized 80% of data for training CNN model and rest 20% data for testing the model.
The proposed model performance has been evaluated by comparing with other SER models
We have used librosa library to export the MFCC s sequences from the audio file presence in
The speech contains the emotions including happy, sad, calm, surprise, disgust, fearful and angry. We have classified the emotions here are angry, disgust, fear, happy, sad. For performance evaluation we have employed confusion matrix for these five emotions.
In this work we have utilized two methods to propose two types of feature extraction. In the first way, we extracted MFCC features and then Mel-spectrogram images have been extracted as feature set. The librosa library is used for extracting features. The preprocessing is done afterwards using sklearn to reduce the high dimensionality. Data Augmentation has done for increasing number of samples. The total of 2800 features has been extracted, and after augmentation samples increased to 8400 and the model is trained for 100 epochs.
The MFCC generation from the audio speech signal is performed. In the MFCC feature the Mel scale is formulated as-

Mel (F)-= 2595*log10 (1+f/700)
F=actual frequency, Mel(f)=perceived frequency

Spectrogram generation of speech signal is best suited to learn the foremost features required for SER[24]. Similarly waveform generation of speech signals are generated[25].
Spectrogram is the visual representation of these speech signals with their frequencies over different time axis.
The extracted spectrogram for different emotions in speech signal is shown below with frequencies (f) and time (t) on x axis and y axis respectively.

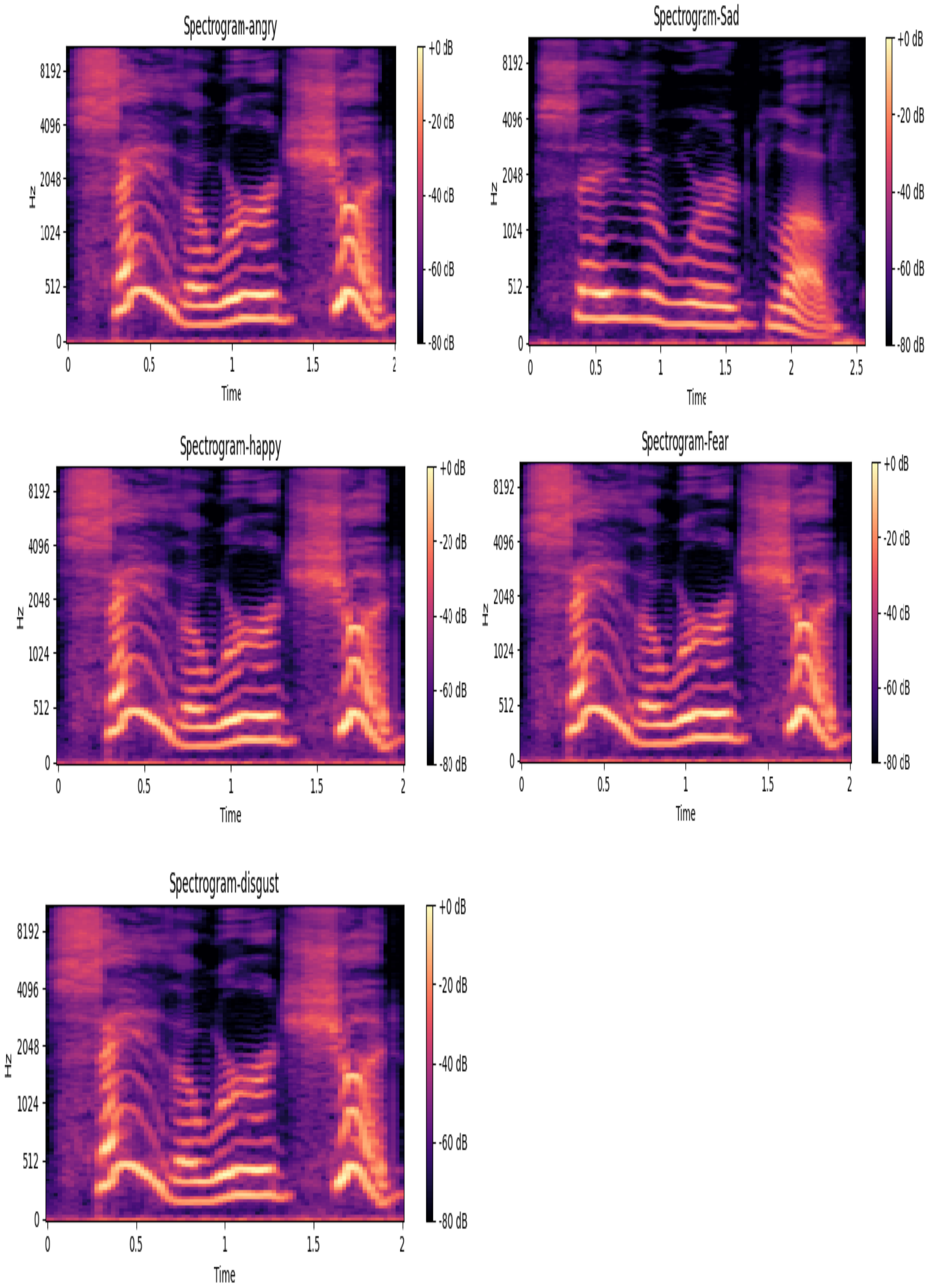


Figure 3. The extracted Spectrograms of different emotions present in speech signal.

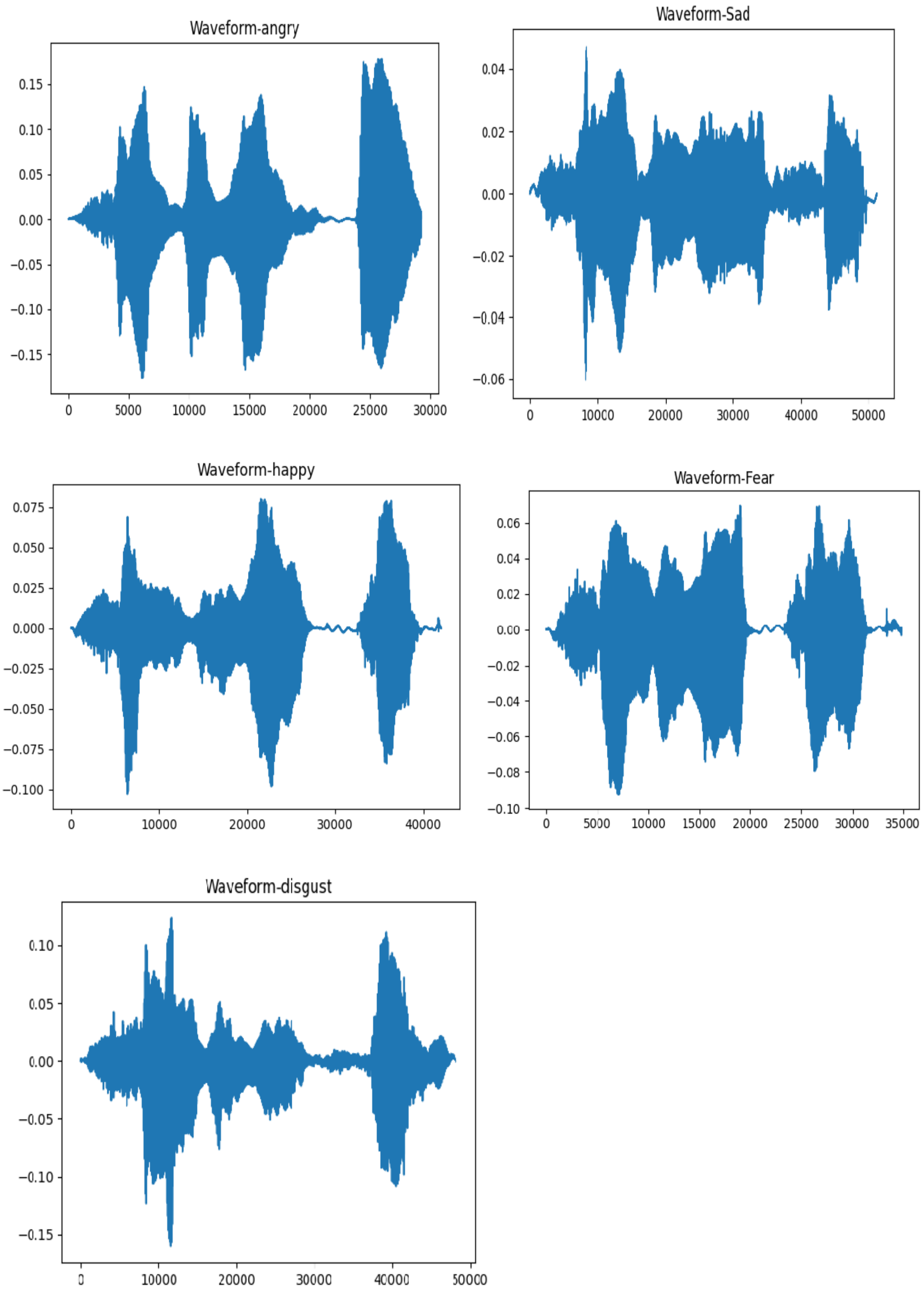


Figure 4. The extracted Waveforms for different emotions present in speech signal

Evaluation Confusion Matrix:

The performance of our experiment has been evaluated with the weighted accuracy. The weighted accuracy is termed as the ratio of the sum of all correctly predicted true positive and true negative values with all predicted values. The formula is:

$$Accuray = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{TP+TN}{TP+TN+FP+FN}\right)_i$$

CNN architecture:

The Convolutional Neural Network is a deep learning algorithm and a type of neural network. It comprises of multiple layers like convolution layers, pooling layers and fully connected layers. The basic architecture of CNN is based on the human brains processing technique. It comprises of artificial neural network. The CNN is doing well in capturing hierarchical pattern in images[26]. The different layers of CNN are:

Convolutional layers: It constitutes the essential part of CNN, which find extensive application in image processing. The function of CNN comprises:

Filters/Kernel: To identify particular features, these tiny matrices such as 3*3 or 5*5 are slid over the input image .Every filter is made to be react to specific patterns like textures or edges. An activation map also known as feature map is produced by repeating the process over the whole image. This layer also maintains the spatial dependencies among pixels.

Pooling Layers: This is the essential layer which assists in numerous ways-

Dimension Reduction: Pooling layer reduces the amount of computation needed for succeeding layers by down sampling the spatial dimensions.

Parameter reduction: By condensing the size of feature maps, pooling reduces the number of parameters and computation.

Feature extraction: Max pooling keeps the most important features (max value) from a group of adjacent pixels to identify the most salient characteristics and also eliminate irrelevant information. Pooling also introduces invariance, when translating input data an concentrates on most important data.

Activation Functions: Rectified Linear Unit (ReLU) and other non-linear activation functions add non-linearity to the model, enabling it to discover more intricate links in the data to adjacent pixel groups.

Fully Connected Layers: These layers use the high-level features that were taught to them by the preceding layers to make predictions. All of the neurons in one layer are connected to all of the neurons in the layer below.

RESULT

In this research we trained the SER model on TESS dataset using Spectrogram and MFCC we conducted the experiment on these two types of features (hybrid features).The data was spitted into 80/20 means 80 percent data is used for training and 20 percent was used for testing the model using the sklearn library. The Data augmentation technique is also utilized here. The extracted samples are 2800 first and increased after augmentation process to 8400.

These hybrid features are fed into the CNN model to evaluate the result. The description of CNN layers we have used here are mentioned, firstly Convolution1D (Conv_1D), then we activated function ‘relu’, after that Dropout layer which is used to avoid over fitting we have used maxpooling layer. We used the flatten layer to convert the n dimensional matrix, which we obtained from the above layers for providing input for the dense layer. Then finally dense layer provide the input to the softmax activation function .The softmax activation function give the different probabilities which we get by training our model.

Therefore in total we have used 13862 trainable parameters in our model.

Layer(type)	Output shape	parameters
conv2d_1(Conv2D)	(None,13,1,32)	320
flatten_1(Flatten)	(None,416)	0
dense_2(Dense)	(None,32)	13,344
Dense_3(Done)	(None,6)	198

Total parameters: 13,862(54.15 KB)
Trainable parameters: 13,862(54.15 KB)
Non-Trainable parameters: 0(0.00 B)

We train our model in 100 epochs, as the number of epochs increases we can see the increased accuracy
The evaluation is done by measuring the model in terms of confusion matrix.
The accuracy shown by our model is 95.21%

The loss graph and accuracy graphs are shown below:

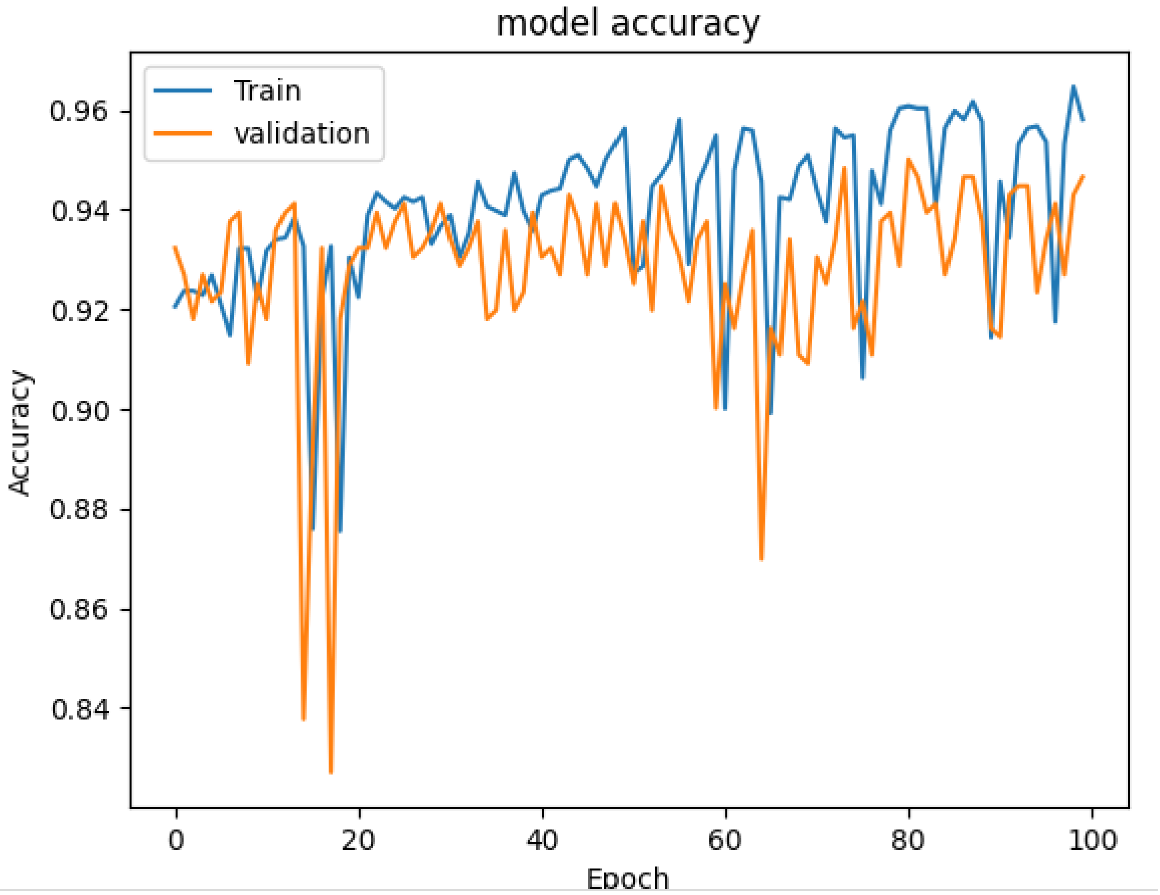


Figure 5(a) Model accuracy graph

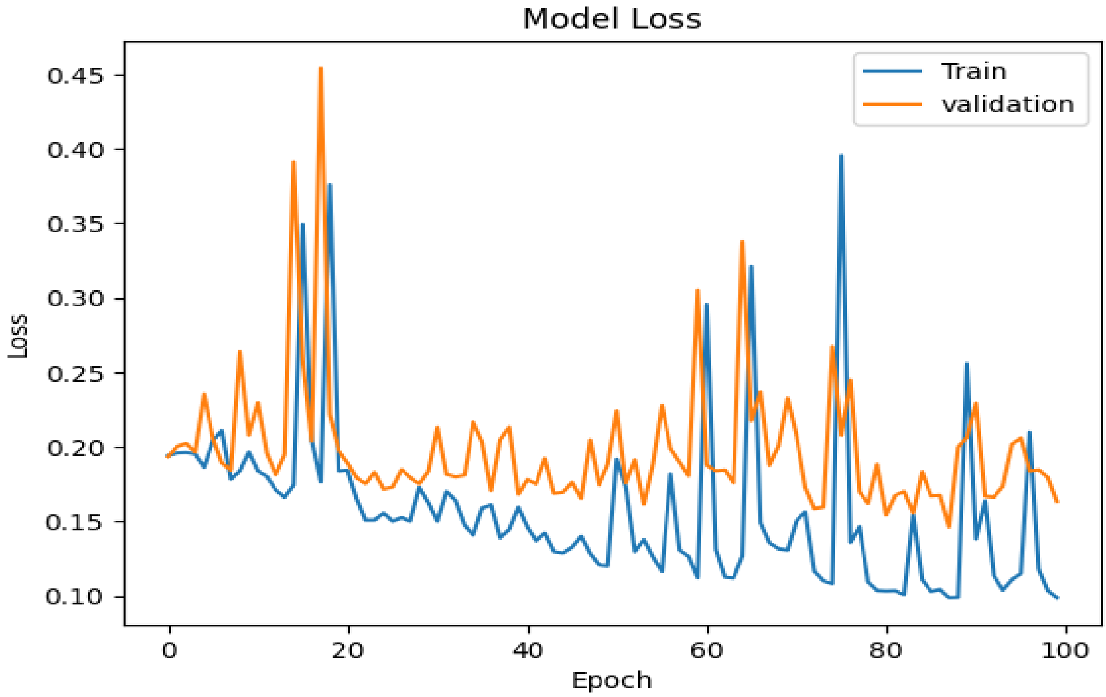


Figure 5(b).Model loss graph

Confusion matrix:

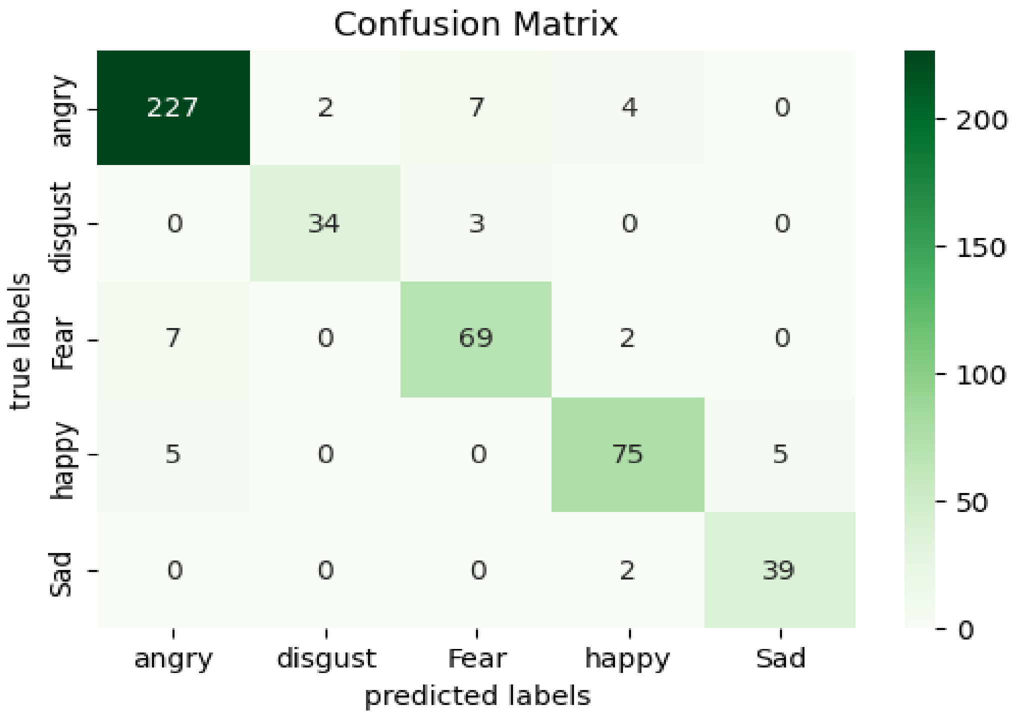


Figure 6. Confusion matrix of different emotions for TESS using CNN

Table 1.Comparative analysis of proposed approach with other approaches using TESS dataset:

Reference	Method	Input features	Dataset	Accuracy
[27]	LSTM-CNN	Mel-Spectrogram, MFCC	TESS, RAVDESS	68%
[28]	LSTM with CNN	MFCC	TESS	88.92%
[33]	CNN	MFCC, Mel Spectrogram	TESS	89%
	Proposed Hybrid MFCC, Spectrogram	MFCC, Mel Spectrogram images	TESS	95.21%

Table 2. Comparative analysis of proposed model with other approach using different datasets

Ref	Year	Model	Input features	Dataset	Accuracy
[29]	2019	Hybrid DNN and SVM based approach	Heterogeneous accoustic features	IEMOCAP	64%
[30]	2019	CNN+Bi-LSTM based model	MFCC,Spectrogram	IEMOCAP EMO-DB	50.05% 82.35%
[31]	2020	DSCNN(Deep Stride CNN)	Spectrogram	RAVDESS IEMOCAP	80.0% 84%
[6]	2020	DCNN+CFS+ML	Log-Mel Spectrogram	RAVDESS IEMOCAP SAVEE EMO-DB	81.30% 83.80% 83.80% 82.10%
[32]	2021	CNN+LSTM	MFCC	IEMOCAP	79.52%
[7]	2022	Combined CNN with LSTM,MLP,SVM (Wav2vec2.0)	Acoustic features based on Wav2vec2.0	IEMOCAP JTES(Japanese Twitter based emotion speech	76.39% 77.25%
[11]	2024	Attention layer, Duel Bi-LSTM	eGeMAPS, deeper text features	IEMOCAP	74.27%
[15]	2025	Pre-trained Wav2vec2.0 and HuBERT model	MFCC, short term features ,Mel Spectrogram and raw audio signal	IEMOP	73.22%
	2025	Proposed Hybrid feature extraction model using CNN	MFCC, Mel Spectrogram images	TESS	95.21%

CONCLUSION

This research work proposed a CNN based method with hybrid feature extraction of audio file for recognition of emotion from speech signal. Using TESS dataset we export Mel Frequency Cepstral Coefficient (MFCCs from audio file and then Mel Spectrogram images, with the use of CNN model, the experimental result shows that the proposed model has achieved a good performance. Therefore this research contributes several directions for future work in this area. Though the large literature on speech emotion recognition system is present, still it has many challenges, SER is not an easy task for several reasons, like the subjective nature of emotions and also different person uses different speech pattern, tones, and vocal signals and also there is a great need to focus on feature extraction and selection.. Therefore the proposed work confirm that it would contribute towards a more robust SER.

REFERENCES

[1] B. Basharirad and M. Moradhaseli, "Speech emotion recognition methods: A literature review," *AIP Conf. Proc.*, vol. 1891, no. October 2018, 2017, doi: 10.1063/1.5005438.

[2] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam, "Emotion and decision making," *Annu. Rev. Psychol.*, vol. 66, no. September 2014, pp. 799–823, 2015, doi: 10.1146/annurev-psych-010213-115043.

[3] X. Huahu, G. Jue, and Y. Jian, "Application of speech emotion recognition in intelligent household robot," *Proc. - Int. Conf. Artif. Intell. Comput. Intell. AICI 2010*, vol. 1, pp. 537–541, 2010, doi: 10.1109/AICI.2010.118.

[4] A. Al-Talabani, H. Sellahewa, and S. A. Jassim, "Emotion recognition from speech: tools and challenges," in *Mobile Multimedia/Image Processing, Security, and Applications 2015*, SPIE, May 2015, p. 94970N. doi: 10.1117/12.2191623.

[5] T. Vogt and E. Andr, "COMPARING FEATURE SETS FOR ACTED AND SPONTANEOUS SPEECH IN VIEW OF AUTOMATIC EMOTION RECOGNITION Augsburg University , Germany Multimedia concepts and applications Applied Computer Science," *Proc. IEEE Inter-Natl. Conf. Multimed. Expo*, pp. 474– 477, 2005.

[6] M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. Bin Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," *Sensors (Switzerland)*, vol. 20, no. 21, pp. 1–18, 2020, doi: 10.3390/s20216008.

[7] B. T. Atmaja and A. Sasou, "Effects of Data Augmentations on Speech Emotion Recognition," pp. 1–14, 2022.

[8] M. Pervaiz and T. Ahmed, "Emotion Recognition from Speech using Prosodic and Linguistic Features," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 8, 2016, doi: 10.14569/ijacsa.2016.070813.

[9] S. K. Pandey, H. S. Shekhawat, and S. R. M. Prasanna, "Deep learning techniques for speech emotion recognition: A review," *2019 29th Int. Conf. Radioelektronika, RADIOELEKTRONIKA 2019 - Microw. Radio Electron. Week, MAREW 2019*, no. July, pp. 1–6, 2019, doi: 10.1109/RADIOELEK.2019.8733432.

[10] Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020, doi: 10.1109/ACCESS.2020.2990405.

[11] S. Zhang *et al.*, "Multi-Modal Emotion Recognition Based on Wavelet Transform and BERT-RoBERTa : An Innovative Approach Combining Enhanced BiLSTM and Focus Loss Function," 2024.

[12] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," *Proc. - IEEE Int. Conf. Multimed. Expo*, vol. 1, no. August 2003, pp. I401–I404, 2003, doi: 10.1109/ICME.2003.1220939.

[13] S. Langari, H. Marvi, and M. Zahedi, "Efficient speech emotion recognition using modified feature extraction," *Informatics Med. Unlocked*, vol. 20, Jan. 2020, doi: 10.1016/j.imu.2020.100424.

[14] K. Sim, "Pattern Recognition Methods for Emotion Recognition with speech signal," no. August, 2014, doi: 10.5391/IJFIS.2006.6.2.150.

[15] J. L. Bautista and H. S. Shin, "Speech Emotion Recognition Model Based on Joint Modeling of Discrete and Dimensional Emotion Representation," pp. 1–20, 2025.

[16] S. Jena, S. Basak, H. Agrawal, B. Saini, S. Gite, and K. Kotecha, "Developing a negative speech emotion recognition model for safety systems using deep learning," *J. Big Data*, 2025, doi:

10.1186/s40537-025-01090-0.

[17] A. Pentari, G. Kafentzis, and M. Tsiknakis, “Speech emotion recognition via graph - based representations,” *Sci. Rep.*, pp. 1–11, 2024, doi: 10.1038/s41598-024-52989-2.

[18] J. H. Chowdhury, S. Ramanna, and K. Kotecha, “Speech emotion recognition with light weight deep neural ensemble model using hand crafted features,” pp. 1–14, 2025.

[19] P. K. S. Raja and P. D. D. Sanghani, “Speech Emotion Recognition Using Machine Learning,” vol. 30, no. 6, pp. 118–124, 2024, doi: 10.53555/kuey.v30i6(S).5333.

[20] C. Barhoumi and Y. Benayed, “Real-time speech emotion recognition using deep learning and data augmentation,” 2025.

[21] S. Byun and S. Lee, “applied sciences A Study on a Speech Emotion Recognition System with Effective Acoustic Features Using Deep Learning Algorithms,” 2021.

[22] A. A. Abdelhamid, S. Member, A. Ibrahim, and M. M. Eid, “Robust Speech Emotion Recognition Using CNN + LSTM Based on Stochastic Fractal Search Optimization Algorithm,” *IEEE Access*, vol. 10, pp. 49265–49284, 2022, doi: 10.1109/ACCESS.2022.3172954.

[23] M. Zielonka, A. Piastowski, A. Czyżewski, P. Nadachowski, M. Operlejn, and K. Kaczor, “Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets,” *Electron.*, vol. 11, no. 22, 2022, doi: 10.3390/electronics11223831.

[24] H. A. O. Meng, T. Yan, F. E. I. Yuan, and H. Wei, “Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network,” *IEEE Access*, vol. 7, pp. 125868–125881, 2019, doi: 10.1109/ACCESS.2019.2938007.

[25] C. Appl, S. Technol, V. No, S. N. Pai, and S. Punnath, “Emotion Classification from Speech Waveform Using Machine Learning and Deep Learning Techniques,” 2024.

[26] R. Begazo, A. Aguilera, and I. Dongo, “A Combined CNN Architecture for Speech Emotion Recognition,” pp. 1–39, 2024.

[27] K. Venkataramanan, “Emotion Recognition from Speech,” pp. 1–14, 2011.

[28] “Journal of Computer Science,” no. 2, pp. 54–67, 2022.

[29] W. Jiang, Z. Wang, J. S. Jin, X. Han, and C. Li, “Speech emotion recognition with heterogeneous feature unification of deep neural network,” *Sensors (Switzerland)*, vol. 19, no. 12, pp. 1–15, 2019, doi: 10.3390/s19122730.

[30] B. J. Abbaschian, D. Sierra-sosa, and A. Elmaghraby, “Deep Learning Techniques for Speech Emotion Recognition , from Databases to Models,” 2021.

[31] S. Kwon, “A CNN-Assisted Enhanced Audio Signal Processing,” *Sensors*, 2020.

[32] J. Liu and H. Wang, “A Speech Emotion Recognition Framework for Better Discrimination of Confusions,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 1, pp. 586–590, 2021, doi: 10.21437/Interspeech.2021-718.

[33] U A and K. V K, "Speech Emotion Recognition-A Deep Learning Approach," *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Palladam, India, 2021, pp. 867-871, doi: 10.1109/I-SMAC52330.2021.9640995.