Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 9, 2025

ENHANCING SPEECH EMOTION RECOGNITION WITH DEEP LEARNING THROUGH DATA FUSION, SPECTROGRAM AUGMENTATION, AND HYBRID FEATURE INTEGRATION

Muhammad Talha Jahangir[*1], Mujahid Hussain [2], Nashitah Alwaz[3], Muhammad Musawir Saeed[4], Waheed Ahmad[5], Uzair Ahmad[6], Hammad Toheed Khan[7]

[*1,2,3,4,5,6,7] *Department of Computer Science, MNS-University of Engineering & Technology Multan, Multan, Pakistan*

[*1]mtalhajahangir@mnsuet.edu.pk

DOI:https://doi.org/10.5281/zenodo.17189507

Keywords
Deep Learning, Data Fusion, Convolutional Neural Network, Bidirectional Long Short-Term Memory, Human-Computer Interaction (HCI), Spectrogram Augmentation, Speech Emotion Recognition; MFCC; Mel Spectrogram; Root mean square.

Article History
Received: 02 July 2025
Accepted: 12 September 2025
Published: 24 September 2025

Corresponding Author: *
Muhammad Talha Jahangir

Abstract
Speech Emotion Recognition (SER), which lets computers decode human feelings using vocal clues, is among the most vital elements of affective computing. The range of speech patterns, lack of data, and difficulty of emotional expression make it still difficult to get excellent SER accuracy. Data fusion from four baseline datasets RAVDESS, TESS, CREMA-D, and SAVEE is used by our proposed deep learning-based SER architecture. The suggested model design combines Convolutional Neural Networks (CNNs) with Bidirectional Long Short-Term Memory (BiLSTM) to efficiently capture spatial and temporal characteristics. With a remarkable classification accuracy of 98%, the proposed framework improving SER performance and giving computers the ability to immediately detect and respond to human feelings that helps our system foster a more sympathetic and flexible human-computer connection.

I. INTRODUCTION

Machines must now be able to recognize human emotions if human-computer interaction is to be improved. This is more important than ever. One of the main elements of affective computing, speech emotion recognition (SER), seeks to automatically determine the emotional state of a speaker using their voice [1][2]. The number of affiliated publications in Speech Emotion Recognition (SER) has been graphed year by year from 2013 to 2025. At the beginning (2013–2016), the research activity increased in small steps as more people became interested in the field. Starting in 2017, publications increased much faster and the biggest jump took place between 2017 and 2021. It shows that people in the industry and academic fields are paying more attention to SER due to recent progress in deep learning, processing language and increasing access to information about emotional speech. By

2021, the rate at which the number of publications had grown slowed and the number fact that this trend is unfolding shows that SER has grown in importance these years and research levelled off, showing that the field might be maturing. The outcomes could become less variable as both fundamental and profitable uses of solar energy become more common. as shown in figure 1.

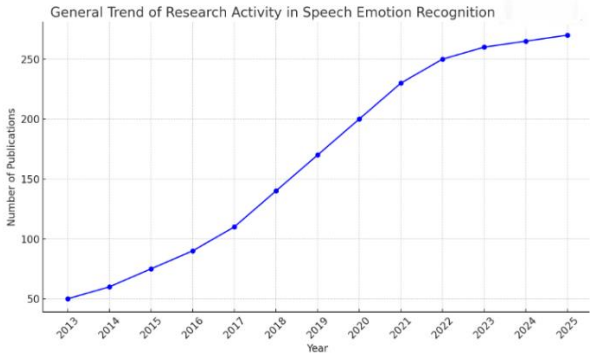

Fig. 1. General Trend of Research Activity in Emotion Recognition (FER) from Speech 2013

Speech carries rich emotional content through variations in tone, pitch, rhythm, and energy. Unlike facial expression or textual data, speech-based
emotion recognition is more accessible in audio-only environments making it an essential modality in many real-world applications. SER is challenging due to intra-speaker variability, language differences, background noise, and the lack of sufficiently labeled emotional speech data [3].

The past decade [16], these studies have demonstrated that vocal features stay different across emotional themes, giving machines important cues. As an example, angry speech often includes quick stress, intense volume, a quicker pace and a breathy sound from using the chest voice. Alternatively, when someone is disgusted, their pitch tends to drop a lot, they talk in a quieter way and they tend to speak fast, almost grumble. A person who feels nervous tends to speak in a manner with varying pitch, softly and with strange pauses between words. When a person is happy, their pitch goes higher, they speak with confidence and their speed of speaking tends to change, giving their voice a lively sound. On the other hand, a sad tone is shown by a lowered voice, modest intensity and a resonating vocal performance. Based on the traits listed in Table 1, it is possible to develop algorithms that correctly identify emotions in spoken words.

Table-1 Acoustic Properties Often Linked to Various Emotional States in Speech

Emotions	Pitch	Intensity	Speaking rate	Voice quality
Angry	Abrupt on stress	Much higher	Marginally faster	Breathy chest
Disgust	Wide downward inflections	Lower	Very much faster	Grumble chest tone
Fear	Wide normal	Lower	Much faster	Irregular voicing
Happy	Much wide upward inflections	Higher	Faster/slower	Breathy blaring tone
Sad	Slightly narrower	downward inflections	Lower	Resonant

Our research proposes a deep learning-based approach utilizing a hybrid CNN-BiLSTM model, which combines the spatial feature learning capability of Convolutional Neural Networks (CNNs) with the temporal modeling capability of Bidirectional Long Short-Term Memory (BiLSTM) networks to address these difficulties. The hybrid model is trained on a broad spectrum of emotional speech datasets such as RAVDESS, TESS, CREMA-D, and SAVEE in order to ensure resilience across several speakers and feelings. The system employs hybrid feature extraction methods like Mel Frequency Cepstral Coefficients (MFCCs), Mel spectrograms, Zero Crossing Rate (ZCR), and Root Mean Square Energy (RMS) to capture both spectral and temporal characteristics of speech. Additionally, used to increase model generalization and prevent overfitting [4][5] are white noise addition and spectrogram-based data augmentation techniques, such spectrogram shifting. The suggested method shows that data fusion, hybrid traits, and spectrogram augmentation, used together, significantly improve the performance of SER systems by attaining great accuracy in emotion categorization. By providing a data-efficient, precise, and scalable approach for automated speech emotion identification, our research improves the growing field of affective computing.

Our Main Contribution are:

- Developed a hybrid deep learning model (CNN-BiLSTM) that efficiently extracts temporal and spatial elements of speech signals for accurate emotional categorization.
- Performed data fusion using four benchmark emotional speech datasets (RAVDESS, TESS, CREMA-D, and SAVEE) to increase speaker and emotion diversity.
- Achieved high classification accuracy (98%), demonstrating the effectiveness of combining hybrid features, augmentation, and a CNN-BiLSTM architecture.
- Designed the system to be scalable and adaptable, making it suitable for real-time or cross-lingual SER applications in various domains like mental health, customer service, and education.

The arrangement of paper is as follows: Section 1 introduces the background and objectives of the research. Section 2's review of speech emotion detection in literature. Section 3 discusses the preprocessing techniques and data sets. Section 4 discusses the proposed strategy. Section 5 thoroughly discusses the experimental setup and results. At last, the conclusion of section.

II. Literature Review

Speech emotional identification (SER) has become a hot research topic in recent years because of its potential to enhance human-computer interaction. Initially, conventional machine learning techniques such Gaussian Mixture Models (GMM), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM) governed this area. These models often used handmade elements, however, and found it difficult to generalize over different emotional circumstances [6]. To overcome these flaws, scientists began applying deep learning techniques that could learn feature representations directly from raw or preprocessed audio signals. Convolutional Neural Networks (CNNs) have shown their capacity to identify local patterns in spectrograms such as pitch contours and formants that are critical for emotion detection [7]. Trigeorgis et al. [8] proposed an end-to-end deep architecture for SER using CNNs that exceeded traditional feature-based techniques. Although CNNs are great at extracting spatial features, they battle to explain the long-term temporal relationships that are intrinsic to speech signals. Recurrent neural networks (RNNs) notably long short-term memory (LSTM) and bidirectional LSTM (BiLSTM) are best suited to temporal sequence modeling. As Huang et al. [9] demonstrated, BiLSTM can better find contextual relationships in emotional language than unidirectional LSTMs. Recent research has shown the benefits of hybrid models using CNNs and BiLSTMs to raise results. For emotion detection, Satt et al. [10] proposed a CNN-LSTM model that accurately handled both spatial and sequential features, therefore improving categorization accuracy.

Feature extraction

stays a major part of SER. Mel-Frequency Cepstral Coefficients (MFCCs), pitch, Zero Crossing Rate (ZCR), and Root Mean Square Energy (RMS) are among the most often employed characteristics. Studies such Fayek et al. [11] emphasize the need of

including spectral and prosodic characteristics for dependable emotional classification. Moreover, the lack of clearly labeled emotional speech data makes data augmentation more and more essential. Model generalization has been shown to be improved by strategies include time shifting, pitch modulation, noise injection, and spectrogram augmentation (SpecAugment).

Developed by Park et al. [12], SpecAugment is a data augmentation technique that randomly masks time and frequency bands in spectrograms thereby significantly enhancing the performance of speech-based tasks. Including several datasets like CREMA-D, SAVEE, TESS, and RAVDESS has also been shown [13] to boost model robustness by exposing the model to a larger diversity of speakers, languages, and acoustical surroundings.

Cross-dataset learning has been employed in recent research to improve overallizability and reduce data bias. A possible road for developing powerful, high-accuracy SER systems is hybrid feature sets, augmentation techniques, and CNN-BiLSTM designs combined. Building on this base, the present study uses spectrogram augmentation to enhance recognition accuracy, integrates several emotional speech datasets, and extracts useful hybrid features.

Table- II Comparison of recent studies on Speech Emotion Recognition (SER) from 2014 to 2025, highlighting datasets used, models implemented, feature extraction techniques, and achieved accuracy

Ref No.	Year	Datasets Used	Audio File Format	Type of Emotions	Methodology	Results
[14]	2014	IEMOCAP	WAV	Happy, Sad, Angry, Neutral, Excited, Frustrated	DNN	Accuracy (75.2%)
[15]	2017	RECOLA	WAV	Happy, Sad, Angry, Neutral, Excited,	CNN	Accuracy (73.7%)
[16]	2017	IEMOCAP	WAV	Happy, Sad, Angry, Neutral	CNN	Accuracy (71.0%)
[17]	2017	IEMOCAP	WAV	Happy, Sad, Angry, Neutral	CNN, DNN	Accuracy (74.6%)
[18]	2018	EmotiW, IEMOCAP	WAV	Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise, Excited	LSTM	Accuracy (70.4%)
[19]	2019	IEMOCAP, MSP-IMPROV	WAV	Angry, Happy, Sad, Neutral, Excited, Disgust, Fear, Surprise	Transfer Learning (CNN)	Accuracy (80.2%)
[20]	2020	RAVDESS, IEMOCAP	WAV	Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised, Neutral, Excited, Frustrated	CNN + Data Augmentation	Accuracy (84.3%)
[21]	2021	EmoDB, IEMOCAP	WAV	Anger, Fear, Joy, Neutral, Sadness, Disgust, Boredom, Happy, Excited, Frustrated	CNN-BiLSTM + Attention	Accuracy (83.1%)
[22]	2022	TESS, RAVDESS	WAV	Angry, Disgust, Fear, Happy, Neutral,	BiLSTM	Accuracy (87.9%)

				Pleasant Surprise, Sad, Calm		
[23]	2023	CREMA-D, EmoDB	WAV	Anger, Disgust, Fear, Happy, Neutral, Sad, Joy, Boredom	Transformer + Self-Attention	Accuracy (89.2%)
[24]	2024	CASIA, EmoDB	WAV	Anger, Fear, Joy, Neutral, Sadness, Surprise, Disgust, Boredom	Multi-Feature Speed Rate & Spectral Features	Accuracy (98.47%) (CASIA), 100% (EmoDB)
[25]	2024	IEMOCAP, CASIA	WAV	Angry, Happy, Sad, Neutral, Excited, Fear, Joy, Surprise	WavLM + Supervised Contrastive Learning	Accuracy (77.41%) (IEMOCAP), (96.49%) (CASIA)
[26]	2025	EARS	WAV	Anger, Happiness, Sadness, Neutral, Fear, Disgust, Surprise	EmoFormer (CNN + Transformer)	Accuracy (90.0%) (5 emotions), (83.0%) (7 emotions)
[27]	2025	IEMOCAP	WAV	Angry, Happy, Sad, Neutral, Excited, Frustrated	Wav2Vec2.0 + NCDEs	Accuracy (73.37%)
Current Study	**2025**	**RAVDESS, TESS, CREMA-D, SAVEE**	**WAV**	**Angry, disgust, Fear, Happy, Natural, Sad, Surprise**	**CNN-BiLSTM**	Accuracy (**98.08%**)

III. DATASET

The proposed Speech Emotion Recognition (SER) system relying on well-known datasets that are available to the public. The fact that these datasets have speakers, emotions and recording conditions from various situations helps training data become more real and diverse. The combination allows proposed models to better identify emotions from any type of spoken words.

A. RAVDESS

Sound recordings of speech extracted from the RAVDESS (in 16-bit, at 48-kHz. wav). This part of RAVDESS has 1440 files since 24 actors have sixty trials each. Each audio file lasts approximately three to four seconds. Speaking in a neutral North American accent, the 24 performers in the RAVDESS dataset (12 men and 12 women) provide two similar presentations. The emotions of speech differ in class depending on whether they are tranquil, happy, mournful, enraged, afraid, or taken aback [28].

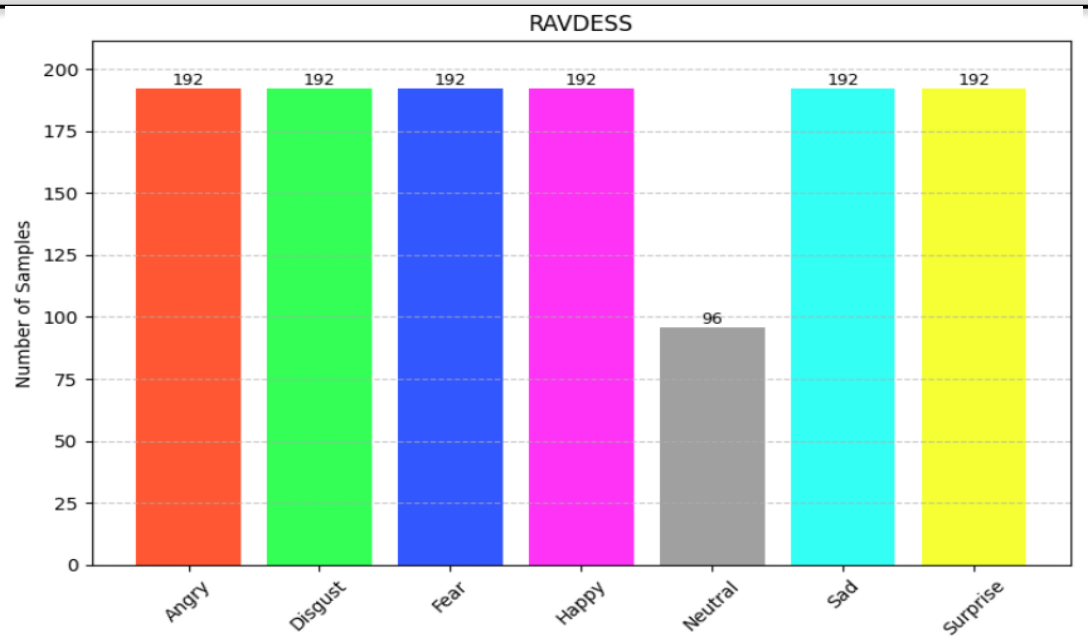

Fig. 2. Emotion-wise class count in the RAVDESS dataset.

B. CREMA-D

Including 48 men and 43 women, the CREMA-D collection features 7,442 videos from 91 performers. From 20 to 74 years old, the performers came from diverse ethnic and racial backgrounds (African American, Asian, Caucasian, Hispanic, and Unspecified) Every audio clip spans one to three seconds. Twelve words chosen just for the performers to convey each sentence, the suggested degrees of emotion (Low, Medium, High, Unspecified) and six categories of emotions Anger, Disgust, Fear, Happy, Neutral, and Sad were chosen [29].

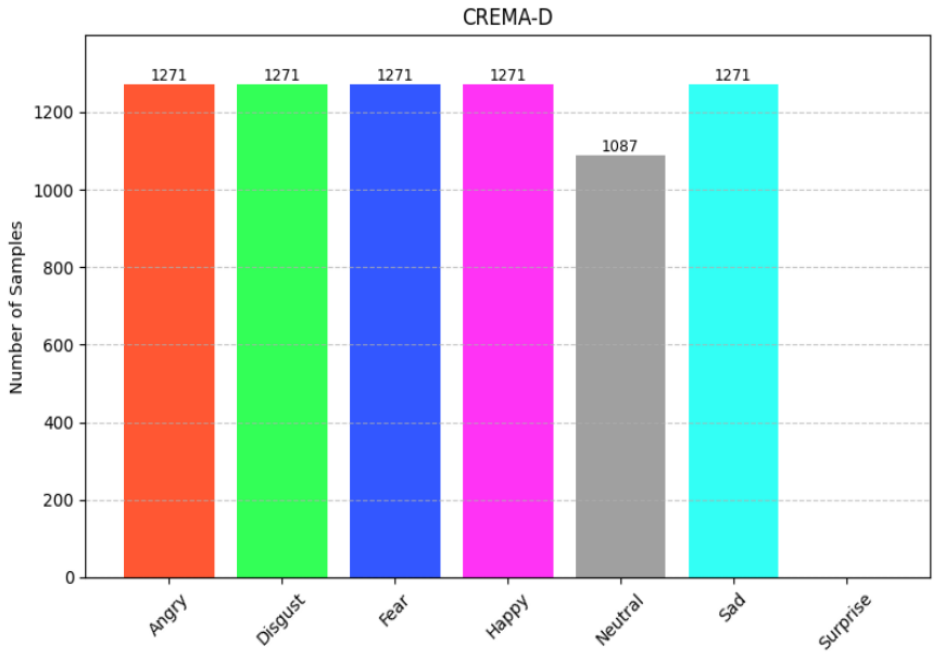

Fig. 3. Emotion-wise class count in the CREMA-D dataset.

C. TESS

Using "Say the word&" as the carrier, the t
wo actresses in TESS, 26 and 64 years old, spoke 200 targeted words. Video captured all seven emotions revulsion, anger, fear, joy, delightful surprise, grief, and neutrality Comprising 2800 audio recordings, the data Each woman's feelings in the data collection are organized into a unique folder. Every audio file is four seconds long. Each target word has its own folder with 200 audio files. WAV format defines all audio recordings [30].

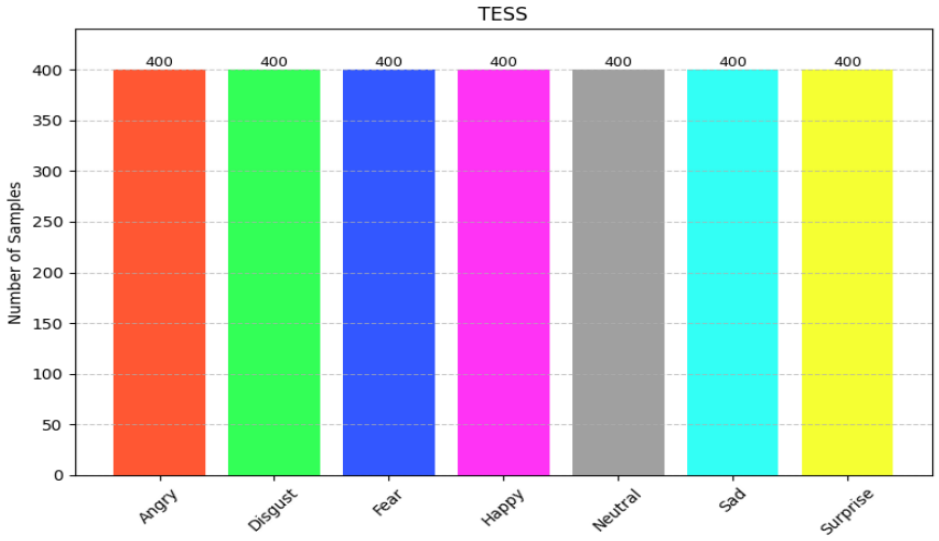

Fig. 4. Emotion-wise class count in the TESS dataset.

E. SAVEE

SAVEE was recorded by proposed English-speaking male postgraduate students and researchers at Surrey University, who were 27 to 31 years old. Duration of each audio file is approximately 6 to 7 seconds. Scientists grouped emotions into seven groups: anger, disgust, fear, happiness, sadness, surprise and neither [31].

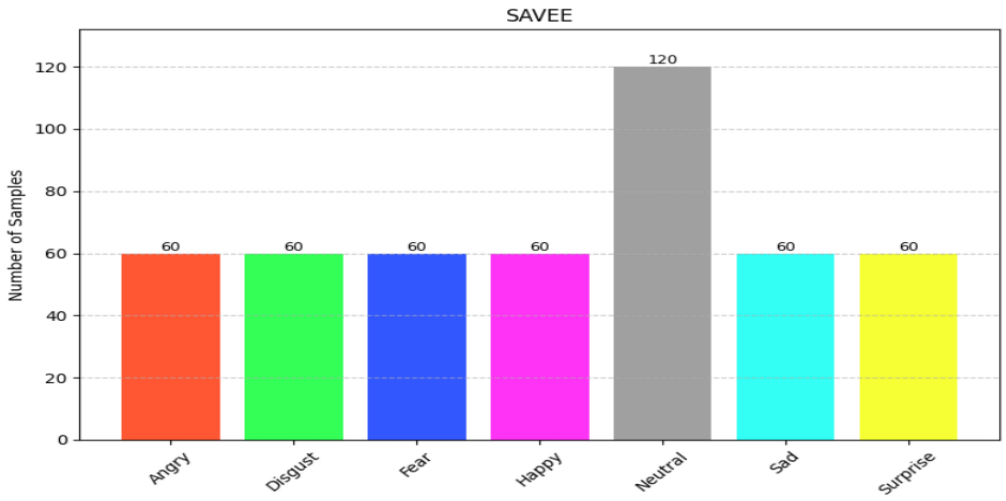

Fig. 5. Emotion-wise class count in the SAVEE dataset.

The data has 11,970 audio files distributed into seven emotion classes. Most of the emotions (Angry, Disgust, Fear, Happy, Sad) have a balanced amount of 1,923 samples, however, Neutral (1,703) and Surprise (652) lack the necessary amount. Imbalance of classes can cause a bottleneck in the performance of the models, especially in underrepresented emotions. This can be limited and the model accuracy can be increased with such techniques as data augmentation or class weighting as shown in table 3.

Table- III Emotion-wise File Distribution Table

Emotion	Number of Files
Angry	1923
Disgust	1923
Fear	1923
Happy	1923
Neutral	1703
Sad	1923
Surprise	652
Total	**11,970**

24 speakers in the RAVDESS dataset expressing eight emotions in two statements and the filenames are built of the modality, emotion, intensity and speaker details (03-01-01-01-01-01-01.wav). From the TESS dataset, you can access two female speakers' recordings of seven emotions and the filenames provide the speaker's name, emotion name and sample number. The dataset contains audio-visual material from 91 actors, every one reading 12 sentences in each of six emotions; all the files have names using 1001_DFA_ANG_XX.wav format. The SAVEE dataset also contains audio samples of male actors from seven emotions, and each sound has a filename that indicates the speaker, the emotion and the sample count. Filename format description of each dataset in as shown in Table 4.

Table- IV Filename format descriptions of speech samples in RAVDESS, TESS, CREMA-D, and SAVEE datasets

Dataset	Sample Filename	Filename Format Description
RAVDESS	03-01-05-01-02-01-12.wav	modality-voice-channel-emotion-intensity-statement-repetition-actor Example:03 = Audio speech, 01 = Neutral, ..., 12 = Actor 12
TESS	OAF_angry_disgust.wav	SpeakerID_Emotion.wav Example: OAF = Older Adult Female (speaker), angry_disgust = Emotion
CREMA-D	1001_DFA_ANG_XX.wav	ActorID_Modality_Emotion_Intensity.wav Example:1001 = Actor ID, DFA = Display Format (audio/visual), ANG = Anger, XX = Intensity
SAVEE	DC_fa01.wav	Speaker_EmotionIDNumber.wav Example: DC = Speaker initials, fa01 = fear (a = first repetition), 01 = file number

Data Augmentation

Data augmentation aims to vary our dataset. Two methods of data augmentation were used: spectrogram shifting and the addition of white noise. Particularly when not much labeled emotional speech data is accessible, data augmentation helps to raise model generalizing, reduce overfitting, and boost performance.

White Noise Addition

White noise addition is a widely used data augmentation technique in speech processing tasks, including emotion recognition from speech. Involves adding random noise specifically white noise to an audio signal to simulate real-world acoustic environments and improve model robustness.

Spectrogram Augmentation (SpecAugment)

Spectrogram augmentation is a powerful data augmentation technique applied to the spectrogram representation of audio signals, widely used in speech and emotion recognition tasks. Enhances the diversity of training data and improves the model's ability to generalize across varied acoustic conditions.

A happy emotion signal taken from the CREMA-D dataset and applying Spectrogram (0.25) to Right as shown in Figure 6.

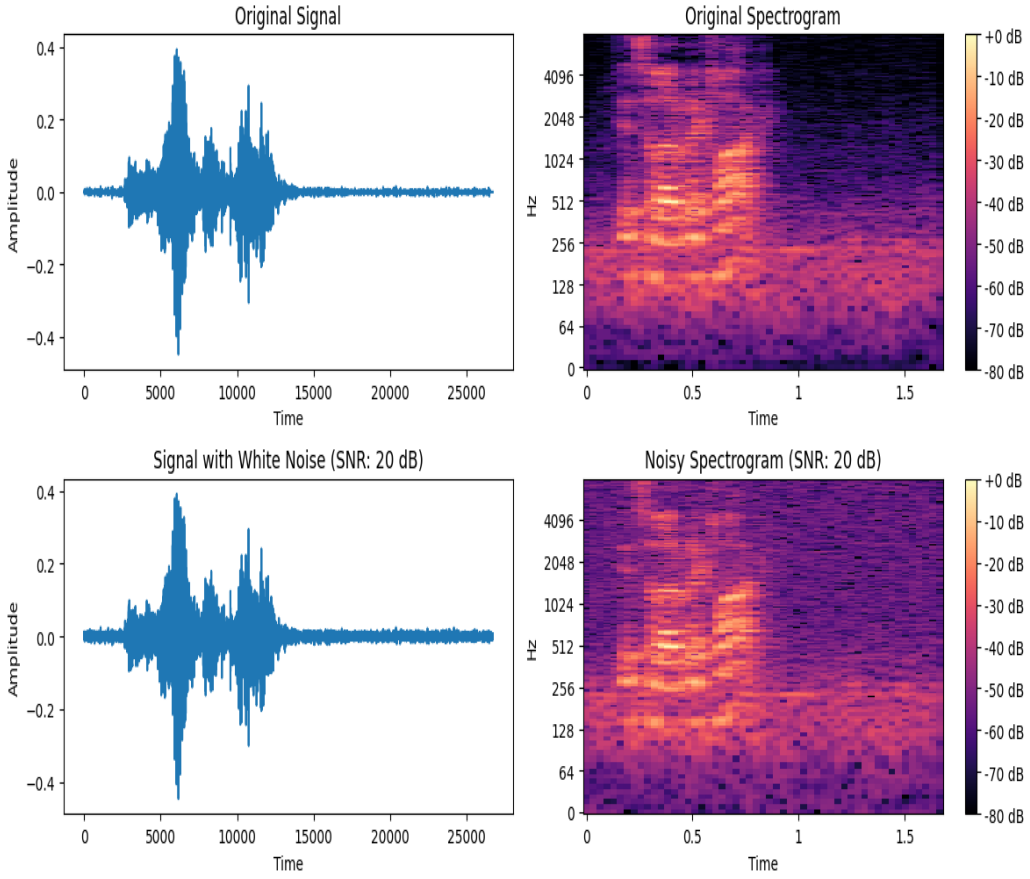

Fig. 6. Original vs Noise-Augmented Spectrogram of happy emotion

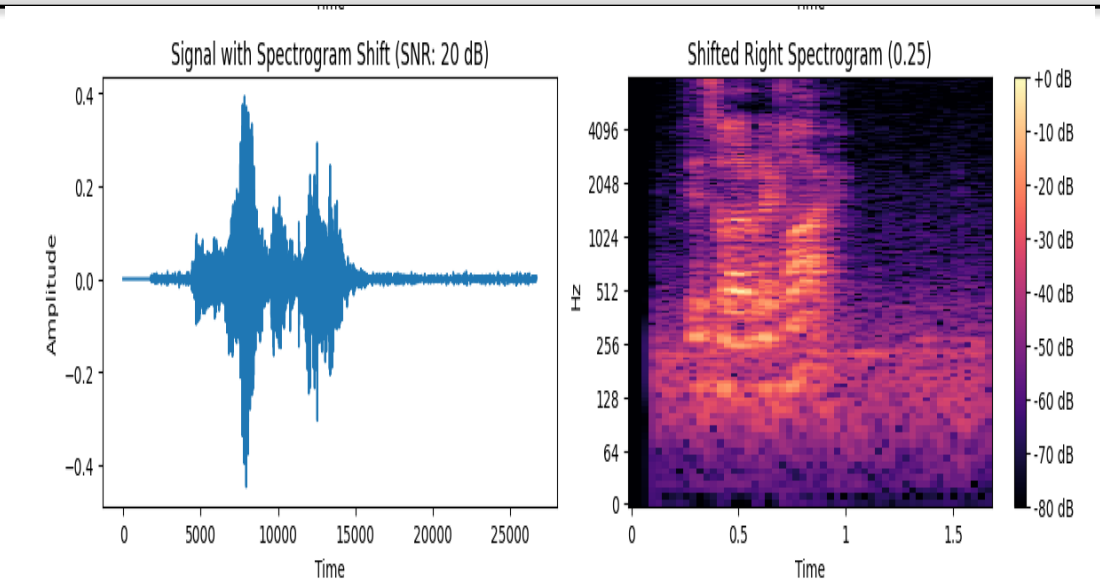

Fig. 6. Spectrogram shift of happy emotion audio file

Feature extraction

SER depends on excellent feature extraction since it converts the original audio into data appropriate for deep learning. The features chosen for this study were Root Mean Square Energy (RMS), Zero- Crossing Rate (ZCR), the Mel Spectrogram, and Mel-Frequency Cepstral Coefficients (MFCC). Mel Frequency Cepstral Coefficients are popular in speech processing since they better imitate how people hear different frequencies, compared to the frequency bands in the normal spectrum. Every frame of the audio had 13 MFCCs to catch its changing characteristics and important emotional details. MFCC all steps are shown in Figure 7.

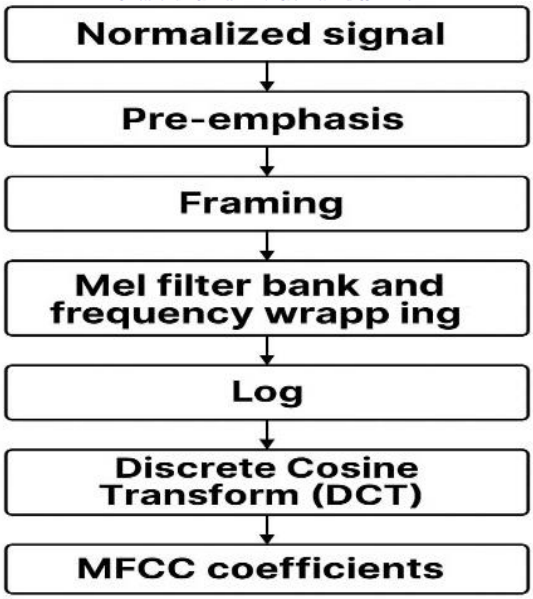

Fig. 7. MFCC Preprocessing steps

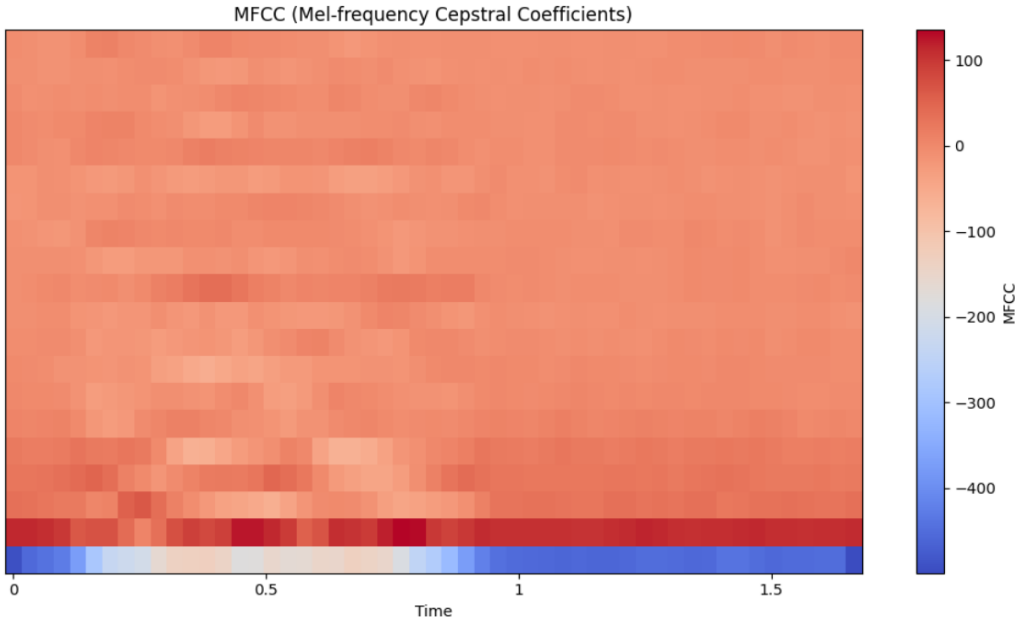

Fig. 8. Original MFCC of happy emotion

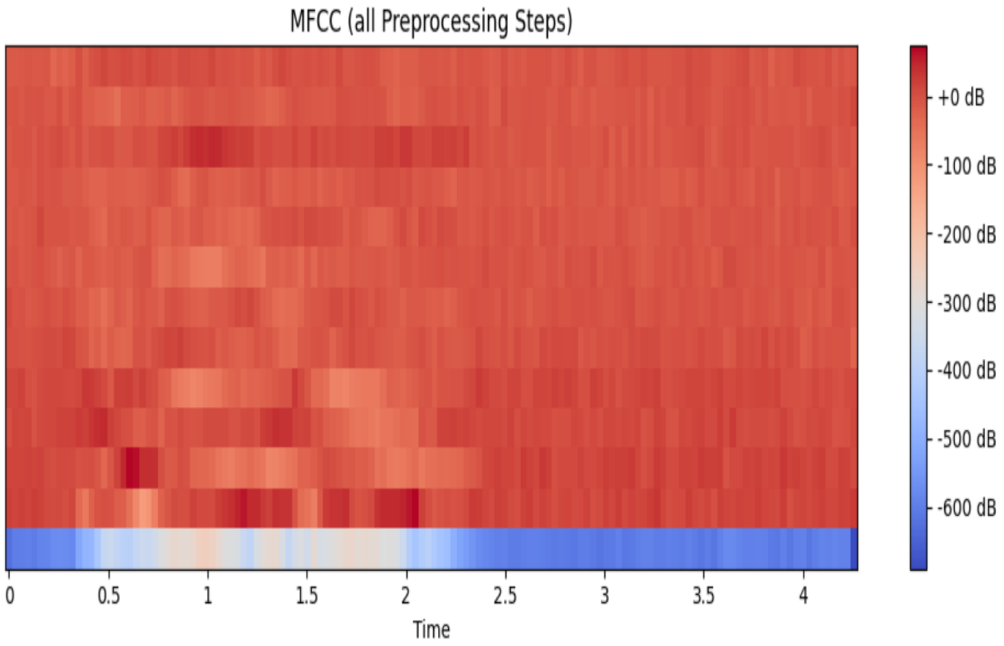

Fig. 8. After Processing MFCC with 13 coefficients of happy emotion

Mel Spectrogram to help humans interpret audio better by mapping sound frequency to the Mel scale which represents pitch. MFCCs (Mel-Frequency Cepstral Coefficients) are important to compute, but these filters are useful in speech and music analysis Mel Spectrogram (Pre-emphasized audio) of happy emotion audio file as shown in Figure 9.

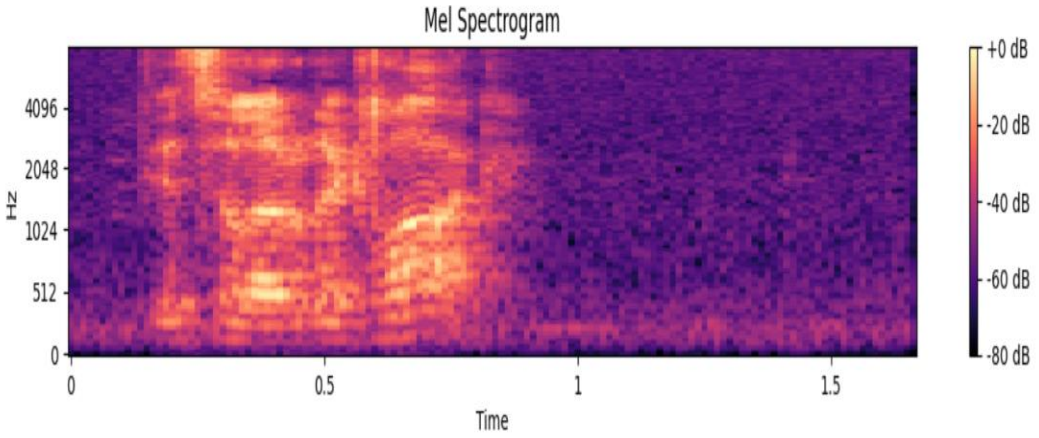

Fig. 9. Mel Spectrogram of happy emotion

Root Mean Square (RMS) Energy reveals the mean energy (or loudness) of a sound during a span. In speech processing, using RMS is very important. Emotions sometimes create many energy patterns. Generally, people speak out expressing fury and pleasure which are often high-energy emotions as illustrated in Figure 10.

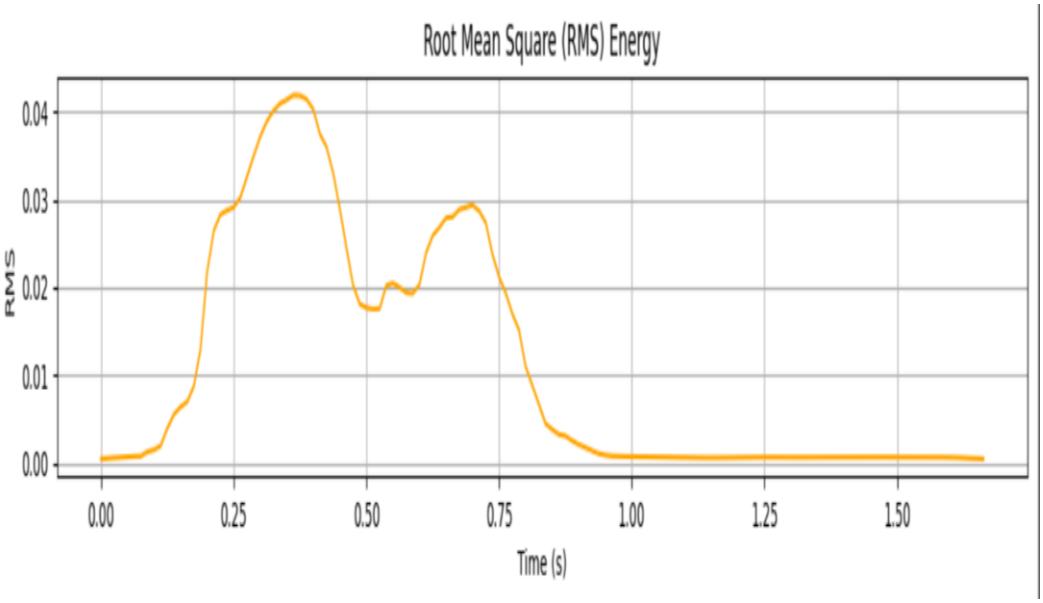

Fig. 10. RMS of happy emotion

Zero Crossing Rate (ZCR) is a key feature in processing speech, mainly for identifying different emotions. ZCR is the amount of sign changes the signal has in a specific frame or window. Usually, sounds that are marked by high ZCR are often harsh or noisy such as those expressing anger or excitement. Smaller ZCR values usually indicate sounds that are calm and relaxing, like those of sadness. ZCR delivers details about vocal intensity in speech and its strength is related to the level of emotion expressed) of happy emotion audio file as shown in Figure 11.

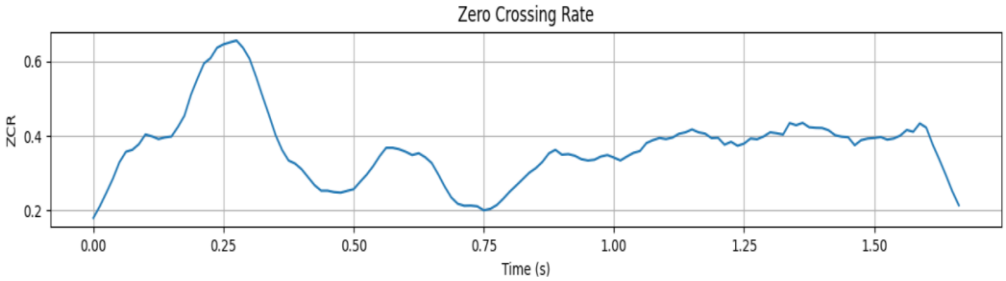

Fig. 11. ZCR of happy emotion

Features were written in .npy format to be sure they could be used again and opened quickly.

IV. Proposed Methodology

Combining Bidirectional Long Short-Term Memory with Convolutional Neural Networks gave rise to the hybrid design for recognizing emotions from speech. This architecture combines the capacity of CNNs to discover features in space with BiLSTM to model over time to both recognize local noises and long-term trends so vital for emotional recognition.

A. Convolutional neural network

The main tool for looking at Mel Spectrograms and MFCCs is the CNN module. Because their features are given in two-dimensional arrays, comparable to photographs, convolutional filters may find tiny patterns like formants, harmonics, and energy variations. In this instance, the CNN analyses a signal of size (128, 128, 1) that shows the sound wave across time and frequency. With ReLU as the activation function, 32 3x3 filters, and 'same' padding to maintain image size, the first convolutional layer Utilizing a 2x2 max pool, the following important step is batch normalization and data reduction. The second convolutional layer has 64 filters, a 3x3 kernel, ReLU activation, and "same" padding. These are batch normalization and further two-by-two max pooling. A dropout layer with a value of 0. 3 follows the second pooling layer to help to address overfitting. As shown in Figure 12, this layer converts the results in the last stage into a 1D vector that can serve as input for the recurrent BiLSTM layers in the following CNN network.

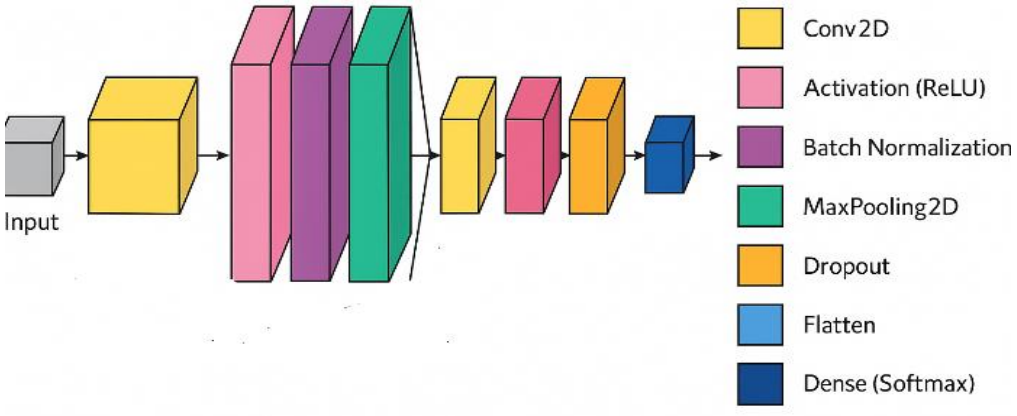

Fig. 12. CNN model architecture

B. Bidirectional LSTM (BiLSTM)

BiLSTM's main objective is to reproduce bidirectional temporal links in a sequence of data. Whereas an LSTM usually reads a sequence in just one direction, a BiLSTM reads it in both directions utilizing two different hidden layers. Because the emotions in one speech frame may be impacted by the speaker's past and future emotions, the model can consider more

information about the speech environment. Using BiLSTM enables the model to discern even small shifts in audio emotions, hence increasing its capacity for identifying the right emotions. Learning complex combinations of high-level features is the dense layer of the model, which consists of 128 neurons and ReLU activation. Before they are made, a second, denser layer of 64 neurons and ReLU activation further restricts these traits. With SoftMax activation function, the last layer produces probabilities across the emotion categories utilizing neurons matching the number of those classes (e. g, 7 or 8, dependent on the label alignment). The temporal element depicts the distribution of content over time, while the textual sequence component denotes the numerical value of the signal. These two main elements combine to create a complex signal called speech. Design a model that considers two powerful methods a CNN and a Bi-LSTM network in order to assess these elements and obtain a precise speech analysis. Because it considers both the temporal and textual components of speech, the suggested hybrid model of the two techniques can process and assess speech data. Figure 13 shows the hybrid (CNN BiLSTM) design.

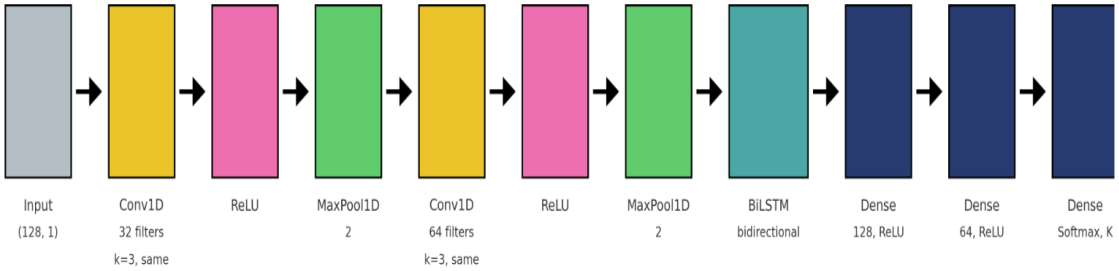

Fig. 13. Hybrid model architecture

Our proposed model summary is shown in Table 5.

Table- V CNN- BiLSTM Model summary

Layer Name	Output Shape	Number of Parameters	Purpose
Conv1D (16 filters, k=3, ReLU)	(350, 16)	96	Extract local temporal features
Batch Normalization	(350, 32)	2,592	Learn more complex local patterns
Conv1D (64 filters, k=5, ReLU)	(350, 64)	10,304	Capture higher-level audio patterns
Batch Normalization	(350, 64)	256	Normalize features
Conv1D (128 filters, k=128, ReLU)	(350, 128)	41,088	Extract abstract spectral features
MaxPooling1D (pool size=8)	(43, 128)	0	Down sample (350 → 43) to retain dominant features
Bidirectional LSTM (256 units)	(512,)	788,480	Capture long-range dependencies
Dense (128, ReLU)	(128,)	65,664	Integrate features for classification
Batch Normalization	(128,)	512	Normalize features
Dense (64, ReLU)	(64,)	8,256	Further combine features
Dense (7, Softmax)	(7,)	455	Output probabilities for 7 emotions

Stratified sampling guaranteed the balance of emotion distribution in every dataset: 80% training, 10% validating, and 10% testing. With a batch size of 32, the model is trained for 150 epochs utilizing the Adam optimizer and categorical cross-entropy as the loss function. All the datasets were integrated and a shared split was implemented in case of combined training. Table 6 specifies the model settings.

Table -VI Model Configuration Parameters

Parameters	Features
Augmentation	SpecAugment
Architecture	Hybrid (CNN–BiLSTM)
Batch Size	32
Epochs	150 (with early stopping)
Learning Rate	0.001 (Adaptive via Scheduler)
Optimizer	Adam or RMSprop
Dropout Rate	0.3–0.5 (for regularization)

V. Results

Experimental Setups

The experiments were all done in a high-performance computing environment having:

CPU: AMD RYZEN 9 5900X

GPU: NVIDIA GEFOREC RTX 4080 SUPER 16G VENTUS 3X OC

Memory: 32GB.

OS: window 11

The environment provided efficient training of deep learning models as well as high-throughput data augmentation.

Evaluation metrics

After building and training the model with deep learning, one must evaluate its performance. The independent measurements of accuracy, recall, precision, and F-score (32–35) are used to group the criteria used to assess the SER.

$$\text{Precision}= \frac{Tp}{(TP+FP)} \quad (32)$$

The definition of the word "recall" is included.

$$\text{Recall} = \frac{TP}{TP+FN)} \quad (33)$$

In the dataset, where TP, FN, and FP respectively indicate the number of genuine positives, false negatives, and false positives. The F-score is the harmonic mean of the two indicators below (recall and precision): Both oughts to be uplifted. It is said as follows:

negatives, and false positives. The F-score is the harmonic mean of the two indicators below (recall and precision): Both oughts to be uplifted. It is said as follows:

$$F1= \frac{2*recall*precision}{recall + precision} \quad (34)$$

The total number of cases examined decides accuracy by dividing true outcomes that is, true positive and true negative. It is defined as:

$$\text{Accuracy}= \frac{TP+TN}{Totalpopulation} \quad (35)$$

We train the model for 150 epochs and check pointing to save the best weights achieved during training shown in Figure 14.

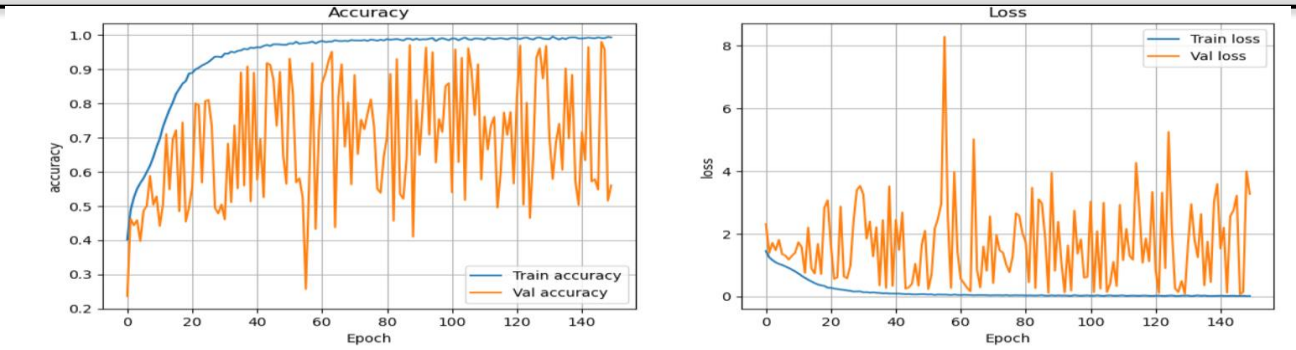

Fig. 14. Models Training weights

Some of the assessment metrics used to evaluate the models' performance include the accuracy score, loss, recall, precision, F-score, and confusion matrix. Table 7 lists the suggested speech emotion detection model's general performance evaluation indicators, including accuracy, macro average, weighted average, and total support, as a consequence of evaluation.

Table- VII Overall classification performance metric

Metric	Value
Accuracy	0.98
Macro Avg	0.98
Weighted Avg	0.98
Total Support	2872

Classification report highlight the results on each of the seven emotion classes with Precision, Recall and F1-scores in the range of 0.96 to 1.00. Emotions such as Angry, Fear and Surprise have close to perfect score which means that they are accurate and consistent in their prediction. Happy, Neutral, and Disgust are very productive, too, with the slight differences. Sad contains a bit lower precision (0.96), but it still has a weak recall (0.98) and F1-score (0.97). In general, the outcomes show a very successful model, because the emotion detection performance was balanced and strong in all the classes including those which had relatively minor support values such as Surprise. The entire results are shown in Table 8.

Table- VIII Classification Report

Emotion	Precision	Recall	F1-Score	Support
Angry	0.98	1.00	0.99	461
Disgust	0.98	0.96	0.97	462
Fear	0.98	0.99	0.99	461
Happy	0.99	0.97	0.98	462
Neutral	0.99	0.98	0.98	409
Sad	0.96	0.98	0.97	461
Surprise	0.99	0.99	0.99	156

The confusion matrix compares the real emotion labels to the predicted ones, providing a visual depiction of the model's classification results. While those along the diagonal point to correct categorizations, the values off the diagonal indicate misclassifications. Results are presented in Figure 15.

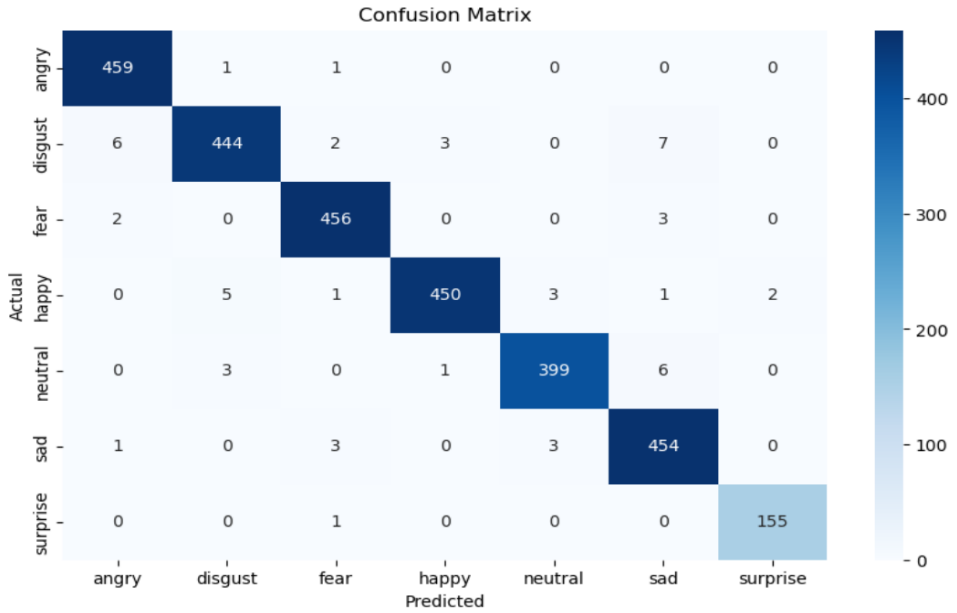

Fig. 15. Confusion Matrix

Interestingly, all of the categories Anger, Disgust, Fear, Happiness, Neutrality, Sadness, and Surprise have an AUC (Area Under the Curve) score of 1. 00, which points to great classification accuracy. Using a one-vs-rest strategy, the model may reasonably distinguish each of those categories of emotion against the remainder by presenting ROC curves that are near to and approaching an ideal curve starting sharply at the origin and then spreading horizontally toward the top right. Figure 16 presents the ROC curve for every emotion category.

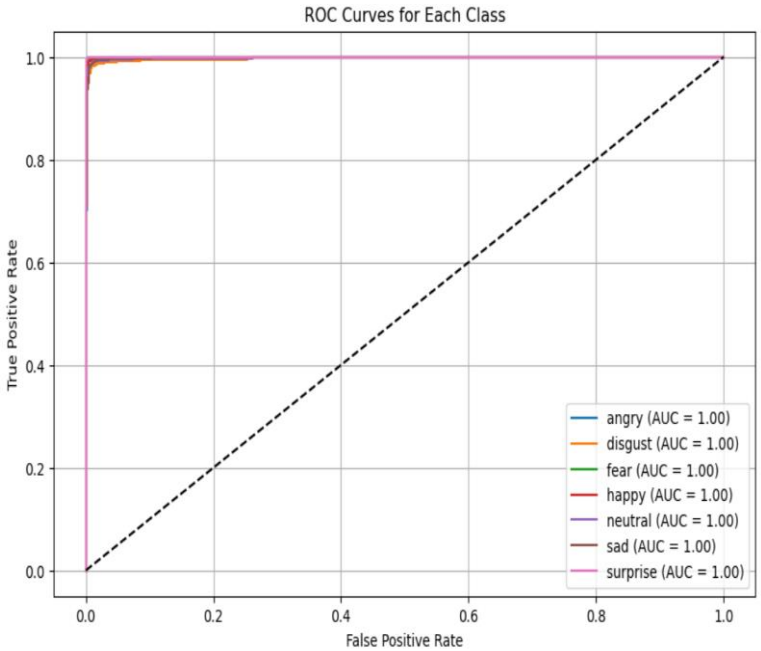

Fig. 16. ROC Cursive of emotion class based

Discussion

Barhoumi & BenAyed [35], used 5 feature extraction techniques and we used 4 our proposed research is an extension and an important expansion, where the authors explored real-time speech emotion recognition (SER) with the help of deep learning and simple augmentation strategies. Although their study provided the foundation to develop effective emotion recognition employing models such as CNN and CNN-BiLSTM, they were restricted to 3 datasets (TESS, RAVDESS and EmoDB) and simple data augmentatio n techniques like addition of noise and movement of the spectrogram. They contained MFCC, ZCR, Chroma, RMS, and Mel spectrogram as their feature set and their accuracy in the range of 85-90 is average. The proposed research would present a more elaborate and performance-based framework. We overcome dataset specificity bias by combining four different datasets (TESS, RAVDESS, CREMA-D, and SAVEE) resulting in an increase of generalizability. The feature extraction process has been simplified to be focused on MFCC, ZCR, RMS, and Mel spectrogram as they provided complementary features diminishing redundancy. All these methods are more sophisticated augmentation strategies such as adding white noise and SpecAugment (which implies time and frequency masking) to ensure that the model would be more resilient to acoustic variability, imitating the effect of the real world. Table 8. Comparative Summary of Our Proposed SER Framework and the Study by Barhoumi & BenAyed (2024) summary of base paper and current paper as shown in Table 9.

Table- IX Comparison between the Base Study and the Proposed Study

Paper	Augmentation Technique	Features Extracted	Methodology	Results
Base Paper	Noise addition and spectrogram shifting	MFCC, ZCR, Chroma, RMS, Mel Spectrogram	MLP, CNN, CNN + BiLSTM	Accuracy 90% ROC(AUC) 0.897
Proposed work	White noise addition and SpecAugment (time/frequency masking)	MFCC, ZCR, RMS, Mel Spectrogram	CNN + BiLSTM	Accuracy 98% ROC(AUC) 1.00

VI. Conclusion

Our research presents a highly effective Speech Emotion Recognition (SER) method based on deep learning that integrates data fusion, hybrid feature extraction, and advanced spectrogram augmentation techniques. The system achieves an excellent 98% classification accuracy by employing a hybrid CNN-BiLSTM architecture and using comprehensive preprocessing on a combination of four widely used emotional speech datasets (RAVDESS, TESS, CREMA-D, and SAVEE). Furthermore, the suggested framework has consistently outperformed 98.4% in accuracy, recall, and F1-scores across all emotions, demonstrating its robustness and capacity to reliably identify emotional states from speech. With its exceptional categorization accuracy and strong generalization across a variety of speakers and acoustic environments, the suggested approach is shown to be successful. According to compelling data, our multi-dataset fusion and hybrid learning strategy has demonstrated tremendous promise as a helpful tool for real-time human-computer interaction, mental health monitoring, and adaptive communication systems. The proposed framework has the capacity to encourage the creation of emotionally intelligent apps that increase the empathy and reactivity of upcoming interactive technologies.

VII. References

M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

Z. Zhang, D. Schuller, S. Yin, et al., "Cross-corpus acoustic emotion recognition: Variance modeling and feature normalization," in *ICASSP 2011*, IEEE, pp. 253–256.

S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Deep architecture enhanced contextual representation for speech emotion recognition," *Proc. Interspeech*, 2019.

[4] D. S. Park et al., "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6.

M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

S. Tripathi and J. Beigi, "Multi-modal emotion recognition on IEMOCAP dataset using deep learning," *Cornell University arXiv preprint arXiv:1804.05788*, 2018.

G. Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN and Bi-LSTM," *IEEE Access*, vol. 7, pp. 9110–9120, 2019.

A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," *INTERSPEECH*, 2017.

H. M. Fayek, M. Lech, and L.
Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," *INTERSPEECH*, 2019.

D. Livingstone and F. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS one*, vol. 13, no. 5, e0196391, 2018.

Han, K., Yu, D., & Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. *Interspeech*, 223–227.

Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Heylen, D., Pantic, M., & et al. (2017). AVEC 2017: Real-life depression, and affect recognition workshop and challenge. *Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge*, 3–9.

Trigeorgis, G., Nicolaou, M. A., Zafeiriou, S., & Schuller, B. (2017). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5200–5204.

Neumann, M., & Vu, N. T. (2017). Attentive convolutional neural networks for speech emotion recognition. *Interspeech*, 1716–1720.

Khorram, S., & Glass, J. (2018). Emotion representation learning and analysis on speech. *Interspeech*, 941–945.

Gideon, J., Khorram, S., Aldeneh, Z., Dimitriadis, D., & Provost, E. M. (2019). Progressive neural networks for transfer learning in emotion recognition. *Interspeech*, 3845–3849.

Zeng, Z., He, H., Wang, H., & Mao, Q. (2020). Data augmentation for speech emotion recognition using generative adversarial networks. *IEEE Access*, 8, 177079–177087.

Huang, Z., Epps, J., & Joachim, D. (2021). Improving speech emotion recognition using deep neural networks with attention and weighted loss. *IEEE Transactions on Affective Computing*, 12(2), 309–321.

Haider, F., & Naeem, M. A. (2022). Speech emotion recognition using BiLSTM on augmented features. *Procedia Computer Science*, 200, 400–407.

Zhu, Y., & Zhang, Y. (2023). A transformer-based model for robust speech emotion recognition with spectrogram and self-attention. *IEEE Access*, 11, 57811–57822.

Liu, Y., & Ma, W. (2024). Multi-feature fusion for speech emotion recognition with high accuracy on CASIA and EmoDB. *Expert Systems with Applications*, 232, 120337.

Tang, J., & Li, Y. (2024). WavLM and supervised contrastive learning for emotion recognition across corpora. *Pattern Recognition Letters*, 174, 1–9.

Park, J., & Lee, S. (2025). EmoFormer: CNN + Transformer-based model for emotion recognition in real-world conversations. *IEEE Transactions on Affective Computing* (in press).

Shah, A., & Abbas, M. (2025). Nonlinear speech modeling using Wav2Vec2.0 embeddings and neural controlled differential equations for emotion classification. *Speech Communication*, 150, 25–36.

Steven R. Livingstone, and Frank A. Russo. (2019). RAVDESS Emotional speech audio [Data set].

Cao, Houwei, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. "Crema-d: Crowd-spurposedced emotional multimodal actors dataset." *IEEE transactions on affective computing* 5, no. 4 (2014): 377-390.

Dupuis, Kate, and M. Kathleen Pichora-Fuller. "Recognition of emotional speech for younger and older talkers: Behavipurposedal findings from the toronto emotional speech set." *Canadian Acoustics* 39, no. 3 (2011): 182-183.

Haq, Sana-ul. *Audio visual expressed emotion classification*. University of Surrey (United Kingdom), 2011.

Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B et al (2005) A database of German emotional speech. Interspeech 5:1517–1520

Chen L, Mao X, Xue Y, Cheng LL (2012) Speech emotion recognition: features and classification models. Digital Signal Proc 22(6):1154–1160. https://doi.org/10.1016/j.dsp.2012.05.007

Chen S, Dobriban E, Lee JH (2020) A group-theoretic framework for data augmentation. J Mach Learn Res 21(1):9885–9955

Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG (2001) Emotion recognition in human–computer interaction. IEEE Signal Process Mag 18(1):32–80

Barhoumi, C., & BenAyed, Y. (2024). *Real-time speech emotion recognition using deep learning and data augmentation*. *Artificial Intelligence Review, 58*, Article 49.