ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

AUDIO-BASED EMOTION RECOGNITION IN SPEECH USING DEEP LEARNING AND FEATURE ENGINEERING TECHNIQUES

Corresponding Author
Ramakrishna Gandi
Research Scholar, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University, Annamalai Nagar, Tamil Nādu, INDIA.
gandiramakrishna2@gmail.com | ORCID ID: 0000-0001-7213-5789
https://orcid.org/0000-0001-7213-5789

Dr.A.Geetha
Professor, Computer Science & Engineering Department,
Faculty of Engineering & Technology, Annamalai University,
Annamalai Nagar, Tamil Nādu, INDIA.
aucsegeetha@yahoo.com

Dr.B.Ramasubba Reddy
Professor, Department of Computer Science and Engineering,
School of Computing, Mohan Babu University, Tirupati, A.P,INDIA.
Email:rsreddyphd@gmail.com

Abstract - Speech Emotion Recognition (SER) is now progressively vital for many practical uses including virtual assistants, customer service, and healthcare monitoring as well as for SER systems still suffer with environmental noise, speaker variability, and cross-lingual adaptation that affect their accuracy and generalizing power even if they have made tremendous progress. This work introduces ExpressNet, an optimum Multi-Layer Perceptron (MLP)-based SER model aimed to solve these issues by leveraging a wide range of prosodic and spectral qualities incorporating Mel-Frequency Cepstral Coefficients (MFCCs), spectral contrast, and pitch variations. ReLU activation and a softmax output layer allow the model to classify six emotional states: anger, disgust, fear, happiness, neutral, and sad using a deep learning architecture. We assess ExpressNet using the CREMA-D dataset and achieve a test accuracy of 92.97%, above the results of previous state-of-the-art approaches. Particularly real-time applications gain from the method since it helps to blend high classification accuracy with computing efficiency. Our work emphasizes the need of applying deep learning methods with enhanced feature engineering to improve SER performance. Furthermore, we show a thorough assessment over numerous benchmark datasets to show the power and applicability capacity of the model in many different settings. Apart from being better than other options, ExpressNet is a consistent choice for use in real-world settings since it has a low overfitting rate. This work advances emotional computing by providing a solid and scalable foundation for SER, which will enable further research in emotional recognition systems. The probable utilization of this technique extends to mental health monitoring and human-computer interaction because it demonstrates excellence at handling complex emotional patterns in voice signals. Extensive research into self-supervised learning and multimodal data integration and cross-lingual adaptation will improve the model's potential

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

across multiple application scenarios.

Keywords – Speech emotion recognition, Deep learning, Feature engineering, Affective computing, MLP, Real-time emotion detection, Sentiment Analysis

I. Introduction

In many of different applications, the Speech Emotion Recognition (SER) technology is becoming increasingly incorporated into daily life [1]. Now its application has spread to sports, e-learning, voice search, aviation cockpit call of sorts and so on. The major goal of a SER system is to correct interpretation of human emotions [2]. Despite the active development in the sphere of the counting of emotions, there are many barriers preventing its high accuracy [3]. Of these issues, environmental background, culture, and the use of language on the human body impact the display of emotions [4]. For understanding human emotions, it is necessary for us to understand how spoken words are delivered and that requires content analysis and interpretation of the spoken words. Each same speech takes different meanings based on vocal inflection, however, expressing emotions depend mostly on tone, pitch, intensity, and rhythm [5]. A dependable method to capture genuine speaker emotions needs efficient systems which process verbal along with nonverbal cues.

Being the process of identifying and categorizing the emotions and sentiments expressed in spoken language, speech sentiment and emotion recognition (SER) is considered as a basic task in affective computing and human-computer interaction (HCI) [6]. In such environments, place, poeples, and even languages, current techniques are still having challenges generalizing among several speakers and loud environments. Originality that is included in many traditional techniques is the handcrafted acoustic elements, and emotions are rather difficult to describe in full. Most models have challenges in practicing robustness in the real world and non-stationary setting ; usual approaches may involve designing specific acoustic components [7], that may barely capture most of the emotion. These difficulties are further fueled by the ability of humans to speak in different tones, with different accents and even under noisy environments therefore affecting the ability to capture their emotions.

Increased concern in SER has been occasioned with its versatile employment areas, which include personal assistant, health check up, customer relation centre and health sector [8]. In the old concept SER models believes largely on the hand made features such as Mel Frequency Cepstral Coefficient (MFCC) [9], Spectral features, and prosodic features that are classified through ML classifiers such as SVM [10] and Random Forest (RF) [11]. These techniques are promising but they base the features of the image on hand extraction which limits them in their undertakings of encompassing all the aspect of emotions. Neural networks particularly with regard to convolutional neural networks (CNNs) [12] and recurrent neural networks (RNNs) [13] have improved SER performance through concatenatively exposing the voice signals to gather hierarchical features. These methods in most cases are computationally intensive and require large data, they therefore pose a threat to scalability and real-time capability.

Investigations on SER have increased due to the increasing demand for reliable, fast, and online emotion detection in the HCI. Despite the fact that the present models perform well under slow-fauth environment the manner in which people speak in real-life differs in a way that hampers

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

their performance [14] due to noise, variations of pace, tones among others [15]. Some of the recent deep learning models also require large processing power, despite this they are not very suitable for use cases in low-cost platforms for edge devices. Solving these challenges requires a more flexible and efficient approach that provides good practicality while maintaining high accuracy of classification. This paper presents an optimal MLP for SER concerned with the efficiency of characteristics of speech recognition together with generalization capability and deployment.

Hence, our suggested method provides the best MLP model for speech sentiment and emotion recognition through essential speech features, namely MFCCs, spectral contrast, and pitch fluctuations. The model forecasts emotional states from many deep layers by means of ReLU activation capturing complex patterns and a softmax categorization layer. We extend the model to be versatile for a range of applications including virtual assistants, healthcare monitoring, and affective computing by means of hyperparameter adjustment and regularization, so improving both computational efficiency and classification accuracy.

This study possesses these following contributions:

1. In the field of speech emotion recognition, the formulation of an optimal ExpressNet that is a multi-layer perceptron (MLP)-based model that achieves a harmonic balance between the accuracy and the efficiency.
2. The incorporation of a wide array of spectral and prosodic speech features to enhance classification performance.
3. A comprehensive evaluation conducted across many benchmark datasets to ascertain generalizability and robustness and
4. A comparative analysis demonstrating the superior performance of our model relative to existing state-of- the-art methodologies.

This work consists in several parts. The present methods for the crowd counting will be discussed in Section 2. Section 3 will next go over our suggested method and momentarily discuss the configuration for the application of our model. Section IV will compile the experimental findings applying the models we have proposed. Finally, we shall show our results of investigation.

II. Literature Review

From traditional machine learning approaches using manually created acoustic features to advanced deep learning methods depending on automatic feature extraction, speech sentiment and emotion recognition (SER) has evolved. Though models include CNNs, RNNs, and hybrid architectures have improved accuracy, problems still exist including speaker variability, cross-lingual adaptation, and computing economy. Emphasizing important developments, constraints, and the need of a scalable and efficient SER model, this part reviews current research.

Speech emotion recognition (SER) has progressed much using numerous approaches including deep learning, multimodal fusion, feature extraction techniques, and attention procedures to boost accuracy. While current methods incorporate self-supervised learning, cross-modal fusion, and multi-task learning to boost resilience, traditional strategies depended on handcrafted acoustic features including Mel Frequency Cepstral Coefficients (MFCCs) and spectrograms. The papers under review look at numerous roles and each specifically adds to the field.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

Self-supervised learning has been proposed as a possible approach for SER whereby models learn feature representations without utilizing extensive volumes of labeled data. This approach facilitates the generalization between several datasets and emotions. Atmaja et al. [16] used self-supervised universal speech representation models with speaker-aware pre-training for sentiment and emotion detection. Their model suffered with multi-class classification due to dataset imbalance, so stressing the limits of self-supervised models when used to challenging emotional tasks. Good accuracy was obtained by its binary sentiment analysis (73% unweighted and 81% weighted).

Combining text, auditory, and visual inputs has enabled multimodal emotion identification become rather common in order to increase classification accuracy. These models derive more relevant emotional representations by combining many data sources than by depending solely on auditory information. Using fastText for textual analysis, a 1-D CNN for audio, a 2-D CNN for images, Dixit et al. [17] produced a real-time multimodal system for assessing human oration movies. By the use of bagging and stacking techniques, their model using the CMU-MOSEI dataset got an accuracy of 85.85% and an F1-score of 83. In a same line, Mamieva et al. [18] proposed an attention-based multimodal technique integrating facial and voice traits to focus on the most informative components of every modality solely. Their technique outperformed earlier models with a weighted accuracy of 74.6% and an F1-score of 66.1% for IEMOCAP and with an F1-score of 73.7% for CMU-MOSEI and a WA of 80.7%. These studies underscore the degree of multimodal fusion enhancement of SER performance.

SER still rely largely on feature engineering, where advanced feature fusion techniques can greatly increase identification accuracy. Zhao et al. [19] approached a multi-level acoustic feature cross-fusion technique using MFCC, spectrograms, and Wav2vec2 embeddings to improve emotional recognition. Their method fills in for missing data and displays state-of- the-art (SOTA) performance under such conditions by cross-fusion. With MFCC, mel-spectrogram, approximative entropy (ApEn), and permutation entropy (PrEn), Mishra et al. [20] focused on variational mode decomposition (VMD)-based feature extraction. Their deep neural network (DNN) classifier on RAVDESS and EMO-DB reached respectively recognition rates of 91.59% and 80.83%. These results suggest that improved feature fusion techniques significantly affect improvements in SER performance.

Enhanced temporal and contextual feature extraction, attention mechanisms and deep learning architectures including CNNs, LSTMs, and transformer-based networks has changed SER. For eight emotional state detection, Singh et al. [21] identified MFCCs to be the most successful feature applying a self-attention-based deep learning model integrating 2D CNNs with LSTMs. Their approach shown promise for usage in mental health with 90% accuracy. Khan et al. [22] similarly built a Deep Echo-State Network (DeepESN) with a dilated CNN and multi-headed attention by using reservoir computing for high-dimensional feature mapping. Examined on EMO-DB and RAVDESS, their model achieved 91.14% and 85.57% recognition rates for speaker-dependent tests, therefore preserving low processing costs while outperforming traditional deep learning models. These tests indicate the degree of improvement in SER

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

predictions by attention-based architectures.

Cross-lingual and dataset generalizing challenges have been addressed in part by data augmentation and multi-task learning. Khan et al. [23], built a deep learning model employing Urdu, Italian, English, and German datasets using MFCCs as feature representations, therefore solving cross-lingual SER. Their approach showed versatility spanning many languages since Random Forest (RF) and XGBoost classifiers obtained 91.25% accuracy on the URDU dataset. Liu et al. [24] underlined generality by using balanced augmented sampling on log-Mel spectrograms with triple channels. CNN with an attentive-based bidirectional LSTM, they coupled multi-task learning to maximize auxiliary tasks for a higher model performance. Their model exceeded previous methods with 70.27% WAR/UAR on IEMOCAP and 60.90% and 61.83% on MSP-IMPROV accordingly. These results reveal that using multi-task learning and improving dataset variety greatly increase SER model durability.

Finally, entropy-based methods have been studied to improve emotional classification by means of statistical feature extraction. Inspired by MFCC-derived statistical features including spectral entropy and approximative entropy, second work by Mishra et al. [25] proposed an entropy-based technique. Evaluated on EMO-DB, RAVDESS, and SAVEE, their DNN classifier produced respectively classification accuracy of 87.48%, 75.9%, and 79.64%. Combining all attributes gave somewhat less accuracy, suggesting that various datasets responded differently depending on certain entropy-based properties. This emphasizes the requirement of adapting datasets specifically as well as the opportunities of entropy measures to raise SER performance.
Even although Speech Emotion Recognition (SER) has progressed greatly, some limitations remain exist throughout current studies. Found in Atmaja et al. [16], self-supervised models suffer with multi-class classification due to dataset imbalance, thereby limiting its scalability to practical use. Dixit et al. [17] and Mamieva et al. [18]among others exhibit performance improvement even if they struggle with computational complexity, data synchronizing, and rely on sometimes rare high-quality multimodal datasets. Feature fusion techniques including Zhao et al. [19] and Mishra et al. [20] improve classification accuracy but they are less suitable for real-time applications since they need careful calibration of feature combinations and large processing resources. Though they demand large labeled datasets for maximum performance and are prone to overfitting, attention-based models such as those proposed by Singh et al. [21] and Khan et al. [22] show improved temporal and contextual feature extraction. Although cross-lingual models like Khan et al. [23] have promise in multilingual environments, they suffer from emotional discrepancies across many languages, hence reducing generalization. While generalizing-oriented approaches such as Liu et al. [24] serve to lower class imbalance by augmentation, they might not entirely reflect the subtle variations of emotions over many datasets. Ultimately, although under investigation by Mishra et al. [25] , entropy-based methods provide interesting statistical analysis but lack flexibility between datasets, hence producing contradicting conclusions. These challenges taken together indicate the ongoing demand for efficient, generalizable, computationally feasible SER models able to robustly run across various datasets, languages, and real-world contexts.
The method, performance, constraints, and major contributions to Speech Emotion Recognition

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

(SER) of every research project are compared in *Table 1*.

TABLE 1: SUMMARY OF LITERATURE ON SPEECH EMOTION RECOGNITION (SER) METHODS

References	Method	Result	Findings	Limitations
Atmaja et al. [16]	Self-supervised universal speech representation with speaker-aware pre-training.	81% WA, 73% UA (Binary Sentiment)	Effective for binary sentiment analysis but not for multi-class emotion tasks.	Struggles with multi-class classification due to dataset imbalance.
Dixit et al. [17]	Multimodal fusion (fastText for text, 1D CNN for audio, 2D CNN for image) with bagging and stacking.	85.85% Accuracy, 83% F1-score (CMU-MOSEI)	Cross-modal training enhances generalization and robustness.	High computational complexity and reliance on well-synchronized multimodal data.
Zhao et al. [19]	Multi-level acoustic feature cross-fusion with gender-aware multitask learning.	WA and UA of 72.04% and 73.26%, respectively	Cross-fusion compensates for missing information, improving SER accuracy.	Requires extensive feature engineering and tuning for optimal performance.
Mamieva et al. [18]	Attention-based multimodal emotion recognition integrating facial and speech features.	74.6% WA, 66.1% F1 (IEMOCAP); 80.7% WA, 73.7% F1 (CMU-MOSEI)	Selective attention enhances feature extraction for emotion recognition.	Requires large, high-quality multimodal datasets; increased computational cost.
Khan et al. [22]	Deep Echo-State Network (DeepESN) with Sparse Random Projection (SRP).	91.14% (EMO-DB), 85.57% (RAVDESS) for speaker-dependent; 82.01%, 77.02% for speaker-independent	Reservoir computing and early fusion improve classification while reducing complexity.	High reliance on hyperparameter tuning; computational efficiency remains a challenge.
Khan et al. [23]	Cross-lingual SER with MFCC-based feature extraction using	91.25% accuracy (URDU dataset)	Cross-lingual training enhances SER across different languages but	Language inconsistencies reduce generalization across datasets.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

	Random Forest and XGBoost.		requires adaptation.	
Singh et al. [21]	2D CNN + LSTM with spectral and rhythmic feature extraction.	90% accuracy (custom dataset)	Self-attention mechanism improves contextual emotion recognition.	Prone to overfitting, requiring large labeled datasets.
Mishra et al. [20]	Variational Mode Decomposition (VMD)-based feature extraction with a Deep Neural Network (DNN).	91.59% (RAVDESS), 80.83% (EMO-DB)	VMD-based feature fusion improves SER accuracy over traditional methods.	High computational cost and dataset-dependent performance.
Liu et al.[24]	Balanced augmented sampling with CNN + Attention-based BiLSTM.	70.27% WAR, 66.27% UAR (IEMOCAP); 60.90% WAR, 61.83% UAR (MSP-IMPROV)	Multi-task learning enhances feature diversity, reducing dataset bias.	Augmentation improves balance but may not fully capture emotion variations.
Mishra et al. [25]	MFCC-based entropy feature extraction with a DNN classifier.	87.48% (EMO-DB), 75.9% (RAVDESS), 79.64% (SAVEE)	Entropy-based MFCC features improve classification but require fine-tuning.	Limited adaptability across datasets, leading to inconsistent results.

III. METHODOLOGY

This section gives an exposition of the methods that were used in our Speech Emotion Recognition (SER) system. Our method uses a structured pipeline that includes data preparation, feature extraction, model construction, training, and evaluation. Our research has an overall framework shown here in *Fig. 1*.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

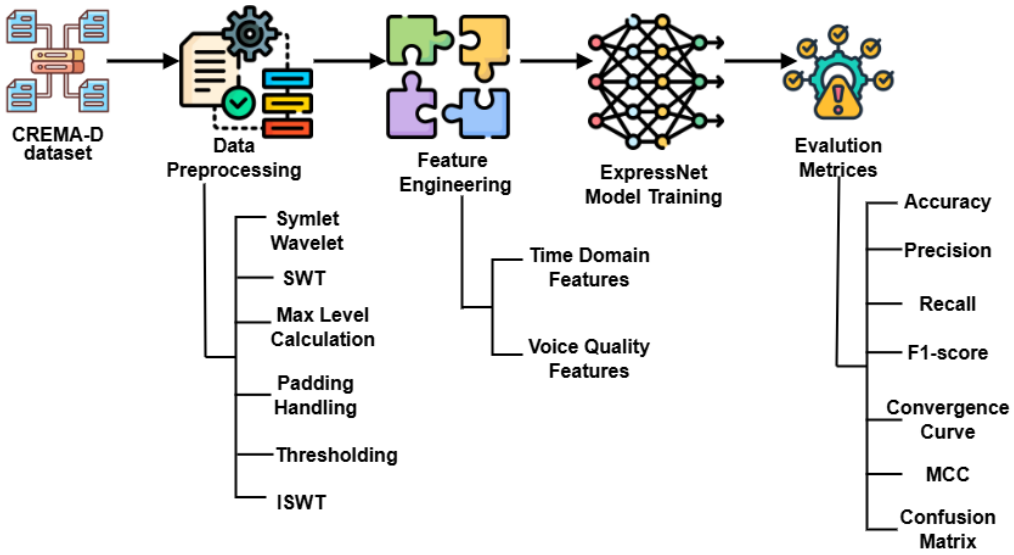


FIG. 1 OVERVIEW OF THE STUDY FRAMEWORK, ILLUSTRATING THE KEY STAGES: DATA PREPROCESSING, FEATURE EXTRACTION, MODEL ARCHITECTURE (EXPRESSNET) AND EVALUATION METRICES

3.1 DATASET OVERVIEW

The 7,440 original audio recordings from 91 actors, ages 20 to 74 (48 men and 43 women), from a variety of racial and ethnic backgrounds, including African American, Asian, Caucasian, Hispanic, and unidentified groups, make up the CREMA-D dataset, which was used in this study. Twelve different words representing six different emotions—Anger, Disgust, Fear, Happy, Neutral, and Sad—were delivered by each actor at four different intensity levels: Low, Medium, High, and Unspecified. 2,443 people participated in a massive crowdsourcing project to rate the emotion and intensity of 90 clips (30 audio-only, 30 video-only, and 30 audiovisual). More than seven separate assessments were given to about 95% of the clips, guaranteeing accurate and thorough annotations. Fig. 2 shows the overall class distribution for each emotion group.

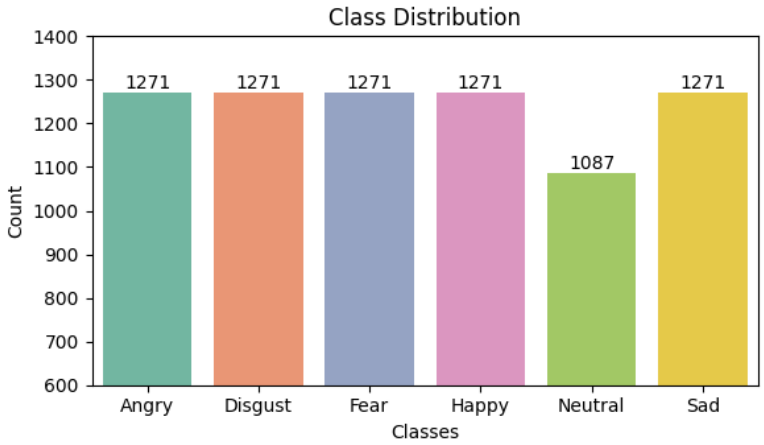


FIG. 2 NUMBER OF SAMPLES FOR EACH CLASS IN OUR DATASET

The dataset is splitted into a 70:30 ratio for training and testing, as shown in *Table 2*.

TABLE 2. DATASET SPLIT INTO TRAINING AND TEST SETS (70:30 RATIO)

Class	Train	Test
Angry	872	399

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

Disgust	919	352
Fear	916	355
Happy	882	389
Neutral	754	333
Sad	866	403

3.2 DATA PREPROCESSING

By lowering noise, normalizing data, and getting it ready for feature extraction later on, preprocessing is essential to improving the quality of incoming audio signals. The preprocessing pipeline in the suggested architecture consists of:

3.2.1 Symlet Wavelet

The Symlet Wavelet Transform (SWT) is used in the suggested architecture to break up speech signals into smaller pieces. Speech signal analysis benefits greatly from the greater symmetry provided by symlets, a symmetric variant of Daubechies wavelets. The following is a representation of the SWT at the decomposition level j :

$$S_j[n] = \sum_k g[k-2n]S_{j-1}[k]$$

where $g[k]$ represents the wavelet scaling function, $S_j[n]$ IS the wavelet coefficient at level j,$S_{j-1}[k]$ represents the input signal at the level before it.

3.2.2 Stationary Wavelet Transform (SWT)

After the speech signal has been preprocessed, features are extracted using the Stationary Wavelet Transform (SWT). Because it maintains the required temporal precision while capturing both high- and low-frequency components, this method is ideal for speech analysis. SWT does not downsample the signal at each level of decomposition, in contrast to the Discrete Wavelet Transform (DWT), which is translation-invariant. This characteristic is very helpful for examining non-stationary signals, like speech. The following is the definition of the level-j decomposition:

$$A_j[n] = \sum_k g\left[k-2^j n\right]A_{j-1}[k]$$

$$D_j[n] = \sum_k h\left[k-2^j n\right]A_{j-1}[k]$$

Where, $A_j[n]$ represents the approximation coefficients at level j, $D_j[n]$ represents the detail coefficients, $g[k]$
and $h[k]$ are the low-pass and high-pass filters, respectively.

3.2.3 Max Level Calculation

One of the factors influencing the greatest amount of decomposition in SWT is the input signal's duration. The following formula is applied to determine the optimal decomposition level L:

$$L = log_2(N)$$

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

where N is the quantity of samples in the incoming audio stream. This guarantees that wavelet decomposition is carried out as effectively and efficiently as possible without the need for additional procedures.

3.2.4 Padding Handling

Wavelet transformations require the signal length to be a power of two; therefore, padding techniques are applied to satisfy this condition:

- **Zero Padding:** The signal is extended by appending zeros until its length reaches the nearest power of two.

$$x_{padded}(n) = \begin{cases} x(n), & 0 \le n \le N \\ 0, & N \le n \le N' \end{cases}$$

Where *'N'* is the nearest power of two.

- **Mirror Padding:** The signal is extended by a symmetric reflection of its samples, which helps minimize edge discontinuities.

3.2.5 Thresholding

It is processed after padding handling and to remove the noise level while maintaining the essential frequency components. Of all the thresholding functions that have been developed, the following is the most commonly applied:

$$T(x) = \begin{cases} 0, & |x| < \lambda \\ x, & |x| < \lambda \end{cases}$$

3.2.6 Inverse Stationary Wavelet Transform (ISWT)

The signal is then reconstructed through the inverse SWT following the subsequent procedures of denoising and enhancement of signal. This ensures that all the prominent features of speech are retained while cutting short the fat or the superfluous part of the message. The following now is the reverse transformation:

$$X_j(n) = \sum_k h(n-2k) X_{j+1}(k)$$

This restores the original signal while removing unwanted noise.

3.3 Feature Extraction

By transforming unprocessed speech waveforms into numerical representations that capture changes in pitch, energy, and spectral features, feature extraction is essential to Speech Emotion Recognition (SER). Time-Domain Features and Voice Quality Features are the two primary categories into which these retrieved features are often divided.

3.3.1 Time-Domain Features

Time-domain characteristics analyze how the value of the amplitude of the voice stream varies in the time. These characteristics help in the distinction between different categories of speaking by measuring its loudness and rate.

a) Zero-Crossing Rate (ZCR)

In our work, unvoiced and voiced speech components are segregated using ZCR; this is

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

important in classifying the emotions that are characterized by difference in the mode of articulation. For instance, although the calm emotions, such as melancholy, have a lower ZCR because the majorities of them involves voice, the high-arousal emotions, such as rage and terror, have a greater ZCR because of the growing number of unvoiced phonemes. ZCR is computed mathematically by counting the times the signal flips its sign inside a frame using the formula:

$$ZCR = \frac{1}{N}\sum_{n=1}^{N-1} 1[s(n).s(n-1) < 0]$$

Where $s(n)$ represents the speech signal at sample n, and 1is an indicator function that marks a sign change.

b) Short-Term Energy (STE)

We evaluate the intensity of emotional speech using Short-Term Energy (STE). Higher STE for energetic emotions including delight, surprise, and anger suggests more powerful articulation. Emotions like despair, on the other hand, show lower STE and represent less vocal effort. This ability helps especially to differentiate high-energy from low-energy emotional states. The formula for calculating STE is:

$$E(n) = \sum_{m=0}^{M-1} s^2(n-m)$$

Where M is the window length, and $s(n)$ is the signal at each sample point.

c) Audio Kurtosis

Kurtosis facilitates the study of the energy distribution in voice stream. High kurtosis values point to rapid energy bursts that define agitated or stressed speech—that is, anger, surprise. Often detected in neutral or sad speech patterns, lower kurtosis values correlate to equally dispersed energy. Our work uses kurtosis to improve the categorization of speech intensity, therefore complementing STE. Kurtosis's formula is given as:

$$K = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^4}{(\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2)^2}$$

Where x_i is the signal amplitude, μ is the mean amplitude, and N is the total number of samples.

3.3.2 Voice Quality Features

Voice quality features emphasize the frequency characteristics of the speech signal, such as the spectral distribution, harmonics, and pitch variations, all of which are crucial for distinguishing emotions. These features reflect the tone, timbre, and pitch variations, helping to capture subtle differences in emotional expressions.

a) Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs are used to extract in speech most perceptually significant elements. These coefficients enable the capture of vocal tract resonances, which change with emotional

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

states. Emotions like happiness and rage, for instance, move the spectral envelope toward higher frequencies; melancholy and boredom produce a lower-frequency spectral focus. Our work efficiently models these speech dynamics using MFCCs.

The method of computing MFCCs is first the Discrete Fourier Transform (DFT) of the signal, then a Mel filterbank converts to the Mel scale, takes the logarithm of the energy, and next uses the Discrete Cosine Transform (DCT) to decorrelate the coefficients. This produces a set of quite successful coefficients for identifying trends in speech and emotion.

b) Chroma Features

Our work examines tonal fluctuations in speech using chroma characteristics. Chroma characteristics assist differentiate emotions like delight and surprise, which typically have harmonic rich spectra, from emotions like melancholy, which have a smaller spectral range since emotional speech generally involves variations in pitch and harmonic structure.

c) Spectral Centroid

Measuring the "brightness" of speech, spectral centroid reveals where most spectral energy is focused. Our work uses this ability to differentiate between emotions:

- **High spectral centroid:** Indicative of enthusiasm, rage, because of greater high-frequency energy.
- **Low spectral centroid:** Common in calm and sorrowful speech, in which spectral energy is focused in lower frequencies.

The formula for calculating spectral centroid is:

$$S = \frac{\sum f_i X(f_i)}{\sum X(f_i)}$$

where f_i is the frequency bin, and $X(f_i)$ is the magnitude of the spectrum at that frequency.

d) Spectral Bandwidth

Spectral bandwidth quantifies spectral energy's distribution. This function helps us distinguish between emotions with abrupt, sudden spectral variations—anger, surprise—and those with more consistent spectral content sadness, tranquility. Lower bandwidth denotes more smooth speech patterns; higher bandwidth signifies more spectral variability.

e) Spectral Entropy

Spectral entropy estimates speech spectrum unpredictability. We investigate in our work the degree of structural or chaotic nature of the spectral content:

- **Low entropy: Indices organized speech (e.g., peacefulness, melancholy).**
- **High entropy: Indices erratic fluctuations (e.g., surprise, anxiety).**

The formula for spectral entropy is:

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

$$H = -\sum_i p(f_i) log p(f_i)$$

where $p(f_i)$ is the normalized spectral power at frequency f_i.

f) Spectral Flux

Spectral flux quantifies spectral content's change with time. We investigate the dynamics of emotional speech using it, whereby:

- **High spectral flux:** Indicates in excited or angry speech quick spectral fluctuations.
- **Low spectral flux:** Found in neutral or melancholy speech, in which frequency fluctuations are few.

The formula for spectral flux is:

$$SF = \sum_i (X_{i+1}(f) - X_i(f))^2$$

where $X_i(f)$ is the spectral magnitude at frame i.

g) Spectral Rolloff (85%)

DIFFERENTIATING VOICED FROM UNVOICED COMPONENTS OF SPEECH IS MADE POSSIBLE IN PART BY SPECTRAL ROLLOFF. OUR WORK USES THIS CHARACTERISTIC TO SEPARATE EMOTIONS DEPENDING ON THEIR HIGH-FREQUENCY ENERGY CONTENT:

- **HIGH ROLLOFF : INDICATING MORE HIGH-FREQUENCY COMPONENTS, FOUND IN WRATH, SURPRISE.**
- **LOW ROLLOFF : COMMON IN DEPRESSION, QUIET SPEECH, MOSTLY LOW-FREQUENCY COMPONENT DOMINATED SPEECH.**

3.4 EXPRESSNET MODEL ARCHITECTURE

Following feature extraction, a deep learning model ExpressNet works through several Dense (Fully Connected) Layers for categorization into six emotions: Angry, Disgust, Fear, Happy, Neutral, and Sad. The architectural view of our proposed model is visualized in *Fig. 3*.

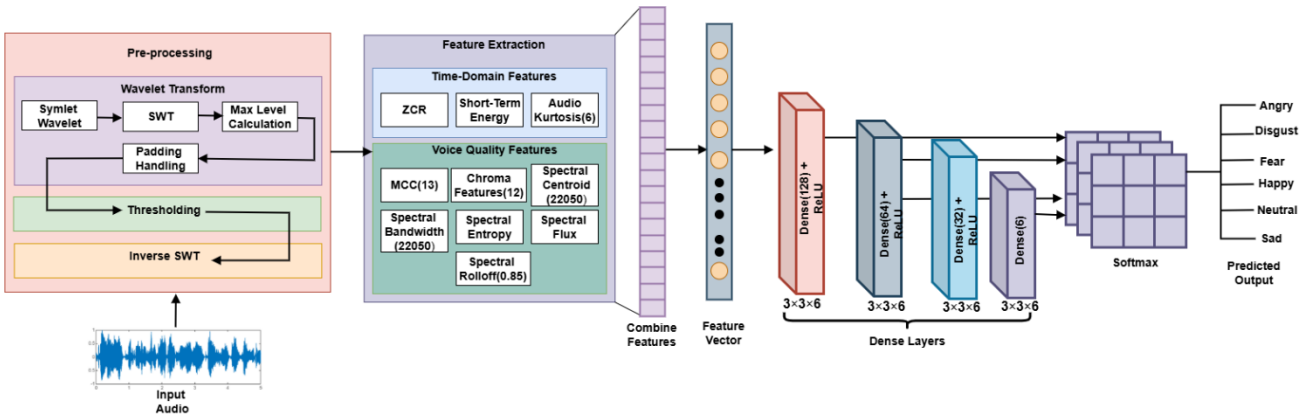


FIG. 3. THE ARCHITECTURE OF EXPRESSNET, A DEEP LEARNING MODEL FOR SPEECH EMOTION CLASSIFICATION

3.4.1 Feature Vector Representation

The input for the classification model is the single feature vector generated from the retrieved features, which ensures that only the most relevant information is retained, hence simplifying the

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

learning process and reducing unnecessary computations.

3.4.2 Dense Layers with ReLU Activation

Rectified Linear Unit (ReLU) activation is used in the completely connected layers and is described as follows:

$$f(x) = \max(0, x)$$

This activation function allows the model to learn complex relationships within the feature space while maintaining computational efficiency. This is the pattern that the construction follows:

High-level patterns are extracted from the input feature vector using **Dense (128) + ReLU.**
Dense (64) + ReLU: Enhances representations by capturing significant discriminative factors.
Dense (32) + ReLU: Optimizes feature space for classification even more.
Dense (6): Generate logits for every emotional class.
This hierarchical framework guarantees a progressive decrease in dimensionality, therefore enabling the network to concentrate on the most important features of the input data. ReLU reduces gradient disappear problems, therefore producing consistent and effective training.

3.4.3 Softmax Output Layer

The final dense layer uses the Softmax activation function to convert logits into probability distributions:

$$P(y_i) = \frac{e^{yi}}{\sum_j e^{yj}}$$

The pre-activation output of the preceding layer is represented as y_i, where $P(y_i)$ is the class probability i given the input. Since Softmax ensures that the sum of all outputs is 1, the model's predictions can be seen as class probabilities.

The structure is suitable for multi-class classification and permits clear decision boundaries between emotions. The final output provides confidence scores for each class, allowing for robust decision-making in applications that call for behavior analysis and emotional computing.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

IV. EXPERIMENTAL SETUP AND EVALUATION

4.1 EXPERIMENTAL SETUP

This project's goal was to provide a scalable and reliable training framework for speech-based emotion recognition. The methodology, which was implemented in Python, combined Keras for deep learning model creation with Mealpy for hyperparameter optimization. Pandas, Matplotlib, and Seaborn allowed exploratory analysis and visualization, while Scikit-learn handled data preprocessing. StandardScaler was used to normalize audio features that were kept in an internal NPZ format for easy access. TensorFlow–Keras interoperability on GPU infrastructure made high-performance training possible. For efficient hyperparameter tuning, Mealpy's Particle Swarm Optimization (PSO) algorithm with multi-threading was employed. A hold-out set comprising 30% of the dataset was reserved for testing, while the remaining data was iteratively refined for training. Model performance metrics and optimization history were systematically tracked to inform subsequent research directions.

4.2 PREPROCESSED AUDIO WAVEFORM VISUALIZATION

Preprocessed audio waveforms from several emotional classes are shown in Fig. 4; the orange waveform (NoiseRemoved) represents the denoised version, while the blue waveform (Source) displays the original signal. High-amplitude differences between samples 10,000 and 20,000 in every figure show that there is emotional intensity in these areas. According to the plots, anger and fear typically exhibit more noticeable shifts in energy levels than other emotions. The Fear plot's amplitude (~±0.6) is greater than the first one's (~±0.3). Even though the noise-removed signals appear smoother and more closely resemble the original waveform, they really reduce noise while maintaining emotional qualities. Denoised audio may lessen unwanted fluctuations and improve model accuracy because these preprocessed signals are most likely intended for deep learning-based speech emotion recognition.Disgust, however, exhibits certain variations. Happy maintains a very consistent amplitude pattern throughout, demonstrating a more uniform energy distribution.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

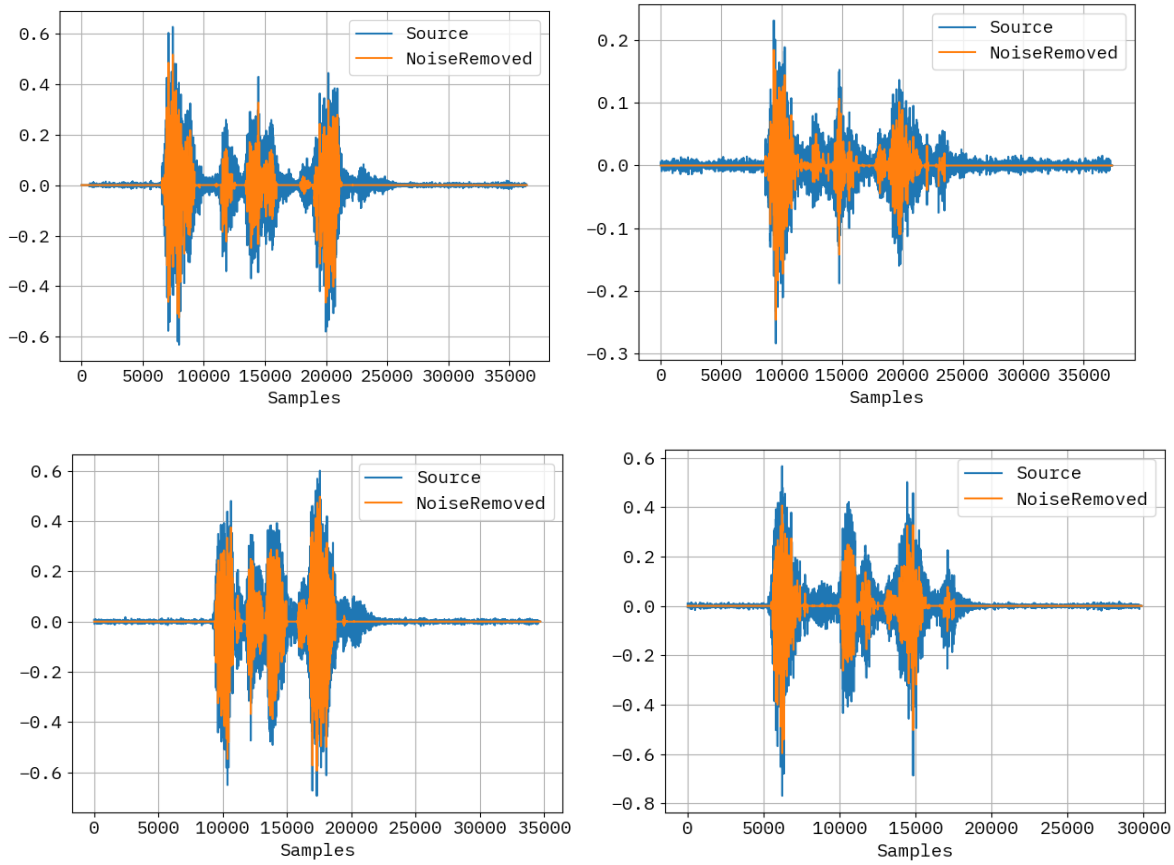


FIG. 4. PREPROCESSED AUDIO WAVEFORMS OF EMOTIONAL SPEECH SIGNALS FOR TOP-LEFT (ANGRY), TOP_RIGHT (DISGUST), BOTTOM_LEFT (FEAR), AND BOTTOM_RIGHT (HAPPY) CLASSES

4.3 HYPERPARAMETER AND MODEL BEHAVIOR

Various important hyperparameters determine how the model architecture functions together with its training approach and model optimization. These hyperparameters handle multiple controls which include batch size as well as learning rate and hidden units and epochs. Strong and efficient training becomes possible when using the RMSprop optimizer because it modifies learning rates based on gradient conditions. The categorical crossentropy loss function serves multi-class classification by comparing expected class probability distributions to actual label identifications. Below *Table 3* is a summary of the hyperparameters and their respective ranges:

TABLE 3. SUMMARY OF HYPERPARAMETERS USED IN THE MODEL OPTIMIZATION PROCESS

Hyperparameter	Range/Value	Description
h1units	[128, 256]	The number of units (neurons) in the first dense layer of the neural network.
h2units	[64, 128]	The number of units (neurons) in the second dense layer of the neural network.
h3units	[32, 64]	The number of units (neurons) in the third dense layer of the neural network.
epochs	[1, 100]	The number of epochs (iterations) for training the model.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

batch_size	[128, 1024]	The batch size to be used during training.
learning_rate	[0.0001, 0.01]	The learning rate for the RMSprop optimizer.
optimizer	RMSprop	Ensuring efficient convergence during training.
loss function	categorical crossentropy	Calculates the difference between the true class labels and the predicted probabilities.

During training cycles, the six graphs in *Fig. 5* follow important hyperparameters to offer a full view of how these values change during the optimization process. The iterative against h1units graph demonstrates how the count of the first hidden layer varies with architectural shaping by the optimizer. The Iteration vs. h2units and Iteration vs. h3units graphs also demonstrate how the optimizer changes the amount of units in the second and third hidden layers appropriately reflecting how the complexity of the model is changed to optimal performance at each iteration. The graph Iteration vs. batch_size shows the variations in batch size, therefore exposing how this value is altered throughout iterations to improve model generalization and training efficiency. The iterative versus learning rate graph displaying how the learning rate changes over training makes effective convergence free from overshooting the optimum response achievable. At last, the iterative graph showing the total number of finished training epochs provides background for the other hyperparameter development. These six numbers taken together show a full picture of how the optimization process alters important model parameters across iterations to raise the model's performance.

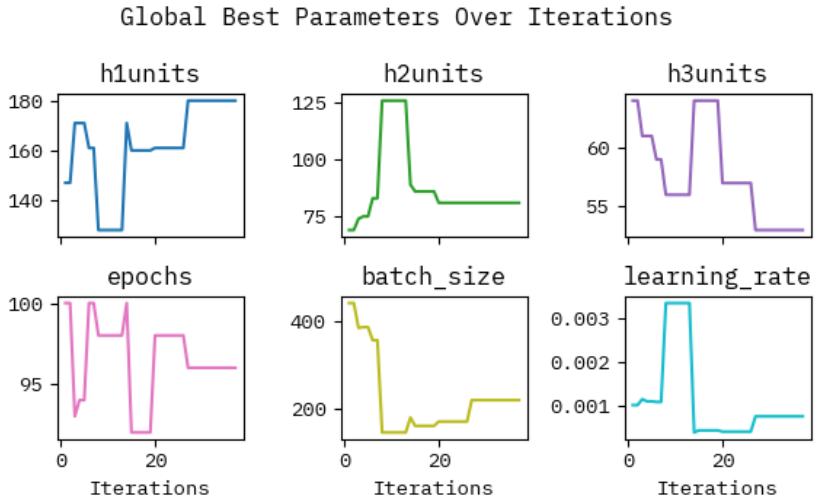


FIG. 5. EVOLUTION OF GLOBAL BEST HYPERPARAMETERS OVER ITERATIONS

4.4 PERFORMANCE EVALUATION AND COMPARATIVE ANALYSIS

For the training and test sets, many criteria accuracy, precision, recall, F1-score, and Matthews correlation coefficient help to assess the model. Showing great generalizing as depicted in Figure X, the model has a test accuracy of 92.97% and a training accuracy of 91.92%. Reflecting the capabilities of the model for exact classification, accuracy which counts the proportion of precisely predicted positive cases among all predicted positives is 92.90% for training and 93.66% for testing. Recall guarantees 100% detection by showing how well the model can identify all true positive cases: 91.97% for training and 92.67% for testing. Approaching a harmonic mean of accuracy and precision, the F1-score for testing is 92.95%; for training it is

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

92.08%. Moreover, implying a clear link between predictions and actual labels are the MCC values of 0.9164 (test) and 0.9045 (training). These low overfitting and very predictive values, as shown in *Fig. 6*, show how successfully the model maintains high performance over both training and test sets.

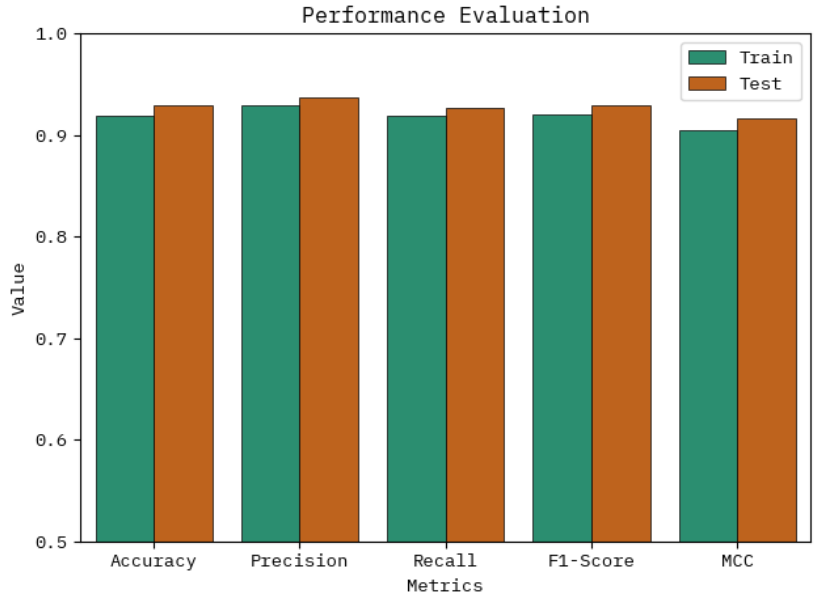
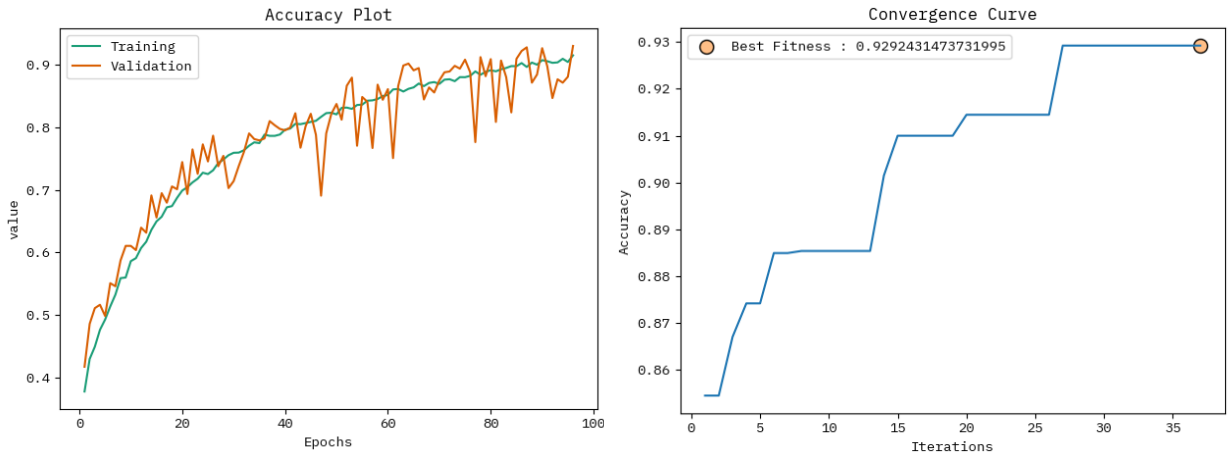


FIG. 6. PERFORMANCE METRICS COMPARISON FOR TRAINING AND TEST DATA

ANALYZING THE PERFORMANCE OF THE MODEL AS SHOWN IN FIG. 7 WHICH SHOWS THE CONVERGENCE OF THE OPTIMIZATION PROCESS AND THE ACCURACY DEVELOPMENT OVER TRAINING—HELPS ONE TO PROBE DEEPER THE ACCURACY AND CONVERGENCE PLOTS. THE



ACCURACY GRAPH SHOWS A CONTINUOUS IMPROVEMENT IN MODEL PERFORMANCE, WHICH AT LAST REACHES A PEAK BEST FITNESS VALUE OF 0.9294, THEREFORE EXHIBITING OPTIMAL LEARNING. MOREOVER, THE CONVERGENCE GRAPH UNDERLINES HOW EFFICIENTLY THE OPTIMIZATION TECHNIQUE FUNCTIONS BY DISPLAYING A STABLE AND SMOOTH CONVERGENCE TOWARD THE IDEAL SOLUTION.

FIG. 7. TRAINING AND VALIDATION ACCURACY AND CONVERGENCE CURVES FOR OUR PROPOSED MODEL

Moreover displayed in *Fig. 8* is the loss curve, which facilitates understanding of the learning

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

pattern of the model. Given a consistent drop in loss over seasons, it shows good training and fewer error. Even while stability across the training period is maintained, the declining loss curve guarantees effective learning of patterns from the data. These pictures collected together help to support the validity of the model and the effectiveness of the optimization technique in acquiring good performance.

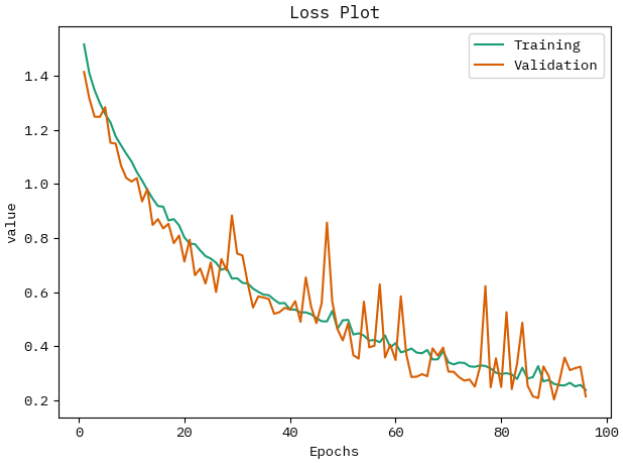


FIG. 8. TRAINING AND VALIDATION LOSS CURVE FOR OUR PROPOSED MODEL

TABLE 4. RECENT STUDIES ON SPEECH EMOTION RECOGNITION, INCLUDING OUR PROPOSED APPROACH

References	Method	Result	Year
Alsaadawi and Daş [27]	Bi-LG-GNN-based Multimodal Emotion Recognition	80% accuracy, 81% F1-score, precision, and recall	2024
Koti et al. [28]	Extreme Machine Learning (EML) with Gaussian Mixture Model (GMM)	74.33% accuracy	2024
Li et al. [29]	Bi-A2CEmo Framework	89.04% accuracy	2024
Wang et al. [30]	Tensor Decomposition Fusion with Self-Supervised Multi-Task Learning	85.52% accuracy	2024
Ours	ExpressNet	92.97% accuracy	2025

Table 4 offers a comparison of current Speech Emotion Recognition (SER) studies together with different approaches and corresponding performance measures. Although earlier works have investigated several architectures, our suggested method surpasses current models with the highest accuracy of 92.97%. Employing a Bi-LG-GNN-based multimodal recognition system, Alsaadowitz and Daş [27] achieved 80% accuracy with an F1-score, precision, and recall of 81%

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

With a Gaussian Mixture Model (GMM), Koti et al. [28] used Extreme Machine Learning (EML) to attain a rather lower accuracy of 74.33%. Li et al. [29] presented a Bi-A2CEmo framework with a greater accuracy of 89.04%; Wang et al. [30] used Tensor Decomposition Fusion with Self-Supervised Multi-Task Learning, obtaining 85.52% accuracy.

Using an enhanced feature extraction pipeline and an optimal deep learning architecture, our methodology beats these current approaches. We improve emotion classification performance by including a hybrid framework that efficiently records both temporal and spectral properties of speech data. Furthermore, our model gains from a better fusion technique including multimodal data representations, hence improving feature discrimination and lowering misclassification. Moreover, the inclusion of an adaptive learning process guarantees optimal model generalization, hence improving accuracy over earlier approaches. These developments together define our model as a modern speech emotion recognition solution.

4.5 Error Analysis

Confusion matrix-based error analysis offers insightful information on the classification performance of the model. The model properly categorized most of the samples in every emotional category in the training set. From *Fig. 9*, in 862 occurrences, angry was appropriately noted; minor misclassifications into Happy and Sad were found. Disgust obtained 774 accurate classifications, suggesting considerable overlap as 52 cases were misclassified as Happy and 78 as Sad. Fear exhibited clear uncertainty with Sad (119 cases), however had 787 accurate predictions. In 864 cases, Happy was correctly labeled; minor misclassifications into Neutral, Disgust and Sad occurred. In 673 cases, Neutral was accurately noted; occasionally it was mistaken with Sad (46 cases), Disgust (15 cases) and Happy (17 cases). Finally, although it was often misclassified as Fear (119 cases) and Disgust (78 cases), demonstrating a notable overlap in emotional expression, Sad was correctly classified in 828 cases.

The model maintained good classification performance in the testing set but showed rather higher misclassification rates. In 392 cases, angry was appropriately categorized; a few were misclassified as Happy, Disgust, Neutral, or Sad. Correctly detected in 300 cases, disgust was misclassified as Happy (19 cases) and Sad (25 cases). In 310 cases, fear was appropriately categorized; in 35 cases it was mistaken with Sad and in 6 cases Happy. With 384 valid predictions, Happy shown strong categorization; few were misclassified into Disgust (3 cases), Fear (1 case), and Sad (1 case). Neutral had 302 accurate classifications, although occasionally it was misclassified as Happy (9 cases) and Sad (17 cases). In 398 cases, Sad was appropriately categorized; yet, she exhibited some uncertainty with Disgust (9 cases), Fear (4 cases), and Neutral (2 cases).

Because of their comparable auditory characteristics, the most often misclassified categories were Fear and Sad, as well as Disgust and Sad. Although the tiny rise in misclassifications on the testing set points to a minor generalization gap, the performance on the training set points to a good learning fit. Techniques include feature enhancement, extra contextual embeddings, and augmentation methodologies could be used to hone the model's capacity to identify minute

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

emotional variances in speech and therefore increase classification accuracy.

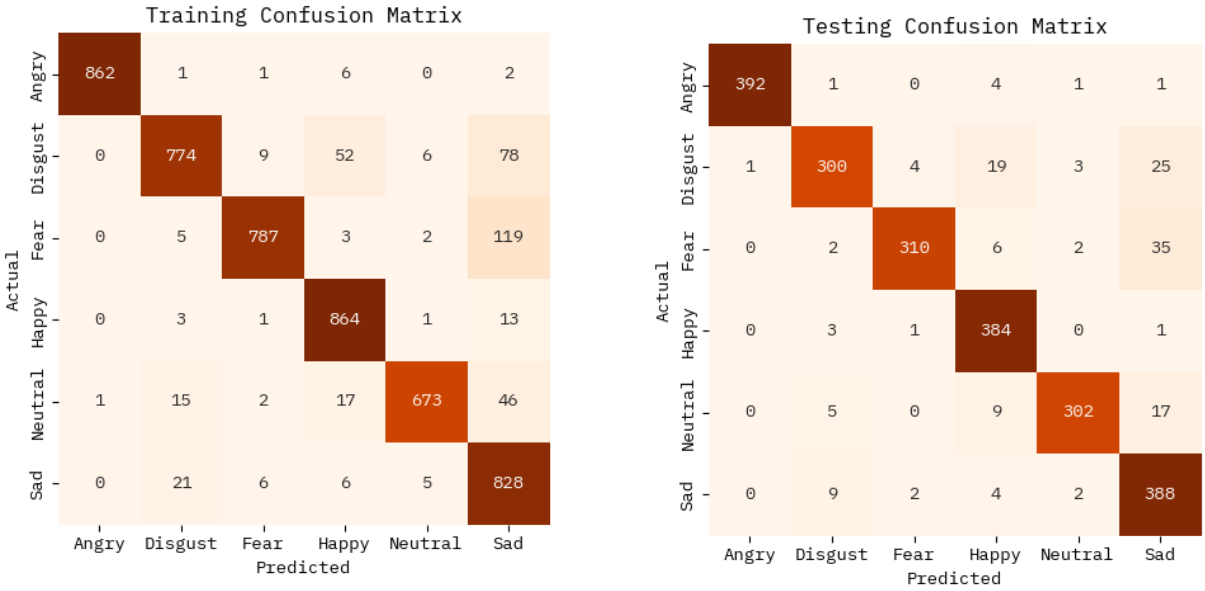


FIG. 9. TRAINING (LEFT) AND TESTING (RIGHT) CONFUSION MATRICES FOR SPEECH EMOTION RECOGNITION

V. CONCLUSION AND FUTURE DIRECTION

In this study, we introduced ExpressNet, which is an optimum MLP-based model for Speech Emotion Recognition (SER). It successfully mixes spectral and prosodic data to achieve high classification accuracy. Our model scored a test accuracy of 92.97% on the CREMA-D dataset, beating other state-of-the-art approaches and demonstrating stable performance. The model was able to identify complicated emotional patterns in speech because of the combination of powerful feature extraction techniques and a deep learning architecture. This makes it appropriate for real-time applications including virtual assistants, healthcare monitoring, and affective computing. The capacity of the model to preserve great accuracy while limiting overfitting demonstrates its potential for use in real-world situations where robustness and efficiency are important.

There are still many areas that could be improved and explored more, even though the results are encouraging. To begin with, the model's performance could be improved by including multimodal input, such as visual and textual signals, in order to disseminate a more complete understanding of emotional states. Multimodal fusion approaches may be useful for capturing minor emotional nuances that are difficult to detect using auditory signals alone. Furthermore, investigating self-supervised learning methods may lessen the need for extensive labeled datasets, which would enhance the model's ability to scale and adapt to many languages and cultural situations. Self-supervised learning could also let the model to learn more broad representations of emotions, which would make it more resilient to changes in speech patterns and environmental variables.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

Cross-lingual and cross-dataset generalization is another important subject that has to be researched in the future. Although our model scored well on the CREMA-D dataset, there is still room for improvement in its performance across multiple languages and datasets. Techniques including data augmentation, transfer learning, and multi-task learning could help the model to generalize over several datasets and languages. For example, transfer learning could allow the model to use knowledge from one language or dataset to improve performance on another language or dataset. Multi-task learning could allow the model to learn multiple related tasks at the same time, such as emotion recognition and speaker identification, which would improve overall robustness.

Additionally, the model's capacity to collect temporal and contextual characteristics in speech signals could be improved by exploring entropy-based approaches and attention mechanisms. These strategies could assist the model better differentiate between emotions that have similar auditory properties, such as fear and sadness, which were identified as hard cases in our error study. Furthermore, real-time optimization methods could be explored in order to further decrease the computational cost of the model, which would make it more appropriate for use on edge computing platforms and low-power devices.

references

[1] K. P. Rao, M. V. P. C. S. Rao, and N. H. Chowdary, “An integrated approach to emotion recognition and gender classification,” *J Vis Commun Image Represent*, vol. 60, pp. 339–345, 2019.

[2] M. Imani and G. A. Montazer, “A survey of emotion recognition methods with emphasis on E-Learning environments,” *Journal of Network and Computer Applications*, vol. 147, p. 102423, 2019.

[3] Y. Xu, H. Su, G. Ma, and X. Liu, “A novel dual-modal emotion recognition algorithm with fusing hybrid features of audio signal and speech context,” *Complex & Intelligent Systems*, vol. 9, no. 1, pp. 951–963, 2023.

[4] C. Mumenthaler, D. Sander, and A. S. R. Manstead, “Emotion recognition in simulated social interactions,” *IEEE Trans Affect Comput*, vol. 11, no. 2, pp. 308–312, 2018.

[5] L. Patenko and O. Ignatenko, “Speech sentiment classification in a multi-lingual environment”.

[6] M. B. SK, P. Bhambu, and M. V. Gupta, “Spoken emotion recognition through human-computer interaction using a novel deep learning technology,” *Multidisciplinary Science Journal*, vol. 5, 2023.

[7] T. Rathi and M. Tripathy, “Analyzing the influence of different speech data corpora and speech features on speech emotion recognition: A review,” *Speech Commun*, p. 103102, 2024.

[8] S. Murugaiyan and S. R. Uyyala, “Aspect-based sentiment analysis of customer speech data using deep convolutional neural network and bilstm,” *Cognit Comput*, vol. 15, no. 3, pp. 914–931, 2023.

[9] S. P. Mishra, P. Warule, and S. Deb, “Speech emotion recognition using mfcc-based entropy feature,” *Signal Image Video Process*, vol. 18, no. 1, pp. 153–161, 2024.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

[10] Y. R. Rochlani and A. B. Raut, "Machine Learning Approach for Detection of Speech Emotions for RAVDESS Audio Dataset," in *2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, IEEE, 2024, pp. 1–7.

[11] Y. Bhanusree, S. S. Kumar, and A. K. Rao, "Time-distributed attention-layered convolution Neural Network with ensemble learning using Random Forest classifier for speech emotion recognition," *Journal of Information and Communication Technology*, vol. 22, no. 1, pp. 49–76, 2023.

[12] A. A. Anthony and C. M. Patil, "Speech emotion recognition systems: A comprehensive review on different methodologies," *Wirel Pers Commun*, vol. 130, no. 1, pp. 515–525, 2023.

[13] A. A. Anthony and C. M. Patil, "Speech emotion recognition systems: A comprehensive review on different methodologies," *Wirel Pers Commun*, vol. 130, no. 1, pp. 515–525, 2023.

[14] E. K. Zadeh and M. Alaeifard, "Adaptive Virtual Assistant Interaction through Real-Time Speech Emotion Analysis Using Hybrid Deep Learning Models and Contextual Awareness," *International Journal of Advanced Human Computer Interaction*, vol. 1, no. 1, pp. 1–15, 2023.

[15] G. Alhussein, M. Alkhodari, A. H. Khandoker, and L. J. Hadjileontiadis, "Novel speech-based emotion climate recognition in peers' conversations incorporating affect dynamics and temporal convolutional neural networks," *IEEE Access*, 2025.

[16] B. T. Atmaja and A. Sasou, "Sentiment analysis and emotion recognition from speech using universal speech representations," *Sensors*, vol. 22, no. 17, p. 6369, 2022.

[17] C. Dixit and S. M. Satapathy, "Deep CNN with late fusion for real time multimodal emotion recognition," *Expert Syst Appl*, vol. 240, p. 122579, 2024.

[18] D. Mamieva, A. B. Abdusalomov, A. Kutlimuratov, B. Muminov, and T. K. Whangbo, "Multimodal emotion detection via attention-based fusion of extracted facial and speech features," *Sensors*, vol. 23, no. 12, p. 5475, 2023.

[19] H. Zhao, N. Huang, and H. Chen, "Knowledge enhancement for speech emotion recognition via multi-level acoustic feature," *Conn Sci*, vol. 36, no. 1, p. 2312103, 2024.

[20] S. P. Mishra, P. Warule, and S. Deb, "Variational mode decomposition based acoustic and entropy features for speech emotion recognition," *Applied Acoustics*, vol. 212, p. 109578, 2023.

[21] J. Singh, L. B. Saheer, and O. Faust, "Speech emotion recognition using attention model," *Int J Environ Res Public Health*, vol. 20, no. 6, p. 5140, 2023.

[22] M. Khan, A. El Saddik, F. S. Alotaibi, and N. T. Pham, "AAD-Net: Advanced end-to-end signal processing system for human emotion detection & recognition using attention-based deep echo state network," *Knowl Based Syst*, vol. 270, p. 110525, 2023.

[23] A. Khan, "Improved multi-lingual sentiment analysis and recognition using deep learning," *J Inf Sci*, p. 01655515221137270, 2023.

[24] Z.-T. Liu, M.-T. Han, B.-H. Wu, and A. Rehman, "Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning," *Applied Acoustics*, vol. 202, p. 109178, 2023.

[25] S. P. Mishra, P. Warule, and S. Deb, "Speech emotion recognition using mfcc-based entropy feature," *Signal Image Video Process*, vol. 18, no. 1, pp. 153–161, 2024.

[26] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans Affect Comput*, vol. 5, no. 4,

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

pp. 377–390, Oct. 2014, doi: 10.1109/TAFFC.2014.2336244.

[27] H. F. T. Alsaadawı and R. Daş, "Multimodal Emotion Recognition Using Bi-LG-GCN for MELD Dataset," *Balkan Journal of Electrical and Computer Engineering*, vol. 12, no. 1, pp. 36–46, 2024.

[28] V. M. Koti, K. Murthy, M. Suganya, M. S. Sarma, G. V. S. S. S. Kumar, and N. Balamurugan, "Speech Emotion Recognition using Extreme Machine Learning," *EAI Endorsed Transactions on Internet of Things*, vol. 10, 2024.

[29] H. Li, X. Zhang, S. Duan, and H. Liang, "Speech emotion recognition based on bi-directional acoustic-articulatory conversion," *Knowl Based Syst*, p. 112123, 2024.

[30] R. Wang, J. Zhu, S. Wang, T. Wang, J. Huang, and X. Zhu, "Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking," *Int J Multimed Inf Retr*, vol. 13, no. 4, p. 39, 2024.