Babylonian Journal of Machine Learning
Vol.2026, **pp**. 333–350
DOI: https://doi.org/10.58496/BJML/2026/019; ISSN: 3006-5429
https://mesopotamian.press/journals/index.php/BJML

Research Article

Speech Emotion Recognition: A Multi-Dataset Experimental Study

Suhad Al-Shoukry [1], Rihab Habeeb Sahib [2], Wurood Mahdi Sahib [3], Abdulkadhem A. Abdulkadhem [4], Mohamed Shabbir Abdulnabi [5, *]

[1] *Department of Software and Artificial Intelligence, Najaf Polytechnic College, Al-Furat Al-Awsat Technical University, Najaf, Iraq.*

[2] *College of Arts, University of Babylon, Babylon, 51002, Iraq.*

[3] *Department of Physiology and Medical Physics, College of Medicine, University of Babylon, Babylon, Iraq.*

[4] *Department of Artificial Intelligence, College of Science, Al-Mustaqbal University, Babylon, Hillah 51001, Iraq.*

[5] *School of Technology, Asia Pacific University of Technology and Innovation (APU), 57000 Kuala Lumpur, Malaysia*

ARTICLE INFO

Article History

Received 06 Sep 2026
Revised 15 Oct 2026
Accepted 11 Nov 2026
Published 17 Nov 2026

Keywords

Speech emotion recognition,

Acoustic feature extraction,

CNN,

BiLSTM.

Class-wise performance evaluation.

ABSTRACT

Speech emotion recognition involves identifying emotions from speech recordings. When some emotions are more frequent than others in a dataset, overall accuracy can mask two kinds of errors: missing an emotion when it is present and predicting it when it is absent. In this study, we evaluated the Enhanced Speech Emotion Recognition Framework (ESERF) on eight emotion categories using recordings from RAVDESS, CREMA-D, and TESS. We assessed performance across three experimental scenarios using overall accuracy, confusion matrices, and class-wise precision, recall, and F1-score. Overall accuracy was 60.74%, 61.02%, and 49.22% in Scenarios 1, 2, and 3, respectively. Although the first two scenarios had similar overall accuracy, recall for Calm increased from 0.80 to 0.87, while precision fell from 0.64 to 0.44. Scenario 2 therefore identified a larger proportion of the Calm recordings, but fewer than half of the recordings it classified as Calm carried that label. The decline in recall from Scenario 2 to Scenario 3 also varied by emotion: recall fell by about 25 percentage points for both Fear and Happy, compared with only two percentage points for Sad.

1. INTRODUCTION

Speech emotion recognition (SER), as illustrated in Fig. 1, is an attempt at identifying emotions based on speech using signal processing and machine learning approaches [1]. This technique draws from ideas of artificial intelligence, psychology, and computational linguistics and has possible uses in the realm of human-computer interaction and affective computing. However, identifying emotions from speech is still challenging because of the variation in emotion expression across different speakers and recording conditions [2], [3].

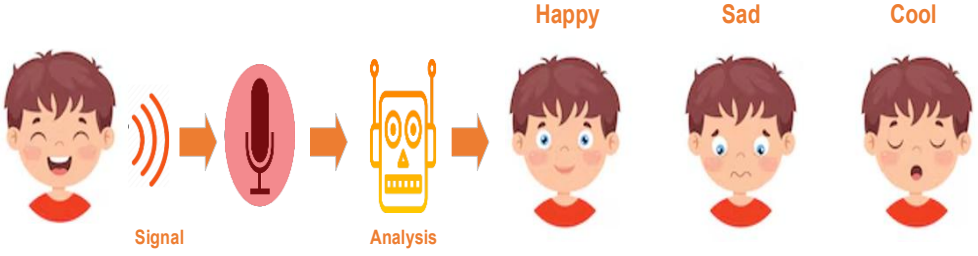


Fig. 1. Typical SER System.

*Corresponding author. Email: mohamed.shabbir@apu.edu.my

Even though there have been advances in SER, its robust implementation across various speakers and recording conditions is still difficult to accomplish. Emotions are expressed via cues such as pitch, rate of speaking, and energy, and the presence of background noise and poor recording may hinder their detection. Therefore, detecting features capable of differentiating emotions that look similar is one of the key aspects of developing SER models [4]. Moreover, variations in language, accent, speaker properties, and class distribution may limit generalization beyond training data [5]. The current research examines the effect of overlapping features on category distinguishability and computational resources that may restrict applications requiring instantaneous reaction. It is possible for two different classifiers to have the same level of overall labeling accuracy but commit different kinds of errors, where one classifier will fail to recognize an emotion if it is there, and another one will assign an emotion to unrelated recordings easier than needed. Hence, the research will examine how the error profile changes among emotion categories when overall accuracy remains the same.

The research contributions can be stated as follows:

(i) Proposal of the ESERF framework that can be used for the classification of emotional speech.

(ii) Eight-category evaluation of the system using recordings from RAVDESS, CREMA-D, and TESS datasets.

(iii) Comparison of class-wise precision, recall, and F1-scores in three different scenarios and error patterns that cannot be discovered based on overall accuracy.

The current research deals with the classification of emotions from speech. In particular, it is focused on the differences in recognition performance for each emotion category. The rest of the paper is structured as follows. Section 2 reviews related works. Section 3 describes the proposed framework, and Section 4 its implementation. Experimental results are presented in Section 5, followed by the discussion of them. Conclusion and future works are provided in Section 6.

2. LITERATURE REVIEW

Speech emotion recognition is an area of affective computing that aims to identify emotional states from speech. Research in this field explores signal-processing techniques, feature representations, and machine learning architectures for emotion classification. Related work also examines emotion and sentiment analysis in other modalities, as well as the robustness of speech-processing systems. Table 1 summarizes selected studies, their methods, datasets, and reported findings .

TABLE I: SUMMARY OF PRIOR STUDIES IN THE CONTEXT OF SER SYSTEM DEVELOPMENT.

Study Author(s)	Year	Methodology	Database Used	Evaluation Criteria	Results (Y)
[9]	2012	Probabilistic word Sentiment Polarity Disambiguation Model	Opinion corpus (mobile phone, digital camera domains)	Accuracy	Mobile: 74.10, Camera: 70.10
[10]	2015	Bayesian model	Opinmine corpus v2	Accuracy	Average: 80.80
[11]	2015	Vocal data mining	Collected data	Recognition rate	Average: 82.43
[12]	2017	Recurrent neural network	IEMOCAP	Accuracy, Unweighted Average Recall	Accuracy: 64.78 , UAR: 60.89
[13]	2018	Transfer learning	Multiple datasets (Literary tales, Election tweets, ISEAR, Headlines, General tweets)	Sensitivity, Specificity	Varied by dataset; e.g. Sensitivity: 68.4 , Specificity: 91.4 (Literary tales)
[3]	2019	Multi-layer perceptron	Emolnt-2017, EmoBank, SemEval-2016	Accuracy	Emolnt-2017: 89.58, EmoBank: 0.635, SemEval- 2016: 57.11
[1]	2020	3D convolutional neural networks	MSP-IMPROV; IEMOCAP	Accuracy	Accuracy 62.60 for IEMOCAP, 55.70 For MSP-IMPROV
[6]	2025	Robust feature attenuation with clean-label poison-only backdoor attack	Speech recognition benchmarks (not explicitly named)	Attack Success Rate, Stealthiness, Defense Bypass (e.g., STRIP, pruning)	>90% success rate; stealthy against multiple defenses

Study Author(s)	Year	Methodology	Database Used	Evaluation Criteria	Results (Y)
[8]	2025	CA-ISS algorithm exploiting resampling and normalization-based vulnerabilities	Six commercial speech systems	Generalizability, Semantic Consistency, Camouflage Effect	High generalization; effective against cloud platforms; cognitive differences exploited
[7]	2025	Random Spectrogram Rhythm Transformation (RSRT) for stealthy poisoning	Two ASR and speaker verification datasets	Stealthiness (human & model), Poisoning Rate, Attack Success	High success with low poisoning rate; imperceptible to human listeners

The studies in Table 1 differ in both the information available to the model and the outcome being predicted. For example, identifying sentiment in written text does not require a model to distinguish emotions from vocal cues. The reported percentages should therefore be interpreted within each study's task rather than used to rank the methods across the table. Recent developments in SER systems have changed their focus to transformer-based architectures, which integrate local convolutional feature representations with global temporal modelling and attention. For example, Tang et al. (Speech Communication, 2025) [14] proposed a CNN-Transformer network having a multidimensional attention mechanism that takes into account the time, channel, and space relationships. It was state-of-the-art on IEMOCAP and Emo-DB, using CNN-based local encoders with lightweight Transformer blocks to learn dependency globally. In addition, Khan et al. (2025) [15] suggested a multimodal cross-modal Transformer, a multimodal system (HuBERT audio features and BERT text features) that combines the two using a cross-modal Transformer (CMT). Their system had reported UW-Acc = 81.3% and W-Acc = 91.8% on ESD corpora and IEMOCAP and exhibited better generalization and cross-modal fusion capacity. Moreover, SwinTSER by Akinpelu et al. (2025) is a better bilingual Transformer model with Shift Window Transformers (SWT) hierarchical feature representation, but achieved 98.3% accuracy on TESS and 66% on hybrid bilingual TESS_EMOVO datasets, and progress in the field of cross-lingual SER [16]. Also, a Multitask Transformer framework of cross-corpus SER was proposed by Ahn et al. (2025) [17] with auxiliary tasks of contrastive learning and information maximization. The multitask design improves cross-domain adaptation across IEMOCAP, MSP-IMPROV, and EMO-DB with a higher level of performance under restricted target labels.

This study investigates a framework combining feature processing with CNN and bidirectional LSTM (BiLSTM) components for speech emotion classification. The reported evaluation uses data from RAVDESS, CREMA-D, and TESS and examines differences in classification performance across emotion categories. Using multiple source datasets broadens the data represented in the study, but does not by itself establish generalisation to unseen speakers, languages, or datasets. Real-time suitability and computational efficiency require separate evaluation and remain considerations for future work.

3. PROPOSED SYSTEM

This section describes the ESERF framework for speech emotion recognition, including feature processing and classification. Table 2 summarizes the reported hyperparameter configurations and architectural settings. The scope of the present evaluation is emotion classification; security properties are not established by the reported experiments.

TABLE II: HYPERPARAMETER CONFIGURATION AND ARCHITECTURAL SETTINGS FOR THE PROPOSED HYBRID CNN–BiLSTM MODEL.

Parameter	Configuration / Value
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1e\text{-}8$)
Initial Learning Rate	1e-3 with ReduceLROnPlateau (factor = 0.5, patience = 5, min_lr = 1e-6)
Batch Size	32
Max Epochs	100
Early Stopping	Patience = 10 (based on validation loss)
Regularization	Dropout = 0.3 (after BiLSTM and Dense layers)
CNN Architecture	Conv1D(64, k = 5) → MaxPool(2) → Conv1D(128, k = 3) → GlobalAvgPool1D; BatchNorm after each Conv
BiLSTM Units	128
Dense Head	Dense (128, ReLU) → Softmax (C classes)

Parameter	Configuration / Value
PCA	Retain 95% variance (typical k ∈ [32, 128])
Weight Initialization	He Normal (CNN/Dense), Glorot Uniform (LSTM)
Data Augmentation (Optional)	Time shift ±10 ms; Gain ±3 dB
Random Seed	42

Furthermore, Figure 2 illustrates the ESERF pipeline beginning with TQWT-based feature decomposition, followed by the TSP for feature transformation and PCA-based dimensionality reduction. Two parallel branches of CNN for spatial feature learning and BiLSTM for temporal sequences modeling are merged through concatenation and passed to dense layers for final emotion classification via a SoftMax output layer.

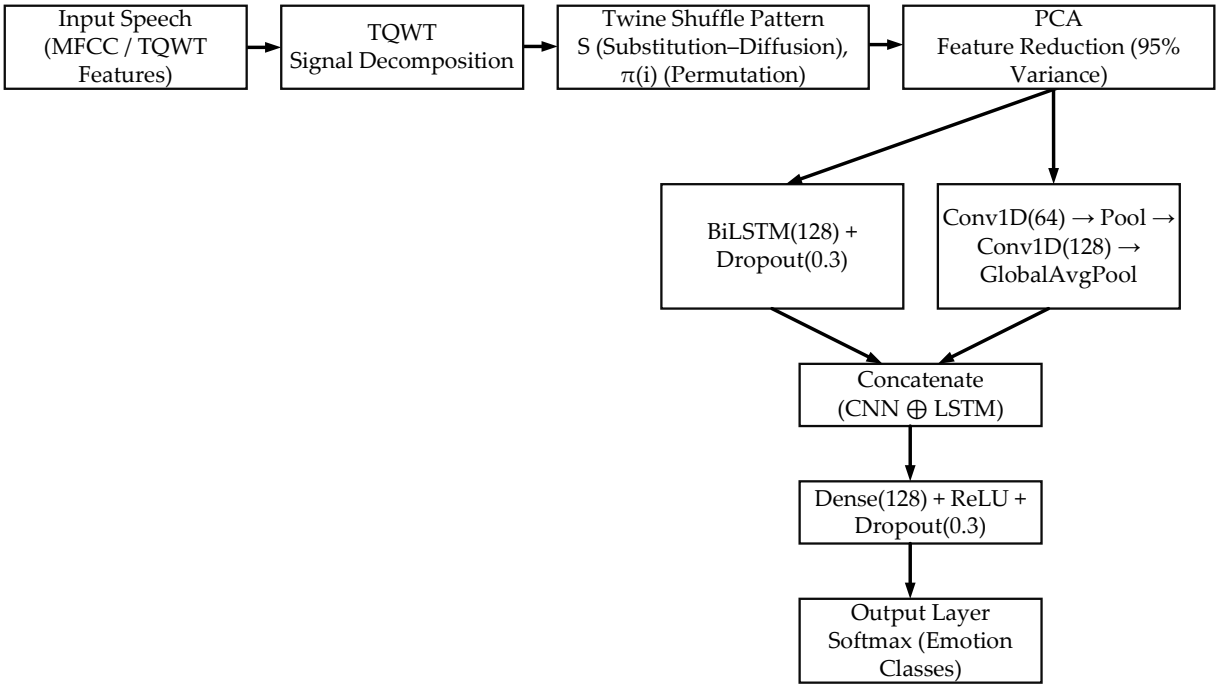


Fig 2. Overall architecture of the proposed ESERF framework illustrating the feature-processing and classification stages.

- TSP for Feature Transformation

TSP transforms the coefficients obtained from TQWT into a feature representation. Equation (1) describes this transformation using the coefficient permutation function and the mapping function S.

$$F_{\text{twine}}(i) = S\left(\text{signum}\left(W_j(i) - W_j(\pi(i))\right)\right) \qquad \mathbf{1}$$

Here,

(i) $S(\cdot)$ represents a cryptographic shuffle box, which performs nonlinear substitution and diffusion operations inspired by lightweight ciphers such as Twine-80. It maps the input signum values into key-dependent substitution outputs.
(ii) $\pi(i)$ is a permutation function that defines the shuffling order of coefficients.
(iii) The signum function serves as a binary comparator, where $\text{signum}(x) = 1$ if $x \geq 0$ and 0 otherwise.

The transformation is discussed here as a feature-processing operation. The present study does not establish cryptographic confidentiality, resistance to feature reconstruction, or protection against tampering.

- PCA for Dimensionality Reduction

PCA is applied to the feature set F to reduce dimensionality while retaining most of the variance. If X is the matrix representing F with zero mean, the covariance matrix C is computed as in Eq. (2):

$$C = \frac{1}{m} X X^T \quad \textbf{2}$$

The following equations describe the feature-processing and classification components of ESERF, including TQWT, TSP, PCA, and LSTM :

- TQWT decomposes a signal $x(n)$ into subbands with coefficients $W_j(n)$, where j indicates the level of decomposition. The Q-factor and redundancy parameters control the bandwidth and overlap between subbands, respectively. The decomposition can be represented as in Eq. (3):

$$x(n) = \sum_{j=1}^{J} W_j(n) + R_J(n) \quad \textbf{3}$$

where $R_J(n)$ is the residual after J levels of decomposition.

- Equation (4) expresses the TSP transformation of the coefficients obtained from TQWT :

$$F_{\text{twine}}(i) = S\left(\text{signum}\left(W_j(i) - W_j(\pi(i))\right)\right) \quad \textbf{4}$$

where $\pi(i)$ is a

permutation function, and $\text{signum}(\cdot)$ is defined as 1 if $W_j(i) \geq W_j(\pi(i))$ and 0 otherwise.

- PCA is applied to the feature set F to reduce dimensionality while retaining most of the variance. If X is the matrix representing F with zero mean, the covariance matrix C is given by $C = \frac{1}{m} X X^T$. The eigenvectors of C, v_i, corresponding to the largest eigenvalues, form the principal components. The reduced feature set F_{PCA} is obtained by projecting F onto the selected eigenvectors as in Eq. (5):

$$F_{\text{PCA}} = X v_i \quad \textbf{5}$$

For LSTMs, the key equations governing the state updates are in Eq. (6):

$$\begin{gathered} f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \\ i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\ C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \\ o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o),\ h_t = o_t * \tanh(C_t) \end{gathered} \quad \textbf{6}$$

where f_t, i_t, and o_t are the forget, input, and output gates, respectively; C_t and h_t are the cell state and hidden state; W s and b s are the weights and biases; and σ and tanh are the sigmoid and hyperbolic tangent activation functions, respectively.

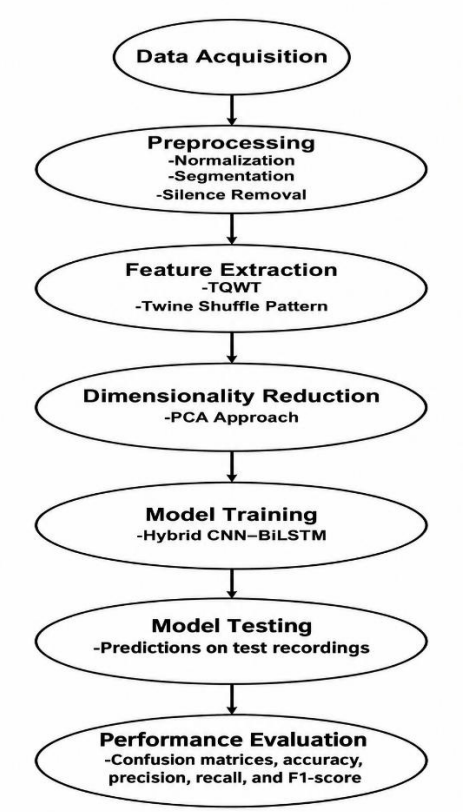


Fig. 3. The Flow Chart of ESERF SER System.

4. IMPLEMENTATION

The implementation comprises data acquisition, preprocessing, feature extraction, and emotion classification. The study uses speech recordings from RAVDESS, TESS, and CREMA-D. Table 3 summarizes the source datasets. Figures 5 and 6 show example waveforms and spectrograms for selected emotion categories, illustrating their amplitude and time–frequency characteristics.

TABLE III: UTILIZED DATASETS RAVDESS, TESS, AND CREMA-D INFORMATION: NO. OF SPEAKERS, NO. OF EMOTIONS, LANGUAGE DURATION, SAMPLING RATE, ANNOTATION METHOD, AND YEAR.

Dataset Name	No. of Speakers	No. of Emotions	Language	Duration	Sampling Rate	Annotation Method	Year
CREMA-D	91	6	English	~10hr	44.1kHz	Crowd-sourced	2018
RAVDESS	24	8	English	1.5hr	48kHz	Actor-reported	2016
TESS	2	7	English	~1hr	16kHz	Actor-reported	2005

Figure 4 summarizes the implementation workflow, from data collection and preprocessing to feature extraction, model training, and evaluation.

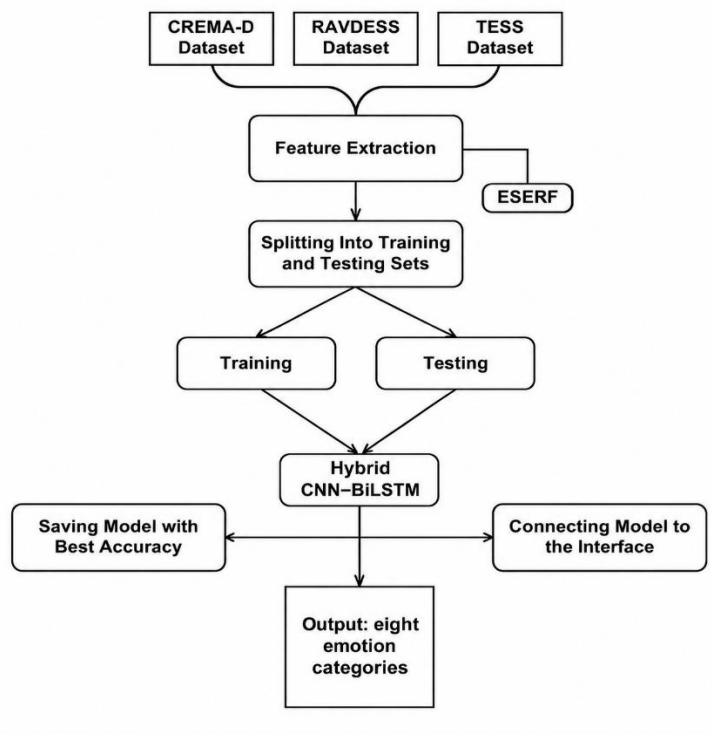


Fig. 4. ESERF System Architecture.

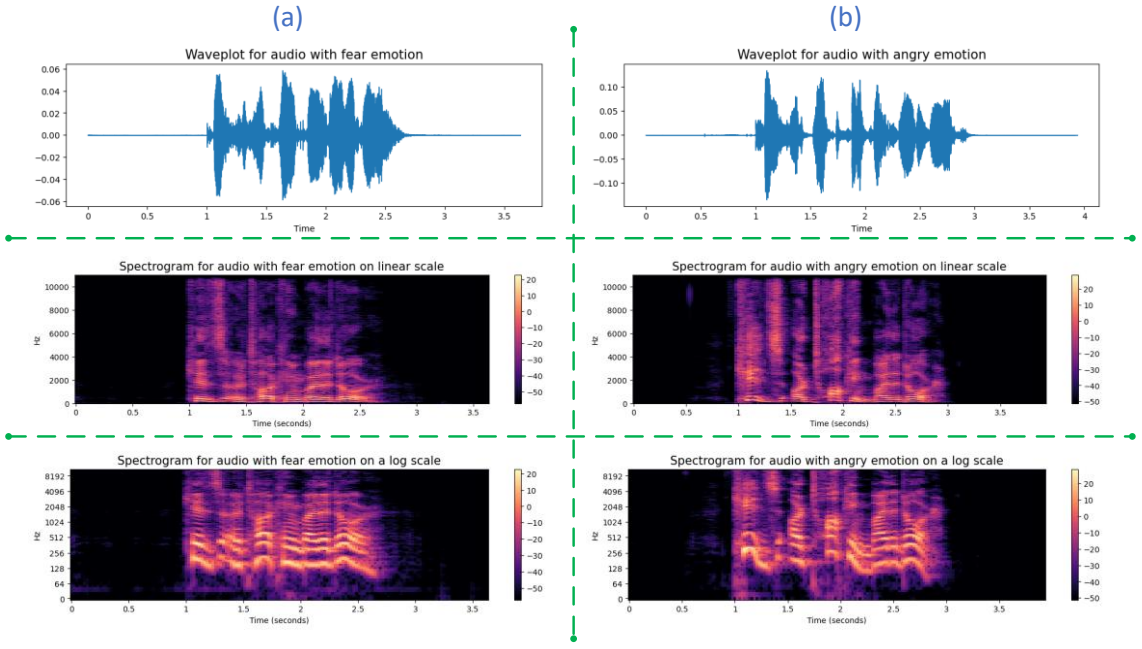


Fig. 5. Audio Wave Plot with Associated Spectrogram of: (a) Fear and (b) Angry Emotions.

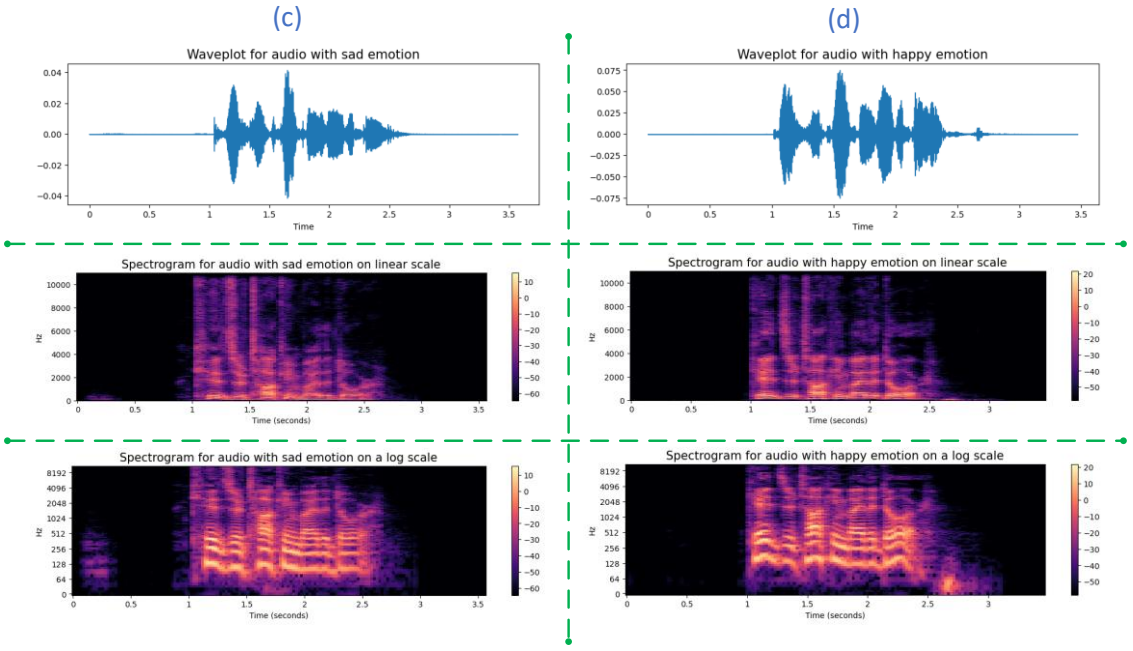


Fig. 6. Audio Wave Plot with Associated Spectrogram of: (a) Sad and (b) Happy Emotions.

The waveform and the spectrogram offer different views of each recording. In particular, the waveform shows temporal changes in amplitude, while the spectrogram represents signal energy distribution in frequency over the same period of time. As can be seen from the examples, there are changes within one utterance but there is no evidence that one of the visual patterns is specific to a certain emotion.

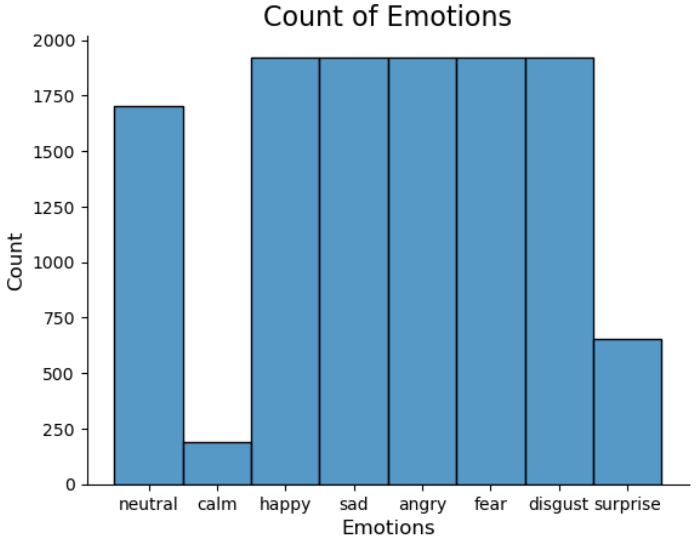


Fig. 7. Distribution of Emotions in the Dataset.

There is an imbalance in emotions distribution because Calm and Surprise are presented less often than the rest of the categories. It should be considered when analyzing overall accuracy along with precision, recall, and F1-score of each class.

4.1 Data Acquisition and Preprocessing

4.1.1 Data Sources

The datasets used in this research are RAVDESS, CREMA-D, and TESS [18] – [20]. Therefore, the choice of datasets was made to cover significant variety of emotions, speech content, and speakers in order to train and test the system of Speech Emotion Recognition (SER).

4.1.2 Preprocessing Steps

Preprocessing is an integral part of the process and consists of a series of techniques aimed at standardizing the audio signal for further processing. In particular:

- Normalization: All audio files are normalized to have a constant playback level, which helps reduce the impact of amplitude changes on feature extraction.
- Segmentation: The audio signal is divided into small frames ranging from 20 to 40 milliseconds to analyze the temporal dynamics of speech relevant to emotional expressions.
- Silence Removal: Silence detection algorithms are employed to remove non-informative segments, focusing the analysis on segments with vocal content.

4.1.3 Combining Preprocessed Data

All data from the source datasets have been used together for the experiments described above. Various factors like differences in recording conditions, speakers' characteristics, and emotions' annotation should be considered while interpreting results based on the combined dataset.

4.1.4 Data Splitting

The combined data were divided into training and test subsets. The training subset was used to fit the model, while the test subset was used to calculate the reported classification metrics.

4.1.5 Hybrid CNN–BiLSTM Model Training

Figure 2 depicts the proposed ESERF classifier as two parallel branches: a convolutional neural network (CNN) for extracting local feature patterns and a bidirectional long short-term memory network (BiLSTM) for modelling temporal relationships. The outputs of the two branches are concatenated and passed to a dense classification layer, followed by a softmax output over the eight emotion categories.

4.1.6. Class Imbalance Handling

The dataset exhibited a significant class imbalance (approximately 10:1, with Calm =192 samples vs. other emotion classes = 1,923 samples). To mitigate bias toward majority classes, we applied class-weighted cross-entropy loss, where each class weight $w_c=N/(K\times n_c)$ depends on the class frequency n_c. In addition, we adopted balanced mini-batch sampling to ensure that each batch contained proportional samples from all classes. Minor oversampling of the Calm class and light data augmentation (time shift ±10 ms, gain ±3 dB) were also used to improve minority representation. Model performance was evaluated using macro-F1, balanced accuracy, and AUROC metrics to ensure fairness across all emotion classes.

4.2. Feature Extraction and Engineering

4.2.1 Tunable Q Wavelet Transform (TQWT)

TQWT is used as it is able to subdivide a signal into sub-bands at distinct resolutions which is essential in attaining corresponding frequency property of different emotions. In the TQWT setup, the custom parameters of Q-factor and redundancy are used to find the optimal balance between the narrow band and interleaving of the sub-bands needed to follow the constantly changing spectral-temporal representation of the speech signal.

4.2.2 Twine Shuffle Pattern (TSP)

This novel property transformation algorithm uses a shuffle-based feature transformation on the coefficients that have been acquired through TQWT that takes into account a nonlinear pattern change which consequently enhances the uniqueness of the features. The physical shuffle relies on the permutation function $\pi(i)$, that is, it rearranges the indices between the coefficients to repeat the way the emotional language is manifested in a particularly different way.

4.2.3 Principal Component Analysis (PCA)

PCA reduces the dimensionality of the extracted feature representation by projecting it onto principal components associated with the largest eigenvalues of the feature covariance matrix. This produces a lower-dimensional representation that retains a selected proportion of the variance.

4.2.4 System Configuration and Experimental Environment

The speech emotion recognition experiments used recordings from RAVDESS, CREMA-D, and TESS. Preprocessing, model training, and evaluation were performed on a workstation equipped with an Intel Core i9-10900K processor, 64 GB of RAM, and an NVIDIA GeForce RTX 3090 GPU with 24 GB of memory. The workstation ran 64-bit Ubuntu 20.04 LTS. The classifier was implemented in Python 3.8 using PyTorch 1.13.0. We used librosa to process the speech recordings and scikit-learn to calculate class-wise precision, recall, and F1-scores.

4.2.5 Performance Metrics

Model performance was evaluated using accuracy, precision, recall, and F1-score, as defined in the following equations :

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{9}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

Accuracy measures the proportion of correctly classified samples. For each emotion category, precision measures the proportion of predictions assigned to that category that are correct, while recall measures the proportion of actual samples in that category that are correctly identified. F1-score is the harmonic mean of precision and recall. Class-wise metrics complement overall accuracy by showing differences in performance across emotion categories.

Support is the number of evaluation examples carrying a given reference label. In Tables 4–6, the reported support is 137 for Calm and 1,468 for Disgust. Consequently, one additional correct prediction changes Calm recall by approximately 0.73 percentage points, compared with approximately 0.07 percentage points for Disgust. The same change in the number of correct predictions therefore has a different effect on the reported recall.

5 RESULTS AND DISCUSSION

This section presents the reported classification results for the three experimental scenarios using data from RAVDESS, CREMA-D, and TESS. Performance is examined using overall accuracy, class-wise metrics, confusion matrices, and learning curves.

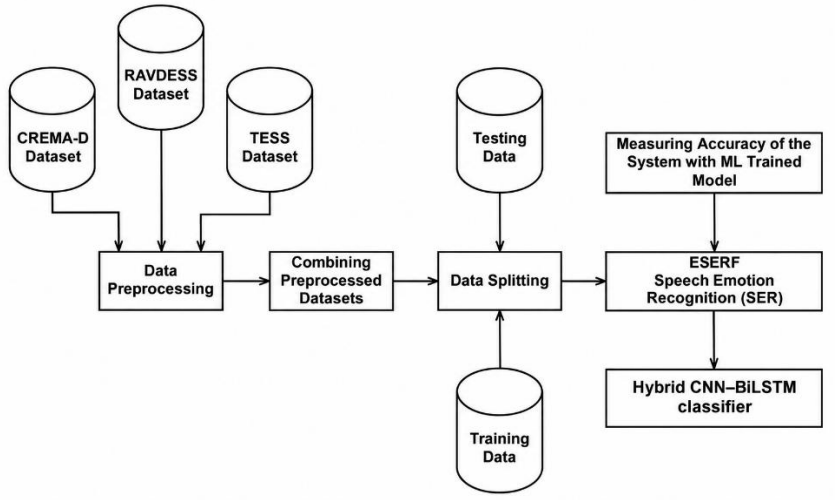


Fig. 8. Dataflow of the Proposed ESERF Speech Emotion Recognition System.

5.1. Model Performance

Overall accuracy was approximately 61% in Scenarios 1 and 2 and 49% in Scenario 3. These results indicate differences in reported performance across the experimental scenarios; they do not establish statistically significant differences.

- Training Dynamics: The training loss progressively reduced with the epoch in all scenarios, with Scenario 1, Scenario 2 and Scenario 3 recording a final training loss of 0.25, 0.23 and 0.35 respectively. Nevertheless, the testing loss was still significantly larger indicating some overfitting, especially in Scenario 3.
- Classification Performance: Classification performance was examined using the confusion matrices and class-wise precision, recall, and F1-scores:

Tables 4–6 report precision, recall, F1-score, and support for each emotion in Scenarios 1, 2, and 3, respectively. Each emotion has the same reported support across the three scenarios, although equal counts alone do not establish that the same recordings were evaluated. In Scenario 1, recall was similar for Surprise (0.83) and Calm (0.80), but precision was higher for Surprise (0.85) than for Calm (0.64). Both emotions received recall equal to 0.87 in Scenario 2, but their precision differed — 0.77 for Surprise and 0.44 for Calm. This means that the classifier successfully classified many instances of Calm emotion but made less than half of all its Calm predictions accurate. In the case of Fear, precision was 0.76, while recall was 0.26, which means that almost three-fourths of all Fear instances were not detected by the algorithm in Scenario 3.

TABLE IV: CLASS-WISE PRECISION, RECALL, F1-SCORE, AND SUPPORT FOR SCENARIO 1.

Emotion	Precision	Recall	F1-Score	Support
Angry	0.73	0.72	0.73	1438
Calm	0.64	0.8	0.71	137
Disgust	0.53	0.52	0.52	1468
Fear	0.57	0.54	0.56	1424
Happy	0.56	0.59	0.58	1462
Neutral	0.56	0.55	0.55	1310
Sad	0.59	0.64	0.61	1400
Surprise	0.85	0.83	0.84	483

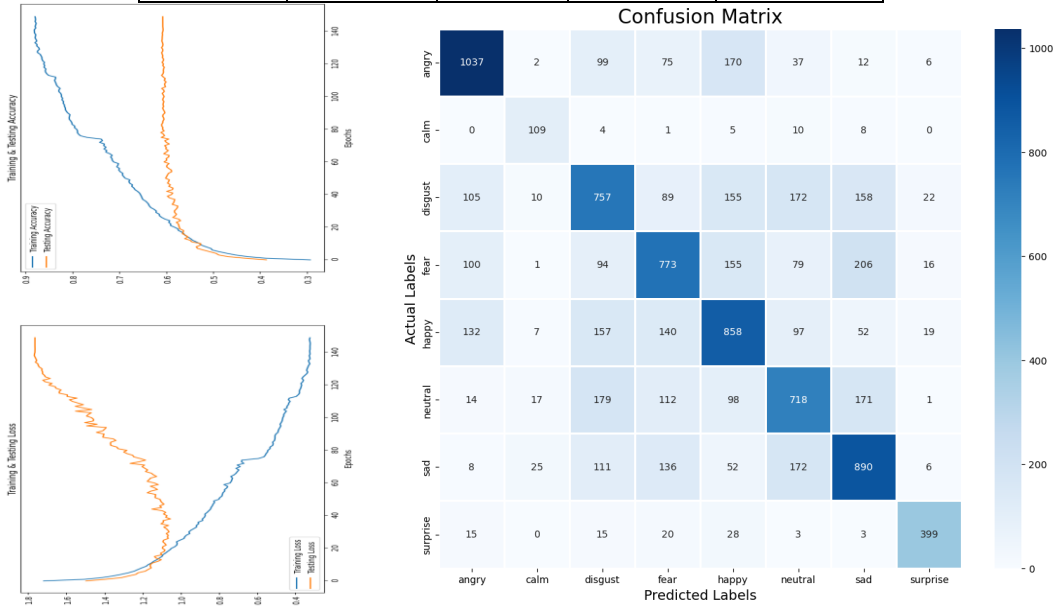


Fig. 9. Illustration of Performance Evaluation of ESERF Speech Emotion Recognition System: Confusion Matrix and Training/Testing Graphs for Scenario 1.

TABLE V: CLASS-WISE PRECISION, RECALL, F1-SCORE, AND SUPPORT FOR SCENARIO 2.

Emotion	Precision	Recall	F1-Score	Support
Angry	0.79	0.73	0.76	1438
Calm	0.44	0.87	0.58	137
Disgust	0.57	0.46	0.51	1468
Fear	0.65	0.51	0.57	1424

Emotion	Precision	Recall	F1-Score	Support
Happy	0.59	0.56	0.57	1462
Neutral	0.57	0.57	0.57	1310
Sad	0.51	0.73	0.6	1400
Surprise	0.77	0.87	0.82	483

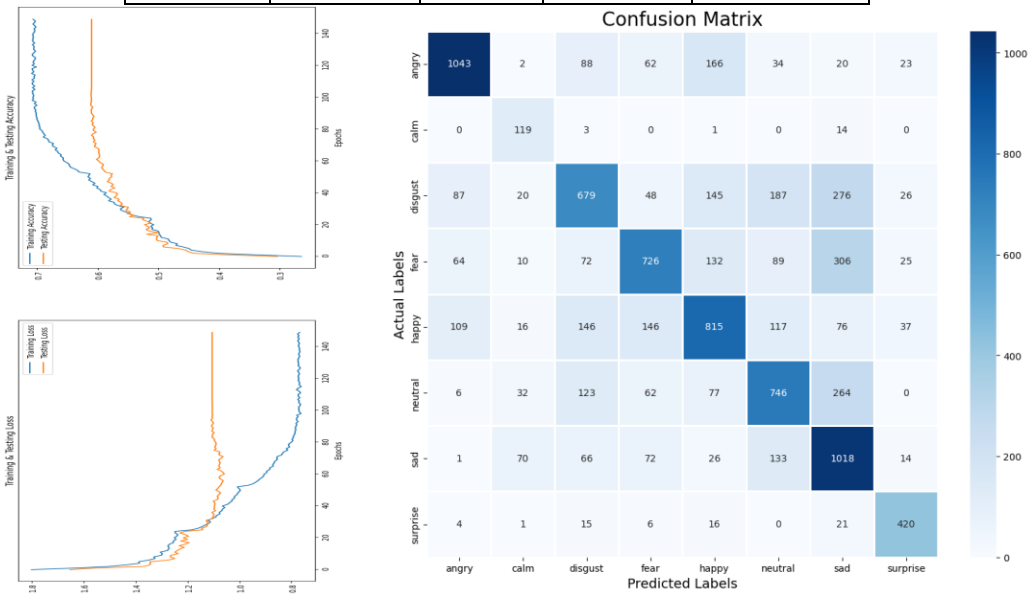


Fig. 10. Illustration of Performance Evaluation of ESERF Speech Emotion Recognition System: Confusion Matrix and Training/Testing Graphs for Scenario 2.

The weighted-average F1-score was approximately equal to 0.48 in Scenario 3. As it can be seen from Table 6, Calm had the recall of 0.83, but precision of only 0.26, which means that there were a lot of false-positive predictions for this emotion. The other emotions that were analyzed, namely Angry, Disgust, and Fear, got 0.66, 0.38, and 0.39 F1-scores respectively. Confusion matrix and learning curves for Scenario 3 are presented in Figure 11 below.

TABLE VI: CLASS-WISE PRECISION, RECALL, F1-SCORE, AND SUPPORT FOR SCENARIO 3.

Emotion	Precision	Recall	F1-Score	Support
Angry	0.65	0.68	0.66	1438
Calm	0.26	0.83	0.4	137
Disgust	0.37	0.38	0.38	1468
Fear	0.76	0.26	0.39	1424
Happy	0.43	0.31	0.36	1462
Neutral	0.51	0.52	0.51	1310
Sad	0.47	0.71	0.57	1400
Surprise	0.51	0.7	0.59	483

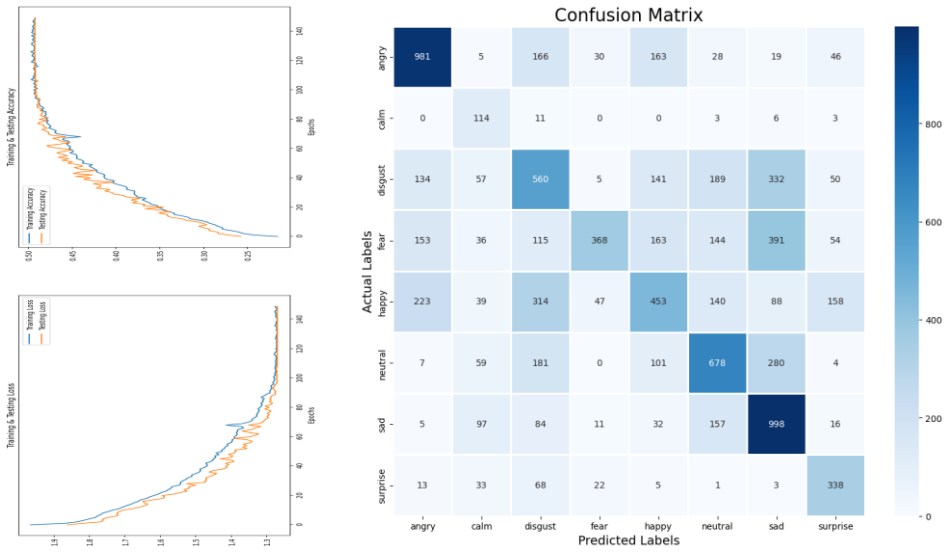


Fig. 11. Illustration of Performance Evaluation of ESERF Speech Emotion Recognition System: Confusion Matrix and Training/Testing Graphs for Scenario 3.

For each emotion, one-versus-rest accuracy was calculated as the sum of true positives and true negatives divided by the 9,122 evaluation examples in that scenario. Because this measure counts correctly excluding an emotion as well as correctly detecting it, it should be interpreted alongside the class-wise precision and recall reported in Tables 4–6.

TABLE VII: ONE-VERSUS-REST ACCURACY FOR EACH EMOTION, CALCULATED FROM THE CONFUSION MATRICES IN FIGURES 9–11.

	Emotion	Accuracy_S1	Accuracy_S2	Accuracy_S3
0	Angry	0.915	0.927	0.891
1	Calm	0.990	0.981	0.962
2	Disgust	0.850	0.857	0.798
3	Fear	0.866	0.880	0.872
4	Happy	0.861	0.867	0.823
5	Neutral	0.873	0.877	0.858
6	Sad	0.877	0.851	0.833
7	Surprise	0.983	0.979	0.948

5.1.1. Explainability Analysis Using SHAP

Figure 12 presents the SHAP (Shapley Additive Explanations) plots [21] for the training and testing stages of ESERF. The plots show how the features labelled f1–f4 contributed to the model's predictions. These contributions help interpret the predictions, but they do not establish which acoustic properties distinguish one emotion from another.

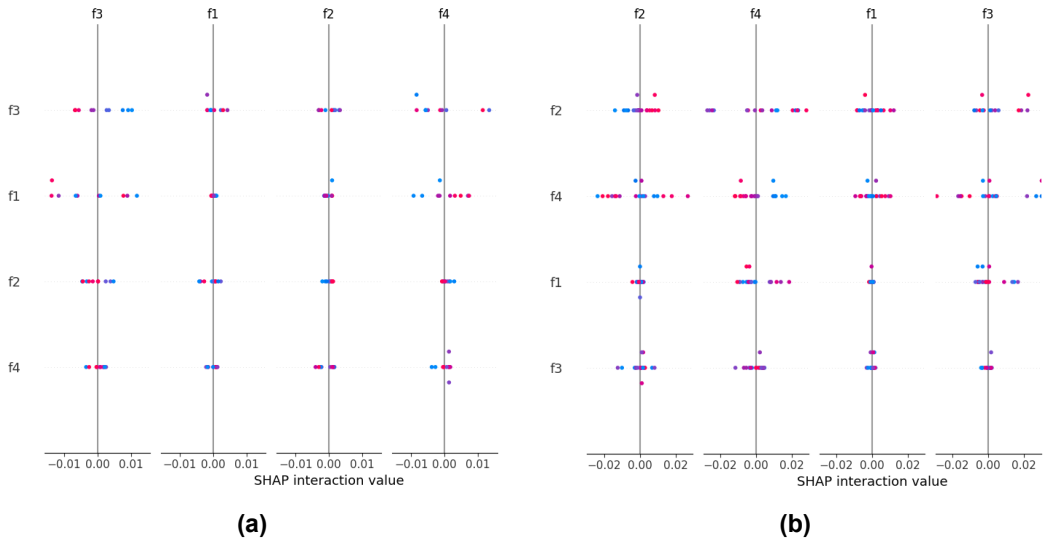


Fig 12. SHAP interaction plots for features f1–f4 during (a) training and (b) testing. The horizontal position of each point indicates its SHAP interaction value relative to zero.

Figure 12 shows the contributions of features labelled f1–f4 to the model outputs. These contributions describe the model's predictions; they do not measure the intensity of an emotion in a recording.

5.1.2. Comparative Analysis with Related Studies

To position our ESERF system within current research, we compare it with these recent studies on speech recognition under various challenges:

(i) Oliveira & Prudinccio (2025) [22] developed a new methodology where an item response theory (IRT)-based model was used to measure the performance of ASR systems based on the level of speech difficulty. Even though their model takes into consideration the performance of the system in controlled conditions and difficulty and noise, ESERF seeks to detect emotions in natural speech based on real-world databases.

(ii) Mena et al. (2025) [23] investigated the performance of ASR models (NeMo, Wav2Vec, Whisper) under babble noise and reverberation and compared the system output with human intelligibility. They present robustness of the environment in their work, and ESERF is concerned with affective variability. Emphasizing the effect of acoustic complexity, both works note that ESERF has a unique way of addressing the overlap in emotion and the existence of actor-specific differences.

(iii) Polok et al. (2025) [24] presented DiCoW, a diarization-conditioned ASR system that incorporates information about the speaker identity to boost speaker-specific recognition in multi-speaker scenarios. Although DiCoW uses label information of speakers to improve transcription, ESERF uses wavelet-based features and a hybrid CNN–BiLSTM model to detect patterns of emotions, which deals with speaker-general emotion classification.

Oliveira et al. [22], Mena et al. [23], and Polok et al. [24] examine how speech difficulty, noise and reverberation, and speaker overlap, respectively, affect automatic speech recognition. Their studies focus on recovering the words spoken, whereas ESERF aims to classify the emotion expressed in a recording. Because the tasks use different target labels, transcription results cannot serve as direct benchmarks for emotion classification. In the present study, Calm and Surprise both achieved a recall of 0.87 in Scenario 2, but their precision differed: 0.44 for Calm and 0.77 for Surprise. Thus, although the classifier detected a similar proportion of the actual recordings in each category, a prediction of Surprise was more often correct than a prediction of Calm. The related studies suggest a useful follow-up question: does this precision gap persist when the same recordings are evaluated under controlled levels of noise or reverberation? Such an experiment would help distinguish the influence of recording conditions from differences between emotion categories.

5.2. Descriptive Comparison and Statistical Limitations

The experimental scenarios are compared descriptively using the reported classification metrics. Table 8 presents the reported class-wise recall values and confidence intervals. These summaries alone do not establish statistically significant differences between scenarios.

5.2.1. Confidence Interval Estimation

For each emotion and scenario, a 95% Wilson score interval was calculated for recall using the number of correctly classified examples and the total number of examples carrying that emotion label.

$$\frac{\hat{p}+\frac{z^2}{2n}\pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z^2}{4n^2}}}{1+\frac{z^2}{n}} \quad (11)$$

Here, $\hat{p} = x/n$, where x is the diagonal count for the emotion in the confusion matrix and n is the corresponding row total. For a two-sided 95% interval, $z = 1.96$. The intervals treat evaluation examples as independent and do not account for possible clustering by speaker. The computed Cls for all scenarios are summarized in Table 8. Scenario 2 had higher reported recall than Scenario 1 for Angry, Calm, Neutral, Sad, and Surprise, but lower recall for Disgust, Fear, and Happy. The differences therefore varied across emotion categories. The displayed intervals do not, by themselves, establish statistically significant differences between the scenarios.

TABLE VIII: REPORTED CLASS-WISE RECALL AND CORRESPONDING 95% CONFIDENCE INTERVALS FOR THE THREE EXPERIMENTAL SCENARIOS.

Emotion	Support per scenario	Scenario 1 recall [95% CI]	Scenario 2 recall [95% CI]	Scenario 3 recall [95% CI]
Angry	1,438	0.721 [0.697, 0.744]	0.725 [0.702, 0.748]	0.682 [0.658, 0.706]
Calm	137	0.796 [0.720, 0.855]	0.869 [0.802, 0.915]	0.832 [0.761, 0.885]
Disgust	1,468	0.516 [0.490, 0.541]	0.463 [0.437, 0.488]	0.381 [0.357, 0.407]
Fear	1,424	0.543 [0.517, 0.569]	0.510 [0.484, 0.536]	0.258 [0.236, 0.282]
Happy	1,462	0.587 [0.561, 0.612]	0.557 [0.532, 0.583]	0.310 [0.287, 0.334]
Neutral	1,310	0.548 [0.521, 0.575]	0.569 [0.542, 0.596]	0.518 [0.490, 0.545]
Sad	1,400	0.636 [0.610, 0.661]	0.727 [0.703, 0.750]	0.713 [0.689, 0.736]
Surprise	483	0.826 [0.790, 0.857]	0.870 [0.837, 0.897]	0.700 [0.657, 0.739]

TABLE IX: REPORTED PERFORMANCE COMPARISON OF CNN-ONLY, LSTM-ONLY, AND HYBRID CNN–BiLSTM MODELS.

Model	Key Layers	Accuracy (%)	Remarks
CNN-only	2 Conv + GAP + Dense	≈ 59	strong at spatial feature extraction, but limited temporal modeling
LSTM-only	Bi-LSTM + Dense	≈ 56	captures temporal patterns but loses spectral locality
Hybrid CNN–BiLSTM (Scenario 1)	Conv + Bi-LSTM + Dense	≈ 61	integrates both spatial–temporal contexts

To isolate the contribution of each component, we sequentially removed modules from the full model as listed in Table 10.

TABLE X: REPORTED ACCURACY OF MODEL VARIANTS WITH INDIVIDUAL COMPONENTS REMOVED OR MODIFIED. DIFFERENCES ARE EXPRESSED IN PERCENTAGE POINTS RELATIVE TO THE FULL-MODEL ACCURACY OF 61%.

Variant	Removed Component	Accuracy (%)	Δ vs Full
A1	Dropout regularization	65	+4 percentage points
A2	Bi-directional LSTM (replaced with uni-LSTM)	57	−4
A3	First convolutional block	55	−6
A4	Dense head (128 → 64 neurons)	59	−2

Table 10 reports accuracy differences associated with modifications to the model. Relative to the stated full-model accuracy of 61%, removing dropout increased the reported accuracy by four percentage points. Replacing the bidirectional LSTM with a unidirectional LSTM, removing the first convolutional block, and reducing the dense-layer size were associated with

decreases of four, six, and two percentage points, respectively. These descriptive results alone do not establish overfitting, statistical significance, or the mechanisms responsible for the observed differences.

5.3. DISCUSSION

Recognition performance varied across the eight emotion categories. In Scenario 2, recall was 0.87 for both Calm and Surprise, compared with 0.46 for Disgust and 0.51 for Fear (Table V). The classifier therefore detected most Calm and Surprise recordings but missed roughly half of the Disgust and Fear recordings. High recall alone does not provide a full evaluation: for Calm, recall increased from 0.80 to 0.87 between Scenarios 1 and 2, while the F1-score fell from 0.71 to 0.58 (Tables IV and V).

A possible reason for the lowered recall of Disgust and Fear is that some recordings contain vocal features of other emotion classes. Recall scores themselves cannot determine which vocal features caused the error and what label the classifier assigned to them instead. Confusion matrices are useful for the first step in answering this question: they might help understand if the missed Fear recordings were consistently categorized as a certain emotion or as different categories. Further, analyzing errors by the source dataset, by speaker, and by emotional intensity (if there are corresponding annotations) would help to differentiate the issue of confusion between emotions from that of errors specific to certain recordings.

There are two research directions to focus on following these observations. First, it is necessary to evaluate the classifier on speakers/datasets not used for training to verify whether the described error patterns remain applicable to the combined dataset. Second, it is possible to compare the use of attention mechanism in classification with the current approach using the same partition for training and testing the model. It is important to determine if such an approach can improve the recall of often missed emotions without increasing the number of false predictions. Precisions, recalls, and F1-scores for each emotion would reflect this trade-off even if the overall accuracy changes slightly.

5.3.1. Accuracy and Model Evaluation

Overall accuracy was approximately 61% in both Scenario 1 and Scenario 2, but recall varied across emotion categories (Tables 4 and 5). Compared with Scenario 1, Scenario 2 achieved higher recall for Angry, Calm, Neutral, Sad, and Surprise, but lower recall for Disgust, Fear, and Happy. For Calm, recall increased from 0.80 to 0.87 while precision fell from 0.64 to 0.44. Scenario 2 therefore detected a larger proportion of Calm recordings, but fewer than half of the recordings it labelled as Calm belonged to that category.

Overall accuracy fell to approximately 49% in Scenario 3. The change in recall from Scenario 2 to Scenario 3 was much larger for Fear, which fell from 0.51 to 0.26, than for Sad, which fell from 0.73 to 0.71 (Tables 5 and 6). These results identify which emotions were missed more frequently, but the aggregate metrics do not establish whether the differences arose from feature processing, model configuration, or other experimental conditions.

5.3.2. Training Dynamics

Training loss decreasing means improving fit of the model on the training data. Training loss/evaluation loss discrepancy can signalize overfitting; however, evaluation accuracy lowering doesn't imply the reason behind this phenomenon. As a result, learning curves must be evaluated together with classification metrics.

5.3.3. Classification Performance

The confusion matrices and classification reports show recognition accuracy discrepancies for different emotion categories. For example, in Scenario 2, both Calm and Surprise have recall equal to 0.87, but Calm has precision of 0.44 and Surprise – 0.77. Therefore, even though most of the Calm samples were recognized, there is still a considerable number of samples from other categories classified as Calm. Thus, the difference shows why both recall and precision need to be taken into account for recognition accuracy evaluation.

In Scenario 3, the overall accuracy is equal to 49%, compared to 61% in scenarios 1 and 2. Fear has precision of 0.76, but recall of 0.26, which means that the model doesn't recognize many Fear samples despite rather precise predictions of Fear samples. These results show recognition accuracy limitations, but the reasons can't be specified from the aggregate metrics alone.

Overall accuracy reducing from Scenario 2 to Scenario 3 is not evenly distributed among all emotions. The recall for Fear decreases from 0.51 to 0.26 and for Happy – from 0.56 to 0.31, which is 25 percentage point reduction in both cases. In contrast to that, the Sad recall is reduced from 0.73 to 0.71 only. This difference reveals the sectors where recognition loses the most without identifying the feature or training factor behind that.

5.3.4. Predictive Insights and Reliability

Error analysis can further focus on two types of error: first, which other emotions were incorrectly classified as Calm in Scenario 2, and second, which labels were assigned to missed Fear samples in Scenario 3. Confusion matrices can reveal these patterns. Changes to the feature set could be evaluated using the same test examples for each configuration to determine whether they reduce these errors. Holding out speakers or an entire dataset for testing would assess whether the model generalizes to unfamiliar voices or recording conditions.

6. CONCLUSION

In this study, we examined speech emotion recognition using recordings from RAVDESS, CREMA-D, and TESS. Across the eight emotion categories, overall accuracy was around 61% for Scenarios 1 and 2 and 49% for Scenario 3. Even though the first two scenarios had similar overall accuracy, they differed in how reliably individual emotions were detected. In particular, for Calm, recall increased from 0.80 in Scenario 1 to 0.87 in Scenario 2, while precision dropped from 0.64 to 0.44. This example demonstrates why both missed detections and false predictions matter when evaluating recognition performance.

Future studies may investigate whether the same error patterns hold when speakers or an entire source dataset are held out for testing. Measuring inference time would also help assess whether the proposed framework is suitable for real-time settings. Future research may explore the security needs of speech emotion recognition systems used in voice applications, including the effects of adversarial audio perturbations, spoofed or synthetic speech, and feature manipulation on classification performance. Separate experiments should assess whether feature-protection mechanisms provide confidentiality or resistance to feature reconstruction and tampering. Such experiments should define the attacker's capabilities, compare protected and unprotected settings, and report security outcomes and the effects of protection on recognition accuracy and computational cost. The current study does not establish these security properties.

CRediT authorship contribution statement

Suhad Al-Shoukry: Data curation, Supervision, Writing – review & editing. Rihab Habeeb Sahib and Wurood Mahdi Sahib: Conceptualization, Methodology, Writing – original draft. Abdulkadhem A. Abdulkadhem and Mohamed Shabbir Abdulnabi: Investigation, Formal analysis, Methodology, Writing – original draft.

Conflict of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Ethical approval

This study used existing recordings from three publicly available speech datasets. The authors did not recruit participants or collect new recordings for this study.

Data Availability

The source recordings are publicly available from RAVDESS [18], CREMA-D [19], and TESS [20]. Access details are provided in the cited dataset records.

References

[1] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends," *IEEE Access*, vol. 8, pp. 16560–16572, 2020.

[2] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *Int. J. Speech Technol.*, vol. 21, pp. 93–120, 2018.

[3] M. S. Akhtar, D. Ghosal, A. Ekbal, P. Bhattacharyya, and S. Kurohashi, "All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 285–297, 2019.

[4] D. Chandola, E. Altarawneh, M. Jenkin, and M. Papagelis, "SERC-GCN: Speech Emotion Recognition In Conversation Using Graph Convolutional Networks," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 76–80.

[5] M. Khan, W. Gueaieb, A. El Saddik, and S. Kwon, "MSER: Multimodal speech emotion recognition using cross-attention with deep fusion," *Expert Syst. Appl.*, vol. 245, p. 122946, 2024.

[6] H. Cai, P. Zhang, Y. Xiao, S. Ji, M. Xiao, and L. Cheng, "Clean-label backdoor attack based on robust feature attenuation for speech recognition," *Expert Syst. Appl.*, vol. 281, p. 127546, 2025, doi:

https://doi.org/10.1016/j.eswa.2025.127546.
[7] W. Yao, J. Yang, Y. He, J. Liu, and W. Wen, “Imperceptible rhythm backdoor attacks: Exploring rhythm transformation for embedding undetectable vulnerabilities on speech recognition,” *Neurocomputing*, vol. 614, p. 128779, 2025, doi: https://doi.org/10.1016/j.neucom.2024.128779.
[8] D. Rong, Q. Sun, Y. Wang, and X. Wang, “A new universal camouflage attack algorithm for intelligent speech system,” *Speech Commun.*, vol. 166, p. 103152, 2025, doi: https://doi.org/10.1016/j.specom.2024.103152.
[9] H. Zhao, Y. Xia, R. Y. K. Lau, and Y. Liu, “Word sentiment polarity disambiguition based on opinion level context,” in *2012 international conference on machine learning and cybernetics*, 2012, vol. 5, pp. 2007–2012.
[10] Y. Xia, E. Cambria, A. Hussain, and H. Zhao, “Word polarity disambiguation using bayesian model and opinion-level features,” *Cognit. Comput.*, vol. 7, pp. 369–380, 2015.
[11] W. Dai, D. Han, Y. Dai, and D. Xu, “Emotion recognition and affective computing on vocal social media,” *Inf. Manag.*, vol. 52, no. 7, pp. 777–788, 2015.
[12] H. M. Fayek, M. Lech, and L. Cavedon, “Evaluating deep learning architectures for speech emotion recognition,” *Neural Networks*, vol. 92, pp. 60–68, 2017.
[13] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger, “Deep learning for affective computing: Text-based emotion recognition in decision support,” *Decis. Support Syst.*, vol. 115, pp. 24–35, 2018.
[14] X. Tang, J. Huang, Y. Lin, T. Dang, and J. Cheng, “Speech Emotion Recognition via CNN-Transformer and multidimensional attention mechanism,” *Speech Commun.*, vol. 171, p. 103242, 2025, doi: https://doi.org/10.1016/j.specom.2025.103242.
[15] M. Khan, P.-N. Tran, N. T. Pham, A. El Saddik, and A. Othmani, “MemoCMT: multimodal emotion recognition using cross-modal transformer-based feature fusion,” *Sci. Rep.*, vol. 15, no. 1, p. 5473, 2025.
[16] S. Akinpelu, S. Viriri, and M. Haroon Yousaf, “SwinTSER: An improved bilingual speech emotion recognition using shift window transformer,” *Cognit. Comput.*, vol. 17, no. 4, p. 129, 2025.
[17] C.-S. Ahn, R. Rana, C. Busso, and J. C. Rajapakse, “Multitask Transformer for Cross-Corpus Speech Emotion Recognition,” *IEEE Trans. Affect. Comput.*, vol. 16, no. 3, pp. 1581–1591, 2025, doi: 10.1109/TAFFC.2025.3526592.
[18] S. R. Livingstone and F. A. 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, Art. no. e0196391, 2018, doi: 10.1371/journal.pone.0196391.
[19] 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 Transactions on Affective Computing, vol. 5, no. 4, pp. 377–390, 2014, doi: 10.1109/TAFFC.2014.2336244.
[20] M. K. Pichora-Fuller and K. Dupuis, “Toronto emotional speech set (TESS),” Borealis, version 1, 2020, doi: 10.5683/SP2/E8H2MF.
[21] N. Das, N. Padhy, N. Dey, H. Paul, and S. Chowdhury, “Exploring explainable AI methods for bird sound-based species recognition systems,” *Multimed. Tools Appl.*, vol. 83, no. 24, pp. 64223–64253, 2024, doi: 10.1007/s11042-023-17982-3.
[22] C. S. Oliveira and R. B. C. Prudêncio, “An item response theory framework to evaluate automatic speech recognition systems against speech difficulty,” *Comput. Speech Lang.*, vol. 95, p. 101817, 2026, doi: https://doi.org/10.1016/j.csl.2025.101817.
[23] C. Mena, A. L. Padilla-Ortiz, and F. Orduña-Bustamante, “Automatic speech recognition in the presence of babble noise and reverberation compared to human speech intelligibility in Spanish,” *Comput. Speech Lang.*, vol. 95, p. 101856, 2026, doi: https://doi.org/10.1016/j.csl.2025.101856.
[24] A. Polok *et al.*, “DiCoW: Diarization-conditioned Whisper for target speaker automatic speech recognition,” *Comput. Speech Lang.*, vol. 95, p. 101841, 2026, doi: https://doi.org/10.1016/j.csl.2025.101841.