IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 10, October, 2025, pp. 2727-2737
Journal homepage: http://iieta.org/journals/isi

Combined Acoustic Features with CNN-BiLSTM-Transformer for Female Emotion Recognition

Dede Kurniadi[1*], Erick Fernando[2], Syahrul Al Zayyan[1], Asri Mulyani[1]

[1] Department of Computer Science, Institut Teknologi Garut, Garut 44151, Indonesia
[2] Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang 15810, Indonesia

Corresponding Author Email: dede.kurniadi@itg.ac.id

https://doi.org/10.18280/isi.301018

Received: 10 September 2025
Revised: 14 October 2025
Accepted: 22 October 2025
Available online: 31 October 2025

Keywords:
acoustic feature fusion, CNN-BiLSTM-transformer, deep learning, female speech emotion recognition, hybrid model

ABSTRACT

Speech Emotion Recognition (SER) is essential for enhancing human-computer interaction by enabling machines to understand user emotional states. However, SER still faces challenges, such as the complexity of audio signals, individual differences, and limited focus on female voices, which often exhibit higher pitch and subtler emotional cues. This study introduces a hybrid model combining Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and Transformer to classify emotions in female speech. The model is trained using the RAVDESS, CREMA-D, and TESS datasets, with stepwise acoustic features: MFCC, ZCR, LPC, RMSE, and ZCPA. Data augmentation techniques are applied to address class imbalance and improve generalization, including the addition of additive noise and pitch shifting to simulate natural variations in female vocal pitch. Additionally, SMOTE is employed to generate synthetic samples for minority classes. Performance is evaluated using 5-fold cross-validation. Results show that the best performance is achieved using the MFCC + ZCR combination, with 88.52% accuracy, 88.80% precision, 88.52% recall, 88.53% F1-score, and 98.95% AUC-ROC. This research advances SER by developing a robust, context-aware model tailored to female vocal traits.

1. INTRODUCTION

Emotions are more than waves of feelings they shape how we think, make decisions, and interact. In spoken communication, elements such as tone, intonation, speech rate, and pauses often convey emotional meaning far beyond the literal content of the words themselves [1, 2]. For instance, the word "fine" spoken slowly in a flat tone may reflect sadness or fatigue, while the same word said with a rising tone and quicker pace might indicate enthusiasm. Subtle variations in acoustic features, particularly in pitch and spectral representations, have been identified as indicators of emotional or depressive states, even when such changes may not be perceptible to the human ear [3].

Voice-based intelligent systems, such as voice assistants, call centers, and voice bots, are widely used today; yet, most still struggle to accurately detect user emotions, particularly those from female voices. A study by Lin et al. [4] revealed that several state-of-the-art Speech Emotion Recognition models showed higher accuracy for male speakers than female speakers, highlighting a persistent gender bias in emotion classification systems. Similarly, Tursunov et al. [5] reported that speech-based recognition systems often exhibit gender bias, where male voices tend to achieve higher recognition accuracy compared to female voices. Their study demonstrated that even advanced CNN-based models could reach up to 96% accuracy for gender classification, yet performance disparities remain evident indicating that acoustic features extracted from female speech are more challenging for models to generalize accurately.

The ability of machines to "sense" these emotional nuances opens up new possibilities for human-computer interaction [6]. Imagine a virtual assistant that immediately offers help upon detecting a frustrated tone, or a mental health application that monitors signs of anxiety from daily phone calls to provide earlier support. In education, an online tutor that can detect student boredom or confusion through vocal cues could dynamically adjust the learning material in real-time. All of these applications depend heavily on how quickly and accurately a model can process highly dynamic audio signals [7].

Although many approaches have been developed for speech emotion recognition (SER), most remain focused on general data without accounting for the differences in vocal characteristics between male and female speakers. Physiologically, female speakers typically have shorter and lighter vocal folds, resulting in a higher average fundamental frequency (F0) range of around 200-260 Hz compared to 85-180 Hz in males, along with more closely spaced formants [4, 5]. These differences affect the spectral contour and energy distribution of speech, often requiring adjustments in thresholds or feature weighting to accurately capture the subtle emotional fluctuations in female voices. Without addressing these distinctions, SER models tend to be biased toward male voice patterns. They may fail to recognize the more delicate and rapidly shifting emotional cues commonly present in female speech.

In the field of deep learning-based audio signal processing, three complementary architectures have gained prominence: the Convolutional Neural Network (CNN) for extracting spatial patterns from spectral representations, the Bidirectional Long Short-Term Memory (BiLSTM) for capturing bidirectional temporal dynamics, and the Transformer for modeling global context in sequences through self-attention mechanisms [7, 8]. While these architectures have demonstrated effectiveness individually, studies integrating all three into a unified processing pipeline, especially for female voice emotion recognition, remain scarce. This is particularly important given the complex frequency and intonation patterns of female speech.

This study proposes and evaluates a hybrid CNN-BiLSTM-Transformer model for female speech emotion recognition. The architecture integrates the three modules hierarchically: CNN serves as the initial stage for extracting local spatial features from acoustic input; its output is then passed to BiLSTM to capture bidirectional temporal sequences; finally, a Transformer layer assigns global attention weights to the most relevant signal segments. This design combines the strengths of all three models while adapting to the fluctuating pitch and intonation patterns unique to female voices.

2. RELATED WORKS

The study by Gomathy [9] focuses on enhancing speech emotion recognition accuracy and efficiency by utilizing an Enhanced Cat Swarm Optimization (ECSO) algorithm. This method helps select only the most important speech features, such as MFCC, LPC, and LPCC, so the system can better recognize emotions while reducing unnecessary data and processing time. ECSO enhances the original Cat Swarm Optimization by incorporating an Opposition-Based Learning (OBL) strategy, which enables the algorithm to explore more possibilities and find the optimal solution more efficiently. The selected features are then analyzed using a Support Vector Neural Network (SVNN) to classify different emotions in speech. When tested in MATLAB, the proposed ECSO-SVNN model achieved impressive results, with 96% accuracy, 0.74 sensitivity, 0.97 specificity, and a 93.4% recognition rate, outperforming other existing methods. In short, this study shows that combining ECSO and SVNN can significantly improve the way machines recognize human emotions through voice.

Anvarjon et al. [6] introduced an efficient method combining RBFN for speech segment selection, CNN for feature extraction, and BiLSTM for temporal modeling. Accuracy results were 85.57% (EMO-DB), 72.25% (IEMOCAP), and 77.02% (RAVDESS). Despite promising results, issues with data imbalance and underrepresentation of emotions persisted.

Kim and Lee [8] developed a hybrid model combining BiLSTM, Transformer, and 2D CNN to enhance emotion recognition using Mel-spectrograms. The model achieved score of 95.65% (EMO-DB) and 80.19% (RAVDESS). Although powerful in capturing emotional representations, its complexity limits real-time application.

The study by Kacur et al. [10] proposes a Convolutional Neural Network model that recognizes emotions directly from log-Mel spectrograms without manual feature extraction. Using the RAVDESS dataset, the model classifies emotions like happy, sad, angry, and neutral. The CNN achieved 93.7% accuracy, outperforming traditional methods such as SVM and KNN. The results show that CNNs effectively capture emotional cues from speech and are suitable for real-time emotion recognition.

Zhao et al. [11] proposed a 1D and 2D CNN-LSTM hybrid to extract local patterns and temporal dynamics from spectrograms. The model achieved 95.33% (Emo-DB) and 89.16% (IEMOCAP) accuracy. While effective, the approach's computational demands make real-time use difficult.

Finch [12] introduced Dynamic CNN with BiLSTM, using adaptive convolutional kernels responsive to emotional changes. It was tested on CISIA, EMO-DB, and IEMOCAP, achieving accuracy of 59.08%, 89.29%, and 71.25%, respectively. Despite its flexibility, the model relies heavily on precise hyperparameter tuning.

Despite notable progress in SER, several key challenges remain. Many existing models overlook the unique vocal traits of female speech, such as higher pitch and dynamic spectral patterns, leading to biased performance due to reliance on gender-agnostic datasets [4]. Additionally, traditional SER studies often use arbitrary acoustic feature combinations without systematically evaluating their impact. Few have adopted a structured, stepwise feature extraction strategy tailored for female voices. Moreover, although hybrid models like CNN BiLSTM and Transformers show strong results [8], their complexity limits real-time deployment, and they are rarely evaluated with feature pipelines designed explicitly for female speech. Common issues such as class imbalance and noise sensitivity are also frequently under-addressed [9]. Data augmentation techniques, such as Gaussian noise injection and pitch shifting, are essential for enhancing model robustness and adapting to real-world conditions. Specifically, pitch shifting is applied to simulate natural variations in female vocal frequency, improving the model's ability to recognize subtle emotional cues across different voice tones. Therefore, an SER framework that combines a hybrid architecture with progressive, female-focused feature extraction and targeted augmentation is crucial for developing more robust, inclusive, and real-world-ready systems.

3. MATERIAL AND METHODS

The research workflow is summarized in Figure 1. It outlines the stages, from data augmentation and progressive feature extraction to k-fold cross-validation, model training using a CNN-BiLSTM-Transformer architecture, and final evaluation using the Confusion Matrix and AUC-ROC.

Figure 1 illustrates the overall workflow and architecture of the proposed Speech Emotion Recognition (SER) system. The process begins with the audio dataset, which undergoes data augmentation techniques such as additive noise and pitch shifting to enhance data diversity and robustness. The augmented data is then used for feature extraction, where five combinations of acoustic features are generated.

Each feature combination is evaluated using a five-fold cross-validation scheme to ensure generalization and avoid overfitting. The modeling stage employs a hybrid deep learning architecture combining CNN, BiLSTM, and Transformer layers. The performance of each model is evaluated using the Confusion Matrix and AUC-ROC metrics, ensuring a comprehensive assessment across all emotional categories.

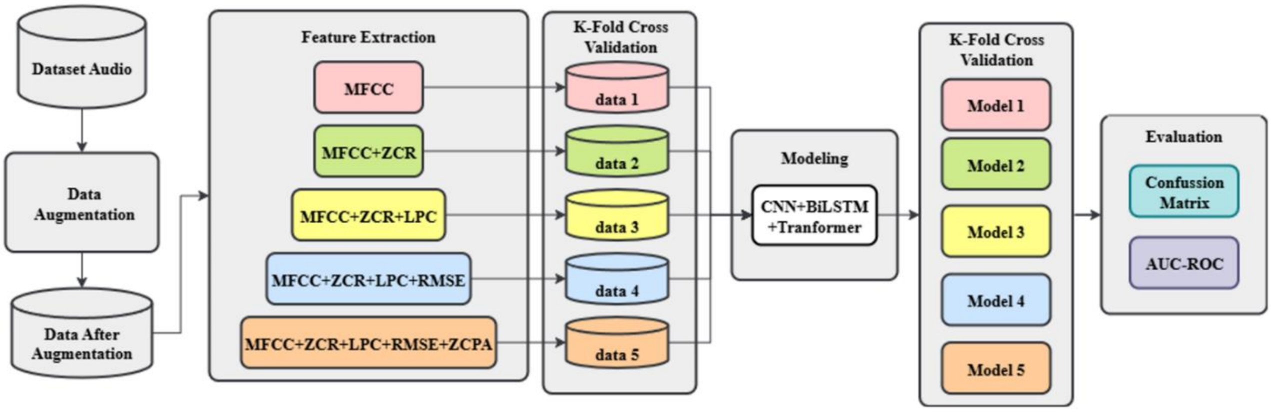


Figure 1. Research method

3.1 Dataset

This study utilized a combination of three publicly available emotional speech datasets: CREMA-D [13], RAVDESS [14], and TESS [15]. To ensure consistency in vocal characteristics and focus on the unique features of female speech, only audio samples spoken by female actors were selected.

The original dataset contained eight emotion categories: angry, fear, disgust, sad, happy, neutral, surprised, and calm. However, to maintain class balance and reduce the impact of underrepresented categories, only six primary emotions were retained: angry, fear, disgust, sad, happy, and neutral. The surprised and calm categories were excluded due to insufficient sample counts. After filtering, the final dataset consisted of 4,002 audio samples, with the distribution presented in Table 1.

Table 1. Total data

Dataset	Angry	Sad	Happy	Disgust	Fear	Neutral
CREMA-D	91	91	91	91	91	91
RAVDESS	192	192	192	192	192	96
TESS	400	400	400	400	400	400

To address the imbalance in the neutral emotion category, which initially had fewer samples than the other classes, this study applied the Synthetic Minority Oversampling Technique (SMOTE) [16]. SMOTE generates synthetic data points for the minority class rather than simply duplicating existing samples, thus introducing more variation and reducing the risk of overfitting. The technique creates new samples using Eq. (1).

$$x_{new} = x_i + \lambda(x_{NN} - x_i), \lambda \sim U(0,1) \qquad (1)$$

where, x_i is the original minority sample, x_{NN} is its nearest neighbor, and λ is drawn uniformly between 0 and 1. Thus, SMOTE generates new points along the line segments connecting minority samples, balancing the class distribution without exact duplication.

To improve model robustness and generalization, especially in real-world scenarios, this study applied two data augmentation techniques: additive noise and pitch shifting. Additive noise involves injecting random Gaussian noise into the original audio signal to simulate environmental disturbances such as background conversations, wind, or electronic hums [17]. This augmentation is mathematically defined in Eq. (2).

$$x_{noisy}(t) = x(t) + \sigma N(0,1) \qquad (2)$$

To simulate environmental variations, the original signal $x(t)$ is added with Gaussian noise $n(t)$ where σ controls the noise level, and $N(0,1)$ represents standard normal noise. This technique enhances the model's robustness against real-world noise.

In addition, pitch shifting was used to simulate variations in vocal frequency, which is particularly relevant for modeling emotional expressions in female voices, which generally have higher pitch ranges. This technique modifies the pitch of the audio signal by a certain number of semitones without affecting its duration, enabling the model to capture pitch-dependent emotional cues [18]. The transformation can be expressed as in Eq. (3).

$$x_{pitch-shifted}(t) = PitchShift(x(t), \Delta p) \qquad (3)$$

where, Δp is the pitch shift amount in semitones. The implementation uses the librosa.effects.pitch_shift() function from the Librosa library. Both techniques enhance the diversity of training data, helping to reduce overfitting and improve the model's sensitivity to subtle emotional variations in female speech. The key parameters used in the data augmentation process are summarized in Table 2, and the results of the augmented data are presented in Table 3.

Table 2. Data augmentation parameter

Technique	Parameter	Symbol	Value
Additive Noise	Standard deviation of noise	σ	0.02
Pitch Shifting	Pitch shift in semitones	Δp	±4 semitones
SMOTE	Number of nearest neighbors	k	5

As shown in Table 2, the applied data augmentation techniques, additive noise, pitch shifting, and SMOTE, were configured to enhance the model's robustness against variations in acoustic conditions.

Table 3. Data after augmentation

Emotion	Original	+50% Augmentation	New Total
Angry	683	+342	1,025
Fear	683	+342	1,025
Disgust	683	+342	1,025
Sad	683	+342	1,025
Happy	683	+342	1,025
Neutral	587	+294	**881**
Total	4,002	+2,004	**6,006**

Based on the data in Table 3, after applying the 50% data augmentation, the dataset size increased from 4,002 to approximately 6,006 audio samples, maintaining proportional class distributions. However, the neutral emotion category remained slightly underrepresented, with 881 samples compared to 1,025 samples in other classes. To address this imbalance, the SMOTE algorithm was applied exclusively to the neutral class, generating 144 synthetic samples. Consequently, the final dataset comprised approximately 6,150 samples across six balanced emotion categories, providing a more uniform data distribution for training and evaluation.

3.2 Feature extraction

In this study, feature extraction was performed on preprocessed and augmented audio data using five acoustic features: MFCC, ZCR, LPC, RMSE, and ZCPA. These features were selected for their ability to capture different aspects of the speech signal relevant to emotion recognition.

3.2.1 Mel-frequency cepstral coefficients (MFCC)

Captures the spectral characteristics of speech and simulates the human auditory system, making it highly effective for distinguishing emotional states [19].

$$MFCC(n) = \sum_{k=1}^{K} log(E_m) \cos\left(\frac{\pi k}{M}(m - 0.5)\right), k = 1, \ldots\ldots, K \tag{4}$$

where, $\boldsymbol{E_m}$ is the energy at the m-th Mel filter and M is the total number of filters. In this study, 13 MFCC coefficients were extracted per frame using a 40-filter Mel-scale filterbank, with a frame length of 25 ms and hop length of 10 ms.

3.2.2 Zero crossing rate (ZCR)

Measures how often the signal crosses the zero-amplitude axis and is helpful in detecting abrupt signal changes, which are usually present in high-arousal emotions [19].

$$ZCR = \frac{1}{N-1} \sum_{n=1}^{N-1} |sgn(x_n) - sgn(x_{n-1})| \tag{5}$$

where, $\boldsymbol{x_n}$ is the amplitude value at time n, $\boldsymbol{N}$ is the number of samples in one frame, and $\boldsymbol{sgn(x)}$ is the sign function, which equals 1 if x is positive and -1 if x is negative. ZCR was computed on each frame (25 ms, 50% overlap) to capture high-frequency variations in the signal.

3.2.3 Linear predictive coding (LPC)

Models the resonant frequencies of the vocal tract and helps to capture phonetic and prosodic patterns tied to emotional expression [20].

$$s(n) = \sum_{i=1}^{p} a_i s(n-1) + e(n) \tag{6}$$

where, $s(n)$ is the speech signal at time n p is the order of the LPC model, a_i are the LPC coefficients representing the characteristics of the speech filter, and $e(n)$ is the residual error value. An LPC model of order p = 10 was used to approximate the vocal tract response, with coefficients estimated on a per-frame basis.

3.2.4 Root mean square energy (RMSE)

Represents the intensity or loudness of the signal, helpful in identifying emotions with strong or weak energy patterns [21].

$$RMS = \sqrt{\frac{1}{N} \sum_{n=1}^{N} x_n^2} \tag{7}$$

where, x_n is the signal amplitude at time n, and N is the number of samples at time n. *RMSE* was calculated per frame (a 25-ms window) to quantify the signal's energy level.

3.2.5 Zero crossing peak amplitude (ZCPA)

Combines ZCR with peak amplitude, adding information about the magnitude of signal changes around zero crossings [19].

$$ZCPA = ZCR \times max(|x_n|) \tag{8}$$

where, *ZCR* is the zero crossing count within one frame, and max($|n|$) represents the peak amplitude in that frame. *ZCPA* was computed using the same frame configuration (25 ms, 50% overlap) to capture both zero-crossing density and amplitude variations.

Algorithm 1 outlines the detailed steps involved in the feature extraction process used in this study. Each audio file undergoes preprocessing, noise augmentation, and the extraction of five key acoustic features: MFCC, ZCR, LPC, RMSE, and ZCPA. These features are then incrementally combined to form five distinct datasets. The specific configurations of each feature set are summarized in Tables 4, 5, and 6, which serve as a reference for understanding the composition of the generated datasets used for model training and evaluation.

Algorithm 1. Feature extraction

```
Step 1: Initialize
-Five empty datasets for each feature combination
-One empty list to store emotion labels
Step 2: For each audio file in the dataset:
1. Preprocessing:
-Load audio file with sampling rate.
-Apply pitch shifting using Eq. (2).
-Add noise using Eq. (3).
2. Feature extraction:
-Compute MFCC using Eq. (4).
- ZCR using Eq. (5).
-Compute LPC using Eq. (6).
-Compute RMSE using Eq. (7).
-Compute ZCPA using Eq. (8).
3. Feature Combination:
Dataset 1: MFCC only
Dataset 2: MFCC + ZCR
Dataset 3: MFCC + LPC + ZCR
Dataset 4: MFCC + LPC + ZCR + RMSE
Dataset 5: MFCC + LPC + ZCR + RMSE + ZCPA
4. Add corresponding emotion label to each dataset entry
5. Display progress for every multiple of batch_size
Step 3: After all files are processed:
-Save all five datasets to separate CSV files
-Each CSV file includes feature values and the emotion label column
```

Theoretically, the MFCC represent the most fundamental and widely adopted features in speech processing, due to their ability to capture spectral contours that closely align with human auditory perception. These features are particularly effective in distinguishing vocal patterns associated with different emotions; for example, sadness typically exhibits flatter, lower-frequency contours [19]. The ZCR contributes additional information by quantifying the number of times a signal transitions from positive to negative within a single frame. This metric is highly sensitive to the signal's texture and is especially useful for identifying high-intensity emotions such as anger or surprise [19]. LPC strengthens the feature representation by modeling the resonant characteristics of the speaker's vocal tract, making it well-suited for capturing phonetic attributes that differentiate emotional expressions [20]. RMSE measures the average energy within a frame, reflecting the loudness and emotional intensity of the speech signal. For instance, angry speech tends to exhibit higher RMSE values, whereas neutral or sad speech typically demonstrates lower energy [21]. Finally, ZCPA combines the sensitivity of ZCR to signal transitions with the peak amplitude values occurring at those transition points [19]. This integration provides an additional dimension for detecting subtle, micro-level emotional variations within the speech signal.

By integrating these features progressively, the model benefits from a comprehensive set of complementary information ranging from global spectral patterns and transition dynamics to vocal resonance, energy intensity, and micro-amplitude fluctuations. Such a feature fusion strategy enhances classification accuracy and offers a deeper understanding of the most salient attributes for recognizing emotions in female speech [22].

Finally, the dataset, undergoing a series of transformations, is randomly divided into training and testing sets using the k-fold cross-validation technique. K-fold cross-validation partitions the dataset into K equally sized, non-overlapping subsets. Each subset is used once as the validation set while the remaining K-1 subsets are used for training. This process is repeated K times so that each subset serves as the validation set exactly once. The final performance is computed as the average of the evaluation metrics across all folds, providing an almost unbiased and more stable error estimate compared to a single train-test split [23]. Typically, K = 5 is chosen to balance bias and variance. In classification tasks, stratified k-fold is commonly used to ensure that the class proportions in each fold reflect the original distribution of the data.

3.3 Hybrid model

The architecture proposed in this study is a hybrid deep learning model that integrates three powerful components: CNN [24, 25], BiLSTM [26, 27], and the Transformer [8, 28]. This combination is designed to simultaneously capture spatial, temporal, and contextual information from female speech signals, which are known to exhibit high pitch and subtle emotional variations. The detailed implementation of this model is outlined in Algorithm 2.

Algorithm 2. Proposed model

```
Step 1: Build hybrid model architecture
Define the input layer according to the feature shape
Step 2: Add CNN block:
Apply 1D Convolutional layer with ReLU activation
    $Yt,k = \sigma(\sum_{i=1}^{F} x_{t+i} \times w_{i,k} + b_k)$
where, $w_{i,k}$ and $b_k$ are the filter weight and bias, respectively, and $\sigma$ denotes the ReLU activation.
Followed by max pooling to reduce dimensionality and dropout for regularization.
Step 3: Add BiLSTM block:
Add Bidirectional LSTM captures bidirectional temporal dependencies, where the state is computed as:
    $h_t^{bi} = [h_t^{forward}; h_t^{backward}]$
Enabling the model to retain past and future contextual information.
Step 4: Add Transformer Block:
Add Positional Encoding further models long-range relationships using positional encoding and multi-head self-attention:
    $Attention(Q, K, V) = softmax\ (\frac{QK^T}{\sqrt{d_k}})V$
which is followed by a feed-forward network to refine learned representations.
Step 5: Flatten the output
The outputs are flattened and passed through a dense layer, with final classification obtained using a softmax function:
    $\hat{y}_i = \frac{\exp\ (z_i)}{\sum_{j=1}^{C} \exp\ (z_i)}$
where, $C$ is the number of emotion classes.
Step 6: Compile model:
Compile model with configuration in Table 2.
Step 7: Train the model:
Perform 5-fold Cross-Validation.
Record each fold's performance metrics (loss, accuracy, precision, recall, F1-score).
Step 8: Save the best-performing model
```

The CNN block extracts local spatial features from the input feature maps, such as formant structures and spectral energy distributions. These patterns are essential for capturing the underlying acoustic structure of emotional speech [24, 25]. Next, the extracted features are passed to a BiLSTM layer, which processes information in forward and backward directions. This bidirectional flow allows the model to retain long-range temporal dependencies, such as speech rhythm and intonation patterns, that are critical in emotion recognition [26, 27]. Following the BiLSTM, the sequence data is input to a Transformer encoder, which employs a self-attention mechanism and positional encoding to model global dependencies across the entire feature sequence [8, 28]. This enables the model to focus on the most informative parts of the speech signal, improving its ability to distinguish subtle emotional cues. The final output from the Transformer is flattened and passed through one or more fully connected (dense) layers, followed by a softmax activation function to classify the input into one of the six emotion categories: angry, fear, disgust, sad, happy, and neutral [1, 2]. The configuration details of each layer in the proposed model are summarized in Table 4.

Table 4. Model Architecture

Layer Type	Parameters / Configuration
Input Layer	Input shape
Conv1D (1)	Filters = 64, Kernel size = 3, Activation = ReLU, Padding = 'same'
Batch Normalization (1)	—
Conv1D (2)	Filters = 64, Kernel size = 3, Activation = ReLU, Padding = 'same'
Batch Normalization (2)	—
MaxPooling1D (1)	Pool size = 2
Dropout (1)	Dropout rate = 0.3
Conv1D (3)	Filters = 128, Kernel size = 3, Activation = ReLU, Padding = 'same'
Batch Normalization (3)	—
MaxPooling1D (2)	Pool size = 2
Dropout (2)	Dropout rate = 0.3
BiLSTM (1)	Units = 128 (bidirectional), return_sequences = True
BiLSTM (2)	Units = 64 (bidirectional), return_sequences = True
Attention Layer	Dense (1, activation = tanh) → Flatten → Softmax → Reshape → Multiply
GlobalAveragePooling1D	—
Dense (1)	Units = 128, Activation = ReLU
Dropout (3)	Dropout rate = 0.5
Dense (2)	Units = 64, Activation = ReLU
Dropout (4)	Dropout rate = 0.5
Output Layer	Units = num_classes, Activation = Softmax

This hybrid architecture leverages the strengths of each component: local pattern detection (CNN), sequential modeling (BiLSTM), and contextual attention (Transformer) to construct a highly expressive and robust model for female speech emotion recognition. The model is trained using the model described in algorithm 2, and optimized using categorical cross-entropy loss [29] and the Adamax optimizer, as further detailed in Table 5.

Table 5. Hyperparameter configuration

Hyperparameter	Value
Loss	Categorical Cross-Entropy
Optimizer	Adamax
Epoch	100
Batch Size	32
Learning rate	0,01

Table 6. Model feature extraction combination

Name	Method	Model
Dataset 1	MFCC	Model 1
Dataset 2	MFCC+ZCR	model 2
Dataset 3	MFCC+ZCR+LPC	Model 3
Dataset 4	MFCC+ZCR+LPC+RMSE	Model 4
Dataset 5	MFCC+ZCR+LPC+RMSE+ZCPA	Model 5

In addition, the combination of feature extraction sets with their corresponding model performance is summarized in Table 6, providing a clear overview of how each feature configuration influences the training and evaluation outcomes of the hybrid architecture.

3.4 Validation Strategy and Performance Metrics

To evaluate the performance of the proposed hybrid model, this study employed a 5-fold cross-validation strategy to ensure consistency and robustness across different subsets of data [30]. In this approach, the dataset is divided into five equal parts, where each part serves as validation data, while the remaining four are used for training. This rotation is repeated until every fold has been used once as a validation set, and the average performance across all folds is computed. The model's effectiveness was assessed using several classification metrics, including accuracy, precision, recall, and F1-score [31, 32]. Accuracy reflects the overall correctness of predictions; precision measures the proportion of relevant positive predictions; recall indicates the model's ability to identify all actual positive instances; and the F1-score balances precision and recall, which is especially useful for handling imbalanced classes.

In addition, the AUC-ROC (Area Under Curve - Receiver Operating Characteristic) [22, 33, 34] was used to evaluate the model's discrimination ability across all classes, providing deeper insights into classification confidence and separability. Confusion matrices and ROC curves were also generated to complement the numerical evaluation, enabling a visual interpretation of the model's classification behavior across emotional categories. This comprehensive evaluation framework ensures that the model is both accurate and reliable, as well as generalizable for female speech emotion recognition tasks.

4. RESULTS

The proposed hybrid model, comprising a CNN, BiLSTM, and Transformer, was trained and evaluated using a stratified 5-fold cross-validation technique to ensure consistent performance across all emotional classes. Each fold employed a balanced dataset, which had been previously augmented and resampled using noise injection, pitch shifting, and SMOTE. Feature extraction was conducted progressively, as outlined in Table 3, resulting in five distinct datasets with varying acoustic feature combinations, ranging from MFCC only to MFCC + ZCR + LPC + RMSE + ZCPA. Each dataset was trained independently using the same model architecture and hyperparameter settings to ensure a fair comparison.

To assess the computational efficiency of each feature configuration, the training time for every model was recorded and summarized in Table 7.

Table 7. Training time

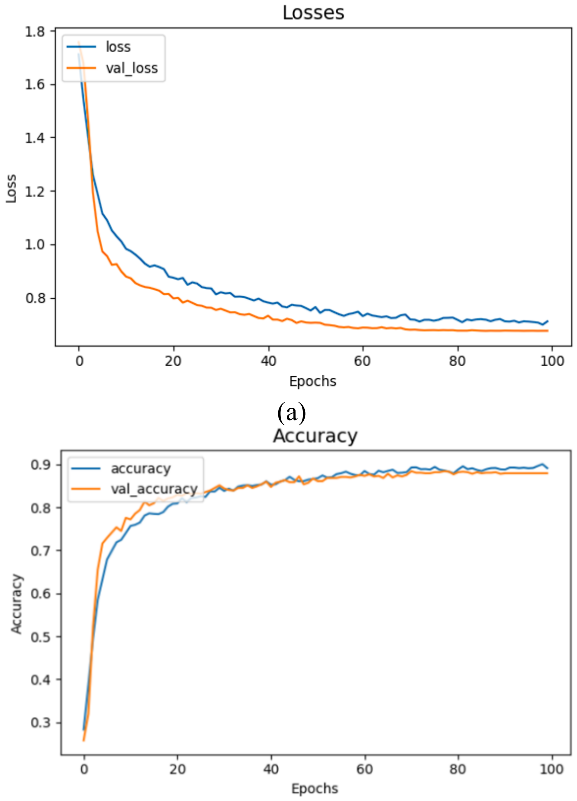
Model	Method	Training Time
Model 1	MFCC	00:03:05
Model 2	MFCC+ZCR	00:02:51
Model 3	MFCC+ZCR+LPC	00:02:55
Model 4	MFCC+ZCR+LPC+RMSE	00:02:57
Model 5	MFCC+ZCR+LPC+RMSE+ZCPA	00:02:55

Table 7 displays the training time required for each model, which was developed using different combinations of acoustic features. Model 1, which used only MFCC, recorded the longest training duration at 3 minutes and 5 seconds. Interestingly, Model 2, which combined MFCC and ZCR, achieved the shortest training time at 2 minutes and 51 seconds, despite incorporating an additional feature. The training durations for Models 3 to 5, which included progressively more features (LPC, RMSE, and ZCPA), remained consistent and showed only minor fluctuations, ranging between 2 minutes and 55 seconds and 2 minutes and 57 seconds. These results suggest that adding acoustic features has a minimal impact on training efficiency, confirming that the proposed hybrid architecture remains computationally stable even as input complexity increases.

Table 8. Training and validation report

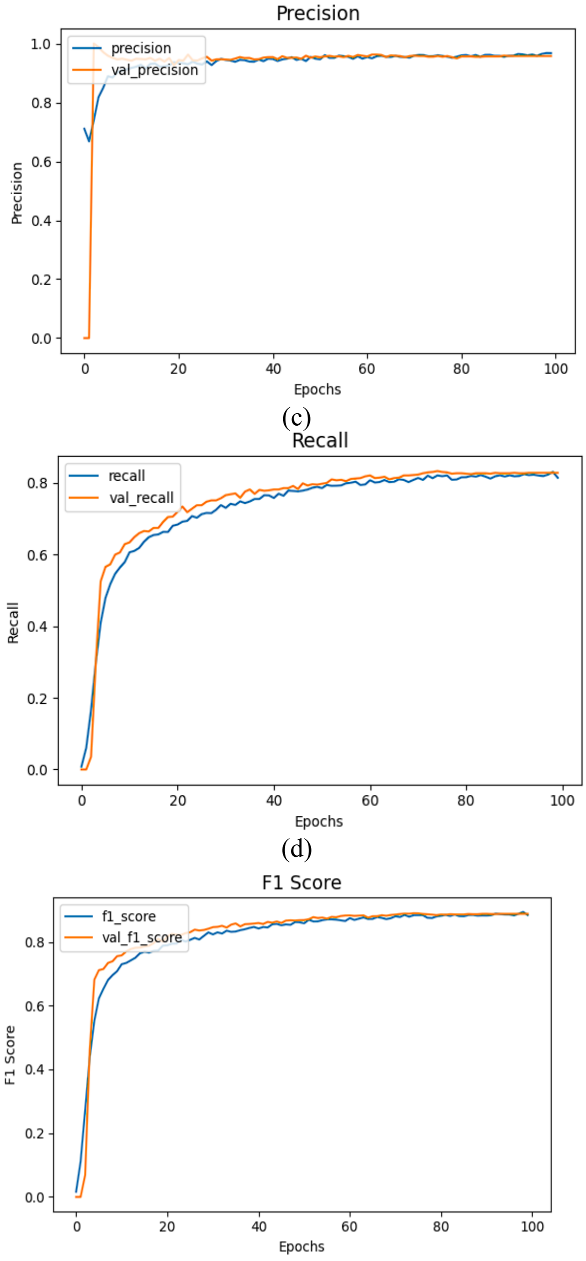
Model	Accuracy (%)	F1-Score (%)	Loss	Precision (%)	Recall (%)	Val Accuracy (%)	Val F1-Score (%)	Val Loss	Val Precision (%)	Val Recall (%)
Model 1	88.78	87.81	0.73	96.02	80.89	87.67	88.32	0.67	96.12	81.68
Model 2	88.49	88.41	0.72	96.47	81.59	87.91	88.80	0.68	95.76	82.78
Model 3	89.05	88.61	0.72	95.26	82.84	88.03	88.10	0.68	94.80	82.30
Model 4	90.18	89.25	0.70	96.02	83.37	86.69	87.83	0.68	94.01	82.42
Model 5	89.90	87.86	0.72	94.73	81.93	87.91	88.59	0.68	94.47	83.39

The training results in Table 8 consistently demonstrate high performance across all models, with training accuracy exceeding 88% and validation accuracy ranging from 86% to 88%. Model 4 (MFCC + ZCR + LPC + RMSE) achieved the highest training accuracy of 90.18% and the highest F1-score of 89.25%, but it also showed the lowest validation accuracy of 86.69%, indicating possible overfitting. In contrast, Model 2 (MFCC + ZCR) exhibited the best balance between training and validation performance, achieving the highest validation F1-score of 88.80%, suggesting that this feature combination is both effective and efficient. Loss values and other metrics remained relatively stable across all models, highlighting the robustness and consistency of the proposed hybrid architecture. The training and validation history of the best-performing model, Model 2, is presented in Figure 2, providing a detailed overview of its learning behavior throughout the training process.



(a)

(b)



(c)

(d)

(e)

Figure 2. Model 2 training & validation history: (a) Loss, (b) Accuracy, (c) Precision, (d) Recall, (e) F1-score

The training and validation results in Figure 2 demonstrate that the proposed hybrid model converges stably, decreasing loss and consistently increasing performance across all metrics. The model achieves high accuracy, near-perfect precision, strong recall, and an F1-score close to 0.9, indicating robust generalization and reliable performance in female speech emotion recognition.

To provide a clearer understanding of each model's classification performance, the evaluation results are illustrated in Figure 3. Figure 3 presents a bar chart comparing the accuracy, precision, recall, and F1-score across all five models, highlighting the relative effectiveness of each feature combination.

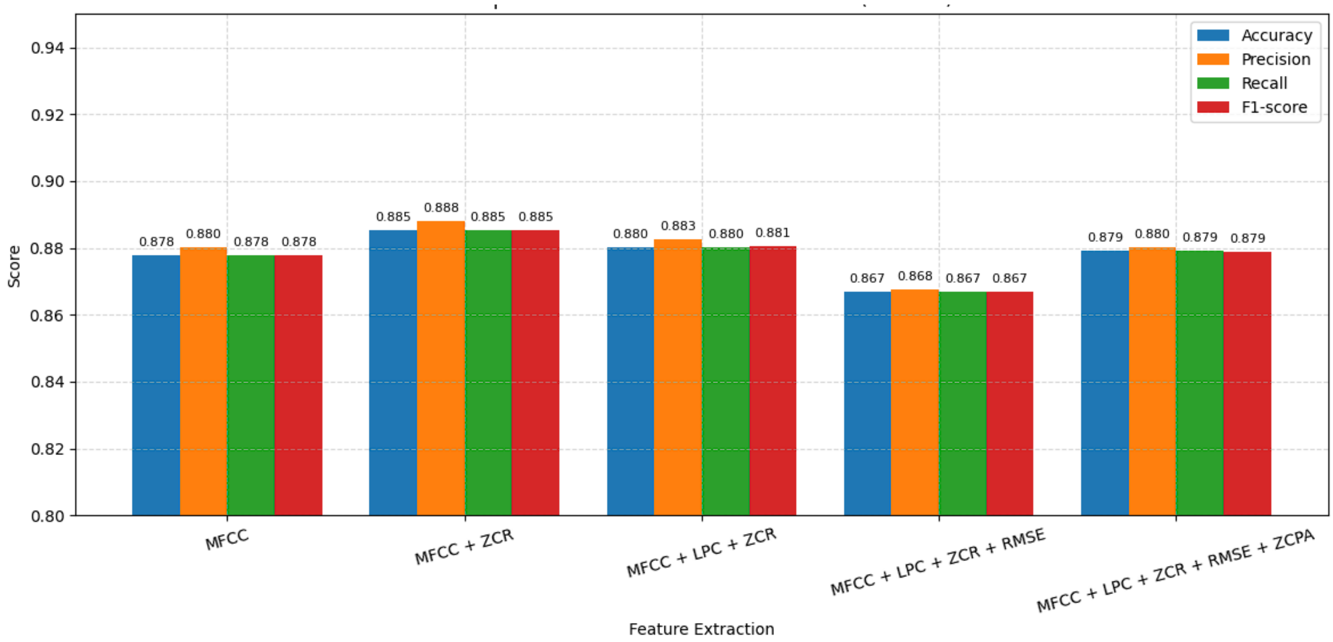


Figure 3. Model evaluation comparison report

The evaluation results, illustrated in Figure 3, highlight that Model 2, which utilizes the feature combination of MFCC and ZCR, consistently outperforms the other models across all key metrics. It achieves the highest accuracy (88.52%), precision (88.80%), recall (88.52%), and F1-score (88.53%), indicating a strong balance between predictive performance and generalization. Other models yield relatively competitive results, but none surpass Model 2 in terms of performance and training efficiency. This confirms that the MFCC + ZCR feature configuration provides an optimal trade-off between computational cost and classification accuracy in the proposed SER system.

In addition to the proposed hybrid CNN-BiLSTM-Transformer architecture, several baseline models were also evaluated to provide a fair performance comparison. These baseline architectures include standalone CNN, BiLSTM, and a combined CNN-BiLSTM model without the Transformer component. The purpose of this comparative analysis is to examine the individual contribution of each network type and to assess whether the integration of convolutional, recurrent, and attention mechanisms offers a measurable improvement in emotion recognition accuracy.

To ensure consistency, the comparison was performed using the best-performing feature combination, namely MFCC and ZCR, which had previously yielded the highest accuracy in the hybrid model. Accordingly, all baseline models were trained and tested using the same MFCC + ZCR feature dataset to ensure that performance differences reflect architectural effectiveness rather than variations in features. The overall comparison framework is illustrated in Figure 4.

As presented in Figure 4, the proposed hybrid CNN-BiLSTM-Transformer model achieved the highest classification accuracy compared to the non-hybrid models. The inclusion of the Transformer component resulted in a significant improvement in performance, demonstrating its effectiveness in capturing long-range temporal dependencies and contextual relationships within speech signals. In contrast, the CNN-BiLSTM and BiLSTM architectures, which rely primarily on sequential modeling, yielded moderate accuracy. The standalone CNN model, however, showed the lowest performance due to its limited ability to model temporal dynamics. These findings demonstrate that the hybrid integration of convolutional, recurrent, and attention-based mechanisms provides a more comprehensive feature representation for emotion recognition tasks.

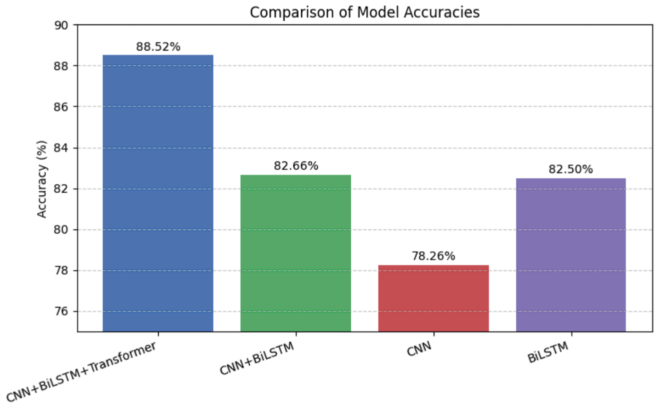


Figure 4. Comparison of baseline model accuracies

The classification effectiveness of the best-performing model, Model 2 (CNN-BiLSTM-Transformer), is further analyzed using a confusion matrix, as depicted in Figure 5. This matrix offers detailed insights into the model's ability to accurately identify each emotion class.

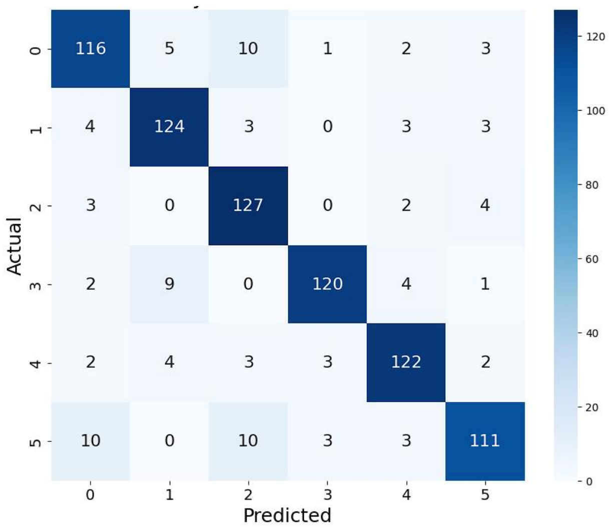


Figure 5. Confusion matrix analysis of the proposed MFCC and ZCR feature combination (Model 2)

As shown in Figure 5, the model demonstrates strong performance across all six emotion classes, with most predictions concentrated along the diagonal, indicating high true positive rates. Minor misclassifications are observed; for instance, some class 0 and class 5 samples were incorrectly predicted as other classes. Nevertheless, the confusion matrix confirms that Model 2 can distinguish between subtle emotional expressions in female speech with minimal error.

The final evaluation metric used to assess the discriminative ability of each model is the AUC-ROC (Area Under Curve-Receiver Operating Characteristic). This metric reflects the model's ability to distinguish between different emotion classes, regardless of classification threshold. The AUC-ROC scores for all five models are summarized in Table 9.

Table 9. AUC-ROC report

Model	AUC-ROC Score (%)
Model 1	98.85
Model 2	98.95
Model 3	98.84
Model 4	98.72
Model 5	98.81

As shown in Table 9, Model 2 again achieves the highest AUC-ROC score of 98.95%, indicating excellent class separation and confirming its superiority across multiple evaluation criteria. The consistently high AUC values across all models suggest that the hybrid architecture is highly effective in learning emotional patterns from female speech data.

To provide a more detailed evaluation of the model's classification performance for each emotion category, the AUC-ROC curves per class for the best-performing model (Model 2) are presented in Figure 6. These curves illustrate the trade-off between the true positive rate (sensitivity) and the false positive rate for each class.

As shown in Figure 6, the model exhibits consistently high AUC scores across all six emotion classes. Five classes, neutral, calm, sad, happy, and fear, achieve an AUC of 0.99, while the disgust class attains a slightly lower AUC of 0.98. These results confirm that the hybrid model can effectively distinguish between emotion categories, demonstrating excellent generalization and robustness in emotion classification tasks.

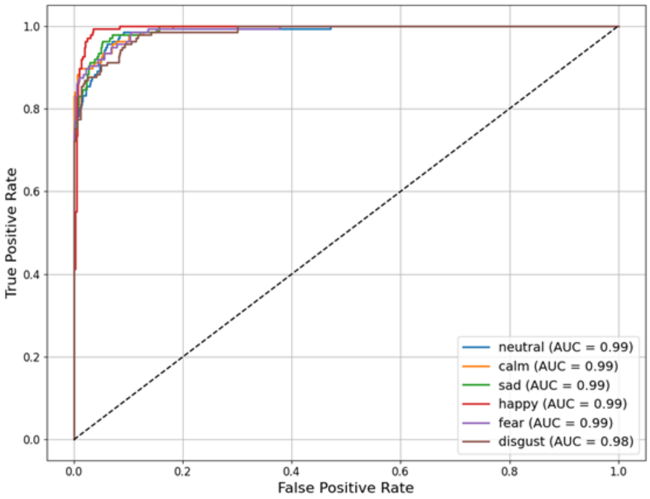


Figure 6. AUC-ROC of the proposed MFCC and ZCR feature combination (Model 2)

5. CONCLUSION

This study successfully developed a hybrid deep learning model based on a CNN-BiLSTM-Transformer architecture designed for recognizing female speech emotions. By implementing a stepwise feature extraction approach that includes MFCC, ZCR, LPC, RMSE, and ZCPA, the model effectively captures complex acoustic patterns relevant to emotional expressions in female voice signals.

The experiments were conducted using an augmented dataset, which was enhanced through pitch shifting and additive noise techniques, and then balanced using the SMOTE algorithm. Training results indicated that Model 2, which utilized the MFCC and ZCR feature combination, achieved the best performance with an accuracy of 88.52% and an AUC-ROC score of 98.95%. Further evaluation using the confusion matrix and per-class ROC curves demonstrated the model's capability to accurately and consistently distinguish between different emotional states.

Additionally, a comparison with baseline architectures (CNN, BiLSTM, and CNN-BiLSTM) was conducted to assess the contribution of each component in the hybrid framework. The results showed that the proposed CNN-BiLSTM-Transformer model outperformed the baseline models, confirming that the integration of convolutional, recurrent, and attention-based mechanisms significantly improves the ability to model both local spectral patterns and long-range temporal dependencies in emotional speech.

Overall, the findings suggest that integrating a hybrid architecture with a stepwise feature extraction and pitch-based augmentation strategy significantly enhances the performance of SER systems, particularly for female speech. This work lays a strong foundation for developing more inclusive and practical emotion recognition systems for real-world human-computer interaction.

However, the model still has certain limitations regarding generalizability and environmental robustness. Since the evaluation was performed on controlled datasets, the performance in real-world conditions, such as those with background noise, varied accents, or spontaneous speech, may differ. Additionally, the absence of cross-dataset validation limits the assessment of the model's transferability to other domains. Future work should focus on cross-corpus testing, noise-robust feature learning, and multi-language adaptation to further enhance the scalability and general applicability of the proposed model.

REFERENCES

[1] Khare, S.K., Blanes-Vidal, V., Nadimi, E.S., Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. Information Fusion, 102: 102019. https://doi.org/10.1016/j.inffus.2023.102019

[2] Jafari, M., Shoeibi, A., Khodatars, M., Bagherzadeh, S., et al. (2023). Emotion recognition in EEG signals using deep learning methods: A review. Computers in Biology and Medicine, 165: 107450. https://doi.org/10.1016/j.compbiomed.2023.107450

[3] Hansen, L., Zhang, Y.P., Wolf, D., Sechidis, K., Ladegaard, N., Fusaroli, R. (2022). A generalizable speech emotion recognition model reveals depression and remission. Acta Psychiatrica Scandinavica, 145(2): 186-199. https://doi.org/10.1111/acps.13388

[4] Lin, Y.C., Wu, H., Chou, H.C., Lee, C.C., Lee, H.Y. (2024). Emo-bias: A large scale evaluation of social bias on speech emotion recognition. In Interspeech 2024, Kos, Greece, pp. 4633-4637. https://doi.org/10.21437/Interspeech.2024-1073

[5] Tursunov, A., Mustaqeem, Choeh, J.Y., Kwon, S. (2021). Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. Sensors, 21(17): 5892. https://doi.org/10.3390/s21175892

[6] Anvarjon, T., Mustaqeem, Kwon, S. (2020). Deep-Net: A lightweight CNN-based speech emotion recognition system using deep frequency features. Sensors, 20(18): 5212. https://doi.org/10.3390/s20185212

[7] Aldeneh, Z., Provost, E. M. (2021). You're not you when you're angry: Robust emotion features emerge by recognizing speakers. IEEE Transactions on Affective Computing, 14(2): 1351-1362. https://doi.org/10.1109/TAFFC.2021.3086050

[8] Kim, S., Lee, S.P. (2023). A BiLSTM–transformer and 2D CNN architecture for emotion recognition from speech. Electronics, 12(19): 4034. https://doi.org/10.3390/electronics12194034

[9] Gomathy, M. (2021). Optimal feature selection for speech emotion recognition using enhanced cat swarm optimization algorithm. International Journal of Speech Technology, 24(1): 155-163. https://doi.org/10.1007/s10772-020-09776-x

[10] Kacur, J., Puterka, B., Pavlovicova, J., Oravec, M. (2021). On the speech properties and feature extraction methods in speech emotion recognition. Sensors, 21(5): 1888. https://doi.org/10.3390/s21051888

[11] Zhao, J., Mao, X., Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control, 47: 312-323. https://doi.org/10.1016/j.bspc.2018.08.035

[12] Finch, E. (2024). Speech emotion recognition with dynamic CNN and Bi-LSTM. Journal of Computer Technology and Software, 3(6): 82.

[13] Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R. (2014). CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing, 5(4): 377-390. https://doi.org/10.1109/TAFFC.2014.2336244

[14] Livingstone, S.R., Russo, F.A. (2018). 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, 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391

[15] Pichora-Fuller, M.K., Dupuis, K. (2020). Toronto emotional speech set (TESS). Scholars Portal Dataverse. https://doi.org/10.5683/SP2/E8H2MF

[16] Dablain, D., Krawczyk, B., Chawla, N.V. (2022). DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. IEEE Transactions on Neural Networks and Learning Systems, 34(9): 6390-6404. https://doi.org/10.1109/TNNLS.2021.3136503

[17] Konduru, A.K., Mazher Iqbal, J.L. (2024). Emotion recognition from speech signals using digital features optimization by diversity measure fusion. Journal of Intelligent & Fuzzy Systems, 46(1): 2547-2572. https://doi.org/10.3233/JIFS-231263

[18] Ahmed, M.R., Islam, S., Islam, A.M., Shatabda, S. (2023). An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition. Expert Systems with Applications, 218: 119633. https://doi.org/10.1016/j.eswa.2023.119633

[19] Prabakaran, D., Sriuppili, S. (2021). Speech processing: MFCC based feature extraction techniques-an investigation. Journal of Physics: Conference Series, 1717(1): 012009. https://doi.org/10.1088/1742-6596/1717/1/012009

[20] Nguyen, T., Shu, R., Pham, T., Bui, H., Ermon, S. (2021). Temporal predictive coding for model-based planning in latent space. arXiv preprint arXiv:2106.07156. https://doi.org/10.48550/arXiv.2106.07156

[21] Patnaik, S. (2023). Speech emotion recognition by using complex MFCC and deep sequential model. Multimedia Tools and Applications, 82(8): 11897-11922. https://doi.org/10.1007/s11042-022-13725-y

[22] Bhardwaj, V., Thakur, D., Gera, T., Sharma, V. (2023). Enhanced dialectal speech recognition in Punjabi using pitch-based acoustic modeling. Ingénierie des Systèmes d'Information, 28(6): 1557-1563. https://doi.org/10.18280/isi.280612

[23] Gorriz, J.M., Segovia, F., Ramirez, J., Ortiz, A., Suckling, J. (2024). Is K-fold cross validation the best model selection method for Machine Learning? arXiv preprint arXiv:2401.16407. https://doi.org/10.48550/arXiv.2401.16407

[24] Ketkar, N., Moolayil, J., Ketkar, N., Moolayil, J. (2021). Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch. Apress.

[25] Talai, Z., Kherici, N., Bahi, H. (2023). Comparative study of CNN structures for Arabic speech recognition. Ingénierie des Systèmes d'Information, 28(2): 327-333. https://doi.org/10.18280/isi.280208

[26] Alkhawaldeh, R.S., Al-Ahmad, B., Ksibi, A., Ghatasheh, N., et al. (2023). Convolution neural network bidirectional long short-term memory for heartbeat arrhythmia classification. International Journal of Computational Intelligence Systems, 16(1): 197. https://doi.org/10.1007/s44196-023-00374-8

[27] Lui, C. F., Liu, Y., Xie, M. (2022). A supervised bidirectional long short-term memory network for data-driven dynamic soft sensor modeling. IEEE Transactions on Instrumentation and Measurement, 71: 1-13. https://doi.org/10.1109/TIM.2022.3152856

[28] Chitty-Venkata, K.T., Mittal, S., Emani, M., Vishwanath, V., Somani, A.K. (2023). A survey of techniques for optimizing transformer inference. Journal of Systems Architecture, 144: 102990. https://doi.org/10.1016/j.sysarc.2023.102990

[29] Gordon-Rodriguez, E., Loaiza-Ganem, G., Pleiss, G., Cunningham, J.P. (2020). Uses and abuses of the cross-entropy loss: Case studies in modern deep learning. arXiv preprint arXiv:2011.05231. https://doi.org/10.48550/arXiv.2011.05231

[30] Ariff, N.A.M., Ismail, A.R. (2023). Study of Adam and adamax optimizers on alexnet architecture for voice biometric authentication system. In 2023 17th International Conference on Ubiquitous Information Management and Communication (IMCOM), Seoul, Korea, pp. 1-4. https://doi.org/10.1109/IMCOM56909.2023.10035592

[31] Tharwat, A. (2021). Classification assessment methods. Applied Computing and Informatics, 17(1): 168-192. https://doi.org/10.1016/j.aci.2018.08.003

[32] Kurniadi, D., Fernando, E., Fauziyah, A., Mulyani, A. (2025). Improving low-light face recognition using DeepFace embedding and multi-layer perceptron. Jurnal RESTI, 9(5):1047-1055. https://doi.org/10.29207/resti.v9i5.6797

[33] Carrington, A.M., Manuel, D.G., Fieguth, P.W., Ramsay, T., et al. (2022). Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1): 329-341. https://doi.org/10.1109/TPAMI.2022.3145392

[34] Saini, D.K.J.B., Shieh, C., Sankpal, L.J., Mehrotra, M., Bhosale, K.S., Raut, Y. (2025). Deep learning-based optimized model for emotional psychological disorder activities identification in smart healthcare system. Journal of Research, Innovation and Technologies, 4(2): 143-157. https://doi.org/10.57017/jorit.v4.2(8).02