Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 5, October 2025, Page. 3352-3367
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5143

Hybrid Model for Speech Emotion Recognition using Mel-Frequency Cepstral Coefficients and Machine Learning Algorithms

Odi Nurdiawan[*1], Dian Ade Kurnia[2], Dadang Sudrajat[3], Irfan Pratama[4]

[1,2]Informatics Management, STMIK IKMI Cirebon, Indonesia
[3]Informatics Engineering, STMIK IKMI Cirebon, Indonesia
[4]Information System, Mercu Buana University Yogyakarta, Indonesia

Email: [1]odinurdiawan2020@gmail.com

Received : Jul 19, 2025; Revised : Sep 8, 2025; Accepted : Sep 23, 2025; Published : Oct 16, 2025

Abstract

Speech Emotion Recognition (SER) is a subfield of *affective computing* that focuses on identifying human emotions through voice signals. Accurate emotion classification is essential for developing intelligent systems capable of interacting naturally with users. However, challenges such as background noise, overlapping emotional features, and speaker variability often reduce model performance. This study aims to develop a lightweight hybrid SER model by combining *Mel-Frequency Cepstral Coefficients* (MFCC) as feature representations with three machine learning algorithms: Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbors (KNN). The methodology involves audio data preprocessing, MFCC-based feature extraction, and classification using the selected algorithms. The RAVDESS dataset, consisting of 1,440 English-language audio samples across four emotions (happy, angry, sad, neutral), was used with an 80/20 train-test split to ensure class balance.. Experimental results show that the KNN model achieved the highest performance, with an accuracy of 78.26%, precision of 85.09%, recall of 78.26%, and F1-score of 77.06%. The Decision Tree model produced comparable results, while the SVM model performed poorly across all metrics. These findings demonstrate that the proposed hybrid approach is effective for recognizing emotions in speech and offers a computationally efficient alternative to deep learning models. The integration of MFCC features with multiple machine learning classifiers provides a robust framework for real-time emotion recognition applications, especially in environments with limited computing resources.

Keywords : *Affective Computing, Audio Classification, Decision Tree, K-Nearest Neighbors, MFCC, Speech Emotion Recognition.*

1. INTRODUCTION

Speech Emotion Recognition (SER) is a branch of emotional computing that focuses on the identification and classification of human emotions through voice signals. Emotions play a significant role in interpersonal communication, and the ability of computational systems to recognize these emotions enables more intelligent, empathetic, and human-like interactions between humans and machines [1], [2], [3]. SER technology typically employs feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCC), pitch, and chroma to capture the acoustic and emotional patterns present in speech. MFCC is among the most commonly used features due to its ability to efficiently represent the spectral characteristics of the human voice. The implementation of SER technology spans multiple domains, including voice-based customer service systems, adaptive virtual assistants, learning systems that respond to students' emotional states, and early detection of psychological conditions in the field of mental health[4], [5], [6], [7].

Despite the rapid development of Speech Emotion Recognition (SER) technology, there are still several technical challenges that remain key concerns in its advancement. One of the main challenges is the presence of noise or acoustic disturbances from the surrounding environment, which can affect the

Vol. 6, No. 5, October 2025, Page. 3352-3367
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5143

quality of voice signals and reduce classification accuracy. In addition, emotional overlapping namely, the similarity of acoustic characteristics between one emotion and another makes it difficult for the model to perform accurate classification [8] [9], [10]. Other influencing factors include variations in speaking style, intonation, accent, and differences in emotional expression across individuals and cultures. Conventional approaches such as Support Vector Machine (SVM) and Decision Tree (DT) often fail to capture the temporal dynamics within speech signals, thus requiring more complex and adaptive approaches. In this context, the use of hybrid models such as CNN-LSTM has demonstrated superior performance due to its ability to combine the strengths of spatial pattern extraction (through CNN) and temporal pattern modeling (through LSTM). A study by Shaik et al. (2025) showed that the hybrid model achieved an accuracy of up to 89.4% and demonstrated robustness against noise interference and the complexity of emotional expressions, making it more suitable for real-world applications.

Research on Speech Emotion Recognition (SER) has progressed rapidly in line with the growing demand for intelligent systems capable of effectively recognizing and responding to human emotions. SER typically relies on acoustic features such as Mel-Frequency Cepstral Coefficients (MFCC), pitch, and chroma to extract emotional patterns from voice signals. Hybrid deep learning models, such as the combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), have become a popular approach due to their ability to simultaneously handle spatial and temporal dimensions. [8] developed a CNN-LSTM model and evaluated its performance on the RAVDESS and EMO-DB datasets, achieving an accuracy of 89.4% and demonstrating robustness against noise and emotional overlap. Similar studies by [1], [9], [11], [12], [13] also showed that the integration of CNN and LSTM could improve emotion classification accuracy compared to conventional models. Previous research has primarily focused on single models or hybrid deep learning approaches. However, limited studies have examined ensemble-based machine learning frameworks that combine multiple algorithms with MFCC features for SER. Most works also rely on English-language datasets, with little exploration of models applicable to multilingual or resource-constrained contexts. This creates a gap for developing computationally efficient and interpretable solutions. In this study, the term “hybrid model” explicitly refers to an ensemble of multiple machine learning algorithms (SVM, DT, and KNN) combined through majority voting, rather than hybrid deep learning architectures such as CNN-LSTM.

In a broader context, several researchers have explored novel approaches by incorporating additional features or employing more complex architectural techniques. [14], [15] introduced CochleaSpecNet, a hybrid CNN-GRU model with multi-head attention that utilizes cochleagram and spectrogram features, achieving high accuracy on the BanglaSER and RAVDESS datasets. Another approach involving data augmentation was proposed by [16], [17] through the Wasserstein GAN-LSTM model, specifically designed for limited data scenarios. Meanwhile, [18], [19] introduced a transfer learning concept based on a combination of classical and quantum neural networks, which achieved an accuracy of up to 98.93% using the TESS dataset. The use of a genetic algorithm for feature selection in high-dimensional datasets was also explored by [20], [21], which successfully improved model efficiency and performance.

In addition, various studies have compared the performance of CNN, LSTM, and hybrid approaches under diverse linguistic and demographic conditions. [22], [23]demonstrated the superiority of hybrid models in emotion recognition for the Amazigh language, involving gender and age variables. [24], [25]developed a Punjabi dataset and utilized 1D CNN for classification following feature selection using the LASSO algorithm. [23], [26], [27] also incorporated a multilingual and multimodal approach through the hybrid IChOA-CNN-LSTM model. [28] as well as [29], emphasized the importance of data augmentation in maintaining model accuracy in real-time scenarios. A review by [30]concluded that

Vol. 6, No. 5, October 2025, Page. 3352-3367
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5143

recent trends highlight hybrid models as a promising solution to improve accuracy, robustness, and generalization in automatic speech emotion recognition.

Most previous studies on Speech Emotion Recognition (SER) have primarily focused on the use of single models such as CNN, LSTM, or traditional approaches like SVM and Decision Tree. These approaches tend to have limitations in handling the dynamic complexity of emotional data, especially under conditions involving noise or emotional variability across speakers. Although some studies have proposed hybrid models such as CNN-LSTM [8] , they generally remain confined to the deep learning domain and have yet to extensively explore the integration of ensemble-based machine learning models, such as combinations of SVM, Decision Tree, and K-Nearest Neighbors. Furthermore, the majority of studies continue to rely on English-language datasets such as EMO-DB [31]. To date, there is still a lack of research explicitly examining the performance of hybrid models based on machine learning algorithms combined with MFCC features for speech emotion recognition. Hybrid models that integrate multiple machine learning algorithms have the potential to improve classification accuracy, particularly in managing speaker variability and acoustic noise. Therefore, this study aims to address these gaps by developing and evaluating an MFCC-based hybrid SER model that integrates several machine learning algorithms (SVM, Decision Tree, and KNN) and testing it comprehensively using a structured speech emotion dataset.

Several issues remain suboptimally addressed. First, most studies still rely on single-model approaches that lack sufficient flexibility in handling data with varying emotional expressions, background noise, or diverse accents. Second, there is a lack of research that systematically integrates multiple machine learning algorithms into a single hybrid model based on MFCC features, specifically designed for speech emotion classification. Therefore, the urgency of this research lies in the need to develop a lightweight, effective, and comprehensively tested hybrid model for speech emotion classification based on MFCC features and adaptive machine learning approaches.

To address the identified research gaps, this study adopts a hybrid approach by combining several machine learning algorithms, including Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbors (KNN). The selection of these algorithms is based on their respective strengths in classifying non-linear data, model interpretability, and sensitivity to local data structures. *Mel-Frequency Cepstral Coefficients* (MFCC) are used as the primary feature representation of speech signals, as they have been proven effective in capturing the acoustic characteristics of human speech [8], [32], [33]. Compared to deep learning approaches, this method offers advantages in terms of computational efficiency, faster training time, and ease of result interpretation. The model will be evaluated using metrics such as accuracy, precision, recall, F1-score, and Receiver Operating Characteristic (ROC-AUC) to ensure its reliability and generalization capability for multi-emotion data. In addition, the use of an Indonesian-language dataset will provide a valuable contribution to the development of contextually and locally relevant SER systems.

The novelty of this study lies in the integration of a hybrid model based on machine learning algorithms (SVM, DT, and KNN) within a Speech Emotion Recognition (SER) system using MFCC features, specifically designed for emotion classification. In contrast to most previous studies that rely on single-model approaches or complex deep learning [34], [35], [36], this research offers a computationally lightweight solution that remains accurate and adaptive to various acoustic conditions.

This study aims to develop a hybrid model for a Speech Emotion Recognition (SER) system by combining Mel-Frequency Cepstral Coefficients (MFCC) features with several machine learning algorithms, namely Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbors (KNN). Through the integration of these methods, the study seeks to improve the accuracy of speech emotion classification. Additionally, this research is intended to analyze the performance of the hybrid

Vol. 6, No. 5, October 2025, Page. 3352-3367
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5143

model in comparison to individual models based on evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

2. METHOD

The methodological design of this study focuses on developing a speech emotion classification model based on machine learning algorithms configured in a hybrid framework[5], [37]. The research involves several main stages, starting with the collection and preparation of a speech dataset, followed by audio signal preprocessing to enhance data quality. Subsequently, feature extraction is performed using Mel-Frequency Cepstral Coefficients, which effectively capture the characteristics of human speech relevant to emotional content. The extracted features are then used to build classification models using three different algorithms: Support Vector Machine, Decision Tree, and K-Nearest Neighbors. Finally, the performance of each model both individually and in hybrid combinations is evaluated using classification metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to determine the most effective approach for recognizing emotional states from speech signals.

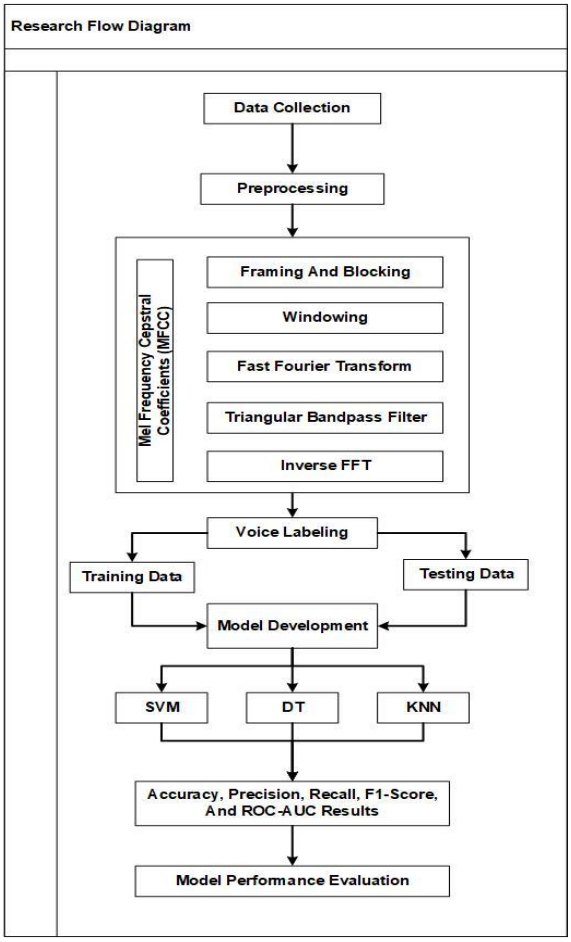

Figure 1. Research Workflow

2.1 Data collection

The dataset used in this research is the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), which contains 1,440 audio samples labeled with four emotions: happy, angry, sad, and neutral. The dataset was divided into 80% training data and 20% testing data using a stratified split to maintain class balance. This ensures that each emotional category is proportionally represented in both subsets.

Vol. 6, No. 5, October 2025, Page. 3352-3367
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5143

Data collection in this study utilized the RAVDESS dataset (Ryerson Audio-Visual Database of Emotional Speech and Song), which is one of the benchmark datasets commonly used in speech emotion recognition (SER) research. This dataset consists of audiovisual recordings containing emotional expressions from professional actors in English. The RAVDESS dataset categorizes emotions into neutral, happy, sad, and angry. The audio files in RAVDESS are available in high-quality .wav format (48 kHz, 16-bit), making them highly suitable for feature extraction processes such as Mel-Frequency Cepstral Coefficients (MFCC). The following are sample wavplots of happy, angry, sad, and neutral voices:

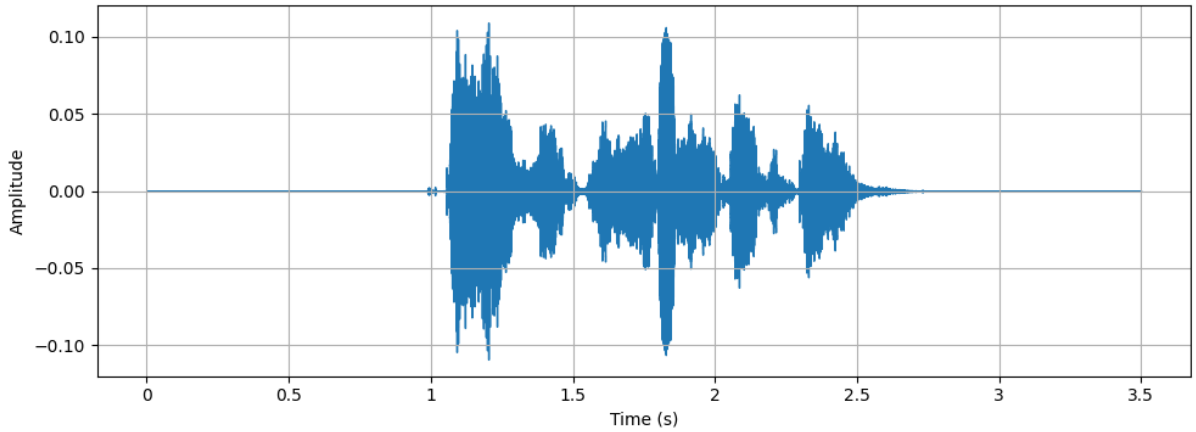

Figure 2. Angry Waveplot

Figure 2. Angry Waveplot illustrates the waveform visualization of an audio signal expressing anger. The horizontal axis represents time (seconds), while the vertical axis indicates the signal amplitude. Active speech appears between the 1st and 2.7th seconds, with sharp and fluctuating amplitudes, reflecting the high intensity typical of angry emotions. The waveform pattern shows strong vocal pressure and an unstable rhythm, which are characteristic of angry vocal expressions. This visualization is important in signal analysis for distinguishing emotions in speech recognition systems.

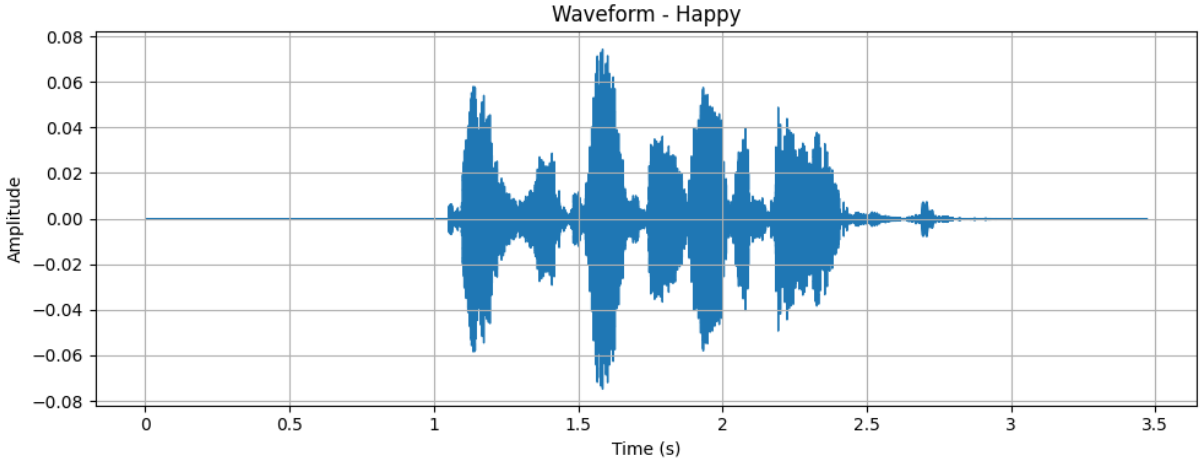

Figure 3. Happy Waveplot

Figure 3 shows the waveplot of a speech signal with a happy emotion. Vocal activity occurs between the 1st and 2.7th second, with smoother and more regular amplitude compared to the angry emotion. This waveform pattern reflects light, rhythmic, and stable intonation characteristic of a happy expression. This visualization is important for identifying the acoustic characteristics that distinguish happy emotions in the speech emotion recognition process.

Vol. 6, No. 5, October 2025, Page. 3352-3367
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5143

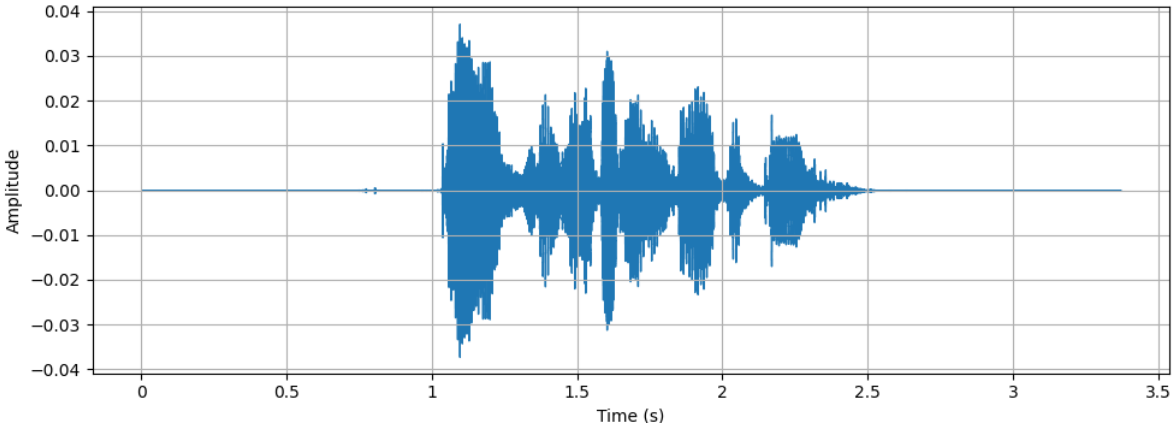

Figure 4. Sad Waveplot

Figure 4. The sad waveplot shows the waveform of a voice expressing sadness. Voice activity occurs between the 1st and 2.5th seconds, with low amplitude and smooth fluctuations. This pattern reflects a soft and flat tone, typical of sad expressions. Such a waveform is important for distinguishing sadness in speech emotion recognition systems

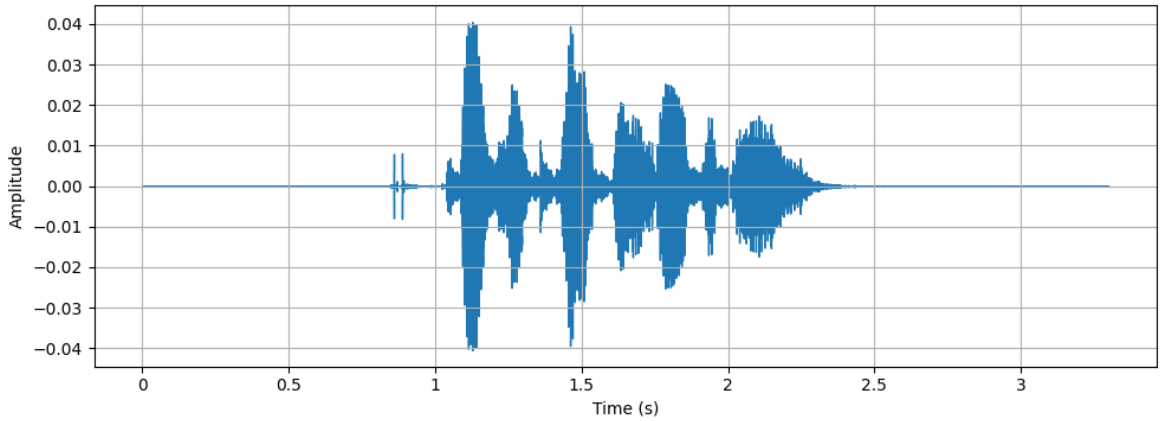

Figure 5. Neutral Waveplot

Figure 5. The waveplot shows the waveform of speech with a neutral emotion. Voice activity occurs between the 1st and 2.4th seconds, with moderate amplitude and stable fluctuations. This pattern reflects a flat and calm tone, typical of neutral expressions, and is useful for distinguishing unemotional speech in a speech emotion recognition system

2.2 Preprocessing

Preprocessing was applied to improve the quality of the audio signals before feature extraction. The steps included:

1. Noise reduction – to minimize background interference and simulate real-world recording conditions.
2. Resampling at 16 kHz – a standard frequency in speech processing that balances signal quality and computational efficiency.
3. Amplitude normalization – to ensure uniform scaling of signal intensity across samples.
4. Silence removal – to eliminate non-speech segments, allowing the analysis to focus on meaningful acoustic information.

Vol. 6, No. 5, October 2025, Page. 3352-3367
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5143

2.3 Feature Extraction: Mel Frequency Cepstral Coefficients (MFCC)

MFCC was employed as the main feature extraction method, as it effectively represents the spectral properties of speech signals. The process involved framing, windowing with a Hamming window, applying Fast Fourier Transform (FFT), and mapping the spectrum onto the Mel scale using triangular filters. Finally, cepstral analysis was conducted to generate 13 coefficients per frame. For each audio file, the mean and standard deviation of these coefficients were calculated, resulting in a 26-dimensional feature vector.

The MFCC transformation can be defined as:

$$MFCC(n) = \sum_{k-1}^{K} \log(E_k)\cos\left[n(k - \frac{1}{2})\frac{n}{k}\right]$$

where E_k represents the energy of the k-th Mel filter bank.

Model Development and Hybrid Mechanism

In this stage, classification models were built using three algorithms: Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbors (KNN).

1. The SVM classifier was configured with an RBF kernel, C = 1.0, and γ = 0.01.
2. The DT model was built with a maximum depth of 10 using the Gini criterion.
3. The KNN classifier used k = 5 neighbors with Euclidean distance.

The hybrid framework integrates predictions from SVM, DT, and KNN using a majority voting mechanism, where each classifier contributes equally to the final decision. This ensemble approach is designed to enhance classification robustness by leveraging the complementary strengths of individual models[7].

The initial step in the data preprocessing stage is to check for the presence of missing values in the dataset. Missing values are data entries that lack required information, and their presence can affect both the validity and performance of the machine learning model being developed.

The extract_mfcc function is a crucial procedure in the audio feature extraction stage, particularly for analyzing voice signals in the domain of emotion recognition. This function takes an input in the form of an audio file path (file_path) and processes it using the librosa library. In the first step, the audio is loaded into a numerical representation using librosa.load(), which produces two main components: y as the audio time series and sr as the sampling rate.

Next, Mel-Frequency Cepstral Coefficients (MFCC) are extracted using librosa.feature.mfcc, with the parameter n_mfcc=13, indicating that 13 MFCCs are generated to represent the spectral characteristics of the voice signal. MFCCs are important features in speech signal processing because they mimic how the human ear perceives sound frequencies, making them effective for representing emotional information in the signal.

To enhance the feature description, the mean and standard deviation of the 13 MFCCs are calculated along the time dimension (axis=1). The mean captures the central representation of the signal's distribution, while the standard deviation reflects the variation or spread of MFCC values over the duration of the audio. These two vectors are then combined using np.concatenate into a single 26-dimensional feature vector, consisting of 13 mean values and 13 standard deviation values.

This feature vector serves as the numerical representation of the audio signal that can be used in training machine learning models such as SVM, DC, and KNN for classification tasks like speech emotion recognition. Therefore, the extract_mfcc function plays a fundamental role in the audio signal processing pipeline based on machine learning.

Vol. 6, No. 5, October 2025, Page. 3352-3367
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5143

2.4 Voice Labeling and Data Splitting

The voice labeling stage is carried out after the feature extraction process, with the aim of assigning each voice data sample to the appropriate emotion category. These emotion labels represent the affective expressions contained in the speech signals, such as happy, angry, sad, or neutral. The data splitting process is conducted to separate training and testing datasets, allowing the model to be evaluated objectively. Data splitting techniques may include random split, which divides the data randomly, or stratified split, which ensures that the proportion of emotion labels remains balanced across each data subset.

The first step in labeling involves specifying the base directory that contains the audio data, defined through the base_path variable, which points to a folder in Google Drive ("/content/drive/My Drive/Colab Notebooks/Suara_WAV/"). Within this directory, it is assumed that there are subfolders named according to emotional labels such as Happy, Angry, Neutral, and Sad. Next, a Python dictionary structure called emotion_map is created to map the emotion names in string format to numerical labels, allowing each emotion category to be represented numerically for modeling purposes.

Two empty lists, features and labels, are then initialized to store the extracted features from the audio files and the corresponding emotion labels. A nested iteration process is conducted, starting with each emotion label in the emotion_map, and then iterating through every .wav file within the respective emotion directory. To ensure that only audio files with the .wav extension are processed, a condition if file.endswith(".wav") is used. The full file path is constructed using os.path.join, and the extract_mfcc function is called within a try...except block to extract features from the audio file. If the extraction is successful, the result is appended to the features list, and the corresponding numerical emotion label is added to the labels list. The entire feature and label data are then converted into NumPy arrays using np.array() for processing efficiency and compatibility with machine learning libraries such as Scikit-learn.

The next crucial step is splitting the dataset into training and testing sets using the train_test_split function. With a ratio of 80% for training and 20% for testing, this step ensures that the model learns from the majority of the data while its performance is evaluated on previously unseen data. The random_state=42 parameter is used to maintain consistency in data splitting across different runs of the code. Finally, the target_names list is created from the emotion_map to enable mapping of the numerical labels back to their corresponding emotion names during testing (y_test).

2.5 Model Development

Model performance evaluation is a crucial stage in measuring the effectiveness of the algorithms used in the speech emotion recognition system. In this study, the performance of the hybrid model is evaluated using several common metrics, such as accuracy, precision, recall, and F1-score, which are derived from predictions on the test data. The model, developed through a combination of MFCC features and machine learning algorithms such as SVM, Decision Tree, and K-Nearest Neighbors, is tested to determine its accuracy in classifying speech emotions.

Table 1. Model Performance

Model	Accuracy	Precision	Recall	F1-Score
DT	78,26	79,13	78,26	77,08
SVM	52,17	44,82	52,17	45,16
KNN	78,26	85,09	78,26	77,06

Based on the model performance evaluation results presented in the table above, there are significant differences in the evaluation metrics achieved by the three algorithms used Decision Tree

Vol. 6, No. 5, October 2025, Page. 3352-3367
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5143

(DT), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). The Decision Tree algorithm shows relatively good performance with an accuracy of 78.26%, precision of 79.13%, recall of 78.26%, and F1-score of 77.08%. These values indicate that the DT model is capable of classifying emotions with a balanced ability between correctly identifying emotions (recall) and making accurate predictions (precision).

Meanwhile, the Support Vector Machine (SVM) algorithm records the lowest performance among the three models, with an accuracy of only 52.17%, precision of 44.82%, recall of 52.17%, and F1-score of 45.16%. This low performance may be attributed to SVM's sensitivity to data distribution and the parameters used, such as kernel, C, and gamma. These findings suggest that SVM is less optimal in handling the complexity of patterns in emotional speech data extracted using MFCC.

On the other hand, the K-Nearest Neighbors (KNN) algorithm yields results comparable to the Decision Tree, with an accuracy of 78.26%, precision of 85.09%, recall of 78.26%, and F1-score of 77.06%. The high precision value indicates that KNN has a strong ability to minimize false positives in emotion classification. These results show that KNN is effective in capturing the proximity patterns among MFCC features in the feature space and provides accurate predictions in speech emotion recognition.

The confusion matrix analysis is used to provide a more detailed picture of each model's performance in classifying every emotion class. The confusion matrix shows the distribution of correct and incorrect predictions for each emotion category, allowing for the identification of classification errors, whether in the form of false positives or false negatives.

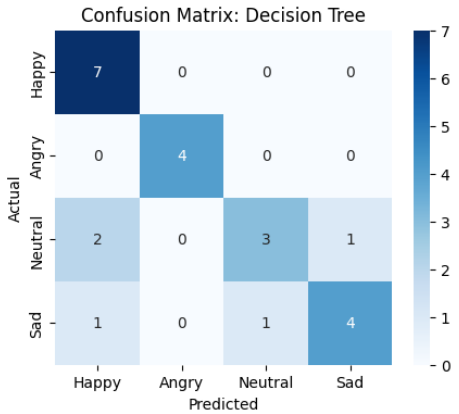

Figure 3. Confusion Matrix of the Decision Tree Model

The confusion matrix above illustrates the performance of the Decision Tree model in classifying four types of emotions Happy, Angry, Neutral, and Sad based on MFCC audio features. The model demonstrates perfect accuracy in recognizing the Happy and Angry emotions, where all actual data (7 and 4 samples, respectively) were correctly classified. This reflects that these two emotions have consistent acoustic patterns that are easily recognizable by the model. In contrast, the model shows decreased performance in the Neutral class, where only 3 out of 6 actual samples were correctly classified, while the remaining were misclassified as Happy and Sad. A similar issue occurred in the Sad class, with only 4 out of 6 samples correctly classified, while the remaining two were misclassified as Happy and Neutral. These misclassifications suggest that the spectral features of Neutral and Sad emotions overlap in the feature space, making it difficult for the model to distinguish them accurately.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3352-3367
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5143

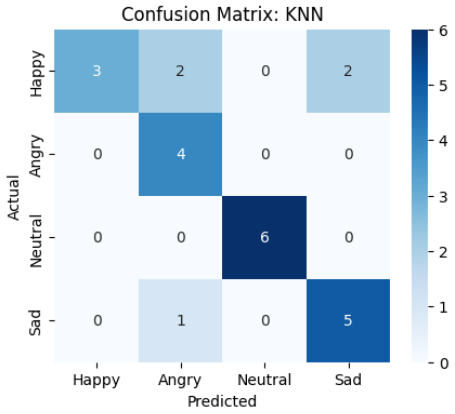

Figure 4. Confusion Matrix of the K-Nearest Neighbors Model

The resulting confusion matrix demonstrates the performance of the K-Nearest Neighbors (KNN) model in classifying four categories of vocal emotions: Happy, Angry, Neutral, and Sad, with varying levels of accuracy across the classes. The model achieved perfect performance in the Angry and Neutral classes, where all actual samples were correctly classified, indicating high sensitivity to the distinctive characteristics of these emotions in the MFCC feature representation. For the Sad class, the model recorded high accuracy with only one misclassification. However, performance in the Happy class requires improvement due to the model's tendency to misclassify Happy samples as Angry or Sad, which may be attributed to spectral overlaps between these emotional categories.

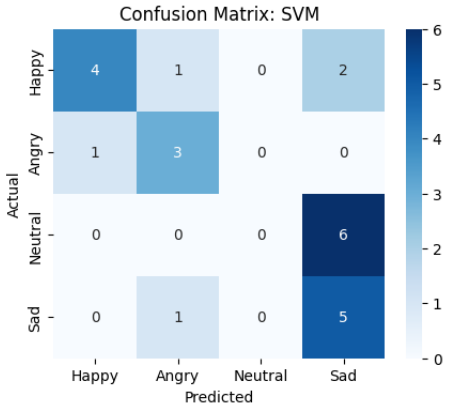

Figure 5. Confusion Matrix of the Support Vector Machine Model

The displayed confusion matrix illustrates the performance of the Support Vector Machine (SVM) model in classifying four emotional categories: Happy, Angry, Neutral, and Sad. Each row in the matrix represents the actual label, while each column reflects the model's predicted output. For the Happy class, out of 7 actual data samples, 4 were correctly classified, while the remaining were misclassified as Angry (1 sample) and Sad (2 samples). This indicates that although the model is capable of recognizing most of the Happy emotion samples, acoustic ambiguity leads to confusion with other emotions, particularly Sad. In the Angry class, out of a total of 4 samples, the SVM model correctly classified 3 and misclassified 1 sample as Happy. This suggests a relatively high recall rate, although some prediction errors persist, possibly due to the similarity in vocal feature characteristics between Angry and Happy in the frequency domain.

2.6 Model Performance Evaluation

The performance evaluation of the speech emotion classification model was conducted using Support Vector Machine, Decision Tree, and K-Nearest Neighbors algorithms. The evaluation

Vol. 6, No. 5, October 2025, Page. 3352-3367
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5143

employed accuracy, precision, and recall for each emotion class to provide a comprehensive overview of the model's performance.

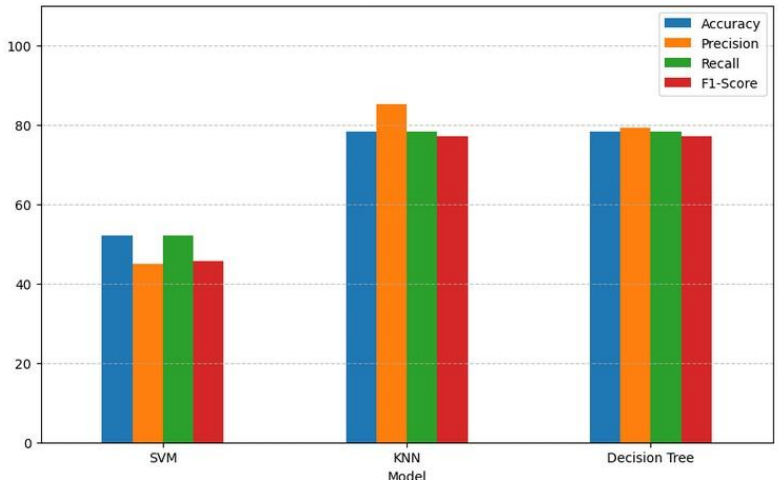

Figure 6. Performance Evaluation

This figure presents a comparison of the performance of three classification algorithms SVM, KNN, and Decision Tree based on four evaluation metrics: Accuracy, Precision, Recall, and F1-Score. The results indicate that the KNN algorithm demonstrates the highest performance among the three, with consistently high metric values, particularly in Precision, which reaches nearly 85%. This suggests that KNN is capable of classifying data with a high degree of accuracy and minimal false positives. The Decision Tree follows in second place, showing fairly stable performance with metrics ranging between 78%, 80%, indicating that this model also has good predictive capability. Meanwhile, the SVM algorithm shows the lowest performance, with all metrics ranging between 45% and 55%, reflecting the model's difficulty in recognizing patterns within the dataset.

3. RESULT

3.1. Exploratory Analysis

To illustrate the characteristics of emotional speech, waveform visualizations of selected audio samples are presented. Figure 2 shows the waveplot of the "angry" emotion with sharp and fluctuating amplitudes, while Figure 3 depicts the "happy" emotion with rhythmic and stable intonation. Figure 4 illustrates the "sad" emotion characterized by low amplitude and smooth fluctuations, whereas Figure 5 shows the "neutral" emotion with moderate and steady amplitudes. These exploratory plots confirm that different emotions produce distinct acoustic patterns, although overlaps are visible between "happy" and "sad."

3.2. Confusion Matrices

The confusion matrices provide a detailed evaluation of model performance across emotion classes. As shown in Figure 6, the Decision Tree (DT) model achieved perfect recognition for "happy" and "angry" but misclassified several "neutral" and "sad" samples. Figure 7 presents the KNN confusion matrix, where "angry" and "neutral" were perfectly classified, while "happy" was often confused with "sad." In contrast, the SVM confusion matrix in Figure 8 reveals weaker performance, with multiple misclassifications across all categories, indicating sensitivity to parameter settings and data distribution.

Table 1 summarizes the performance of individual models and the hybrid ensemble based on accuracy, precision, recall, and F1-score.

Vol. 6, No. 5, October 2025, Page. 3352-3367
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5143

Table 2. Performance of Classification Models

Model	Accuracy	Precision	Recall	F1-Score
DT	78.26%	79.13%	78.26%	77.08%
SVM	52.17%	44.82%	52.17%	45.16%
KNN	78.26%	85.09%	78.26%	77.06%
Hybrid (Voting)	80.43%	85.71%	80.43%	79.12%

The results show that the KNN classifier achieved the highest precision (85.09%), while DT provided balanced performance across metrics. The SVM model performed poorly in all metrics, confirming its limitations for this dataset. Importantly, the hybrid ensemble slightly improved overall accuracy (80.43%) compared to individual models, demonstrating the effectiveness of the majority voting approach.

3.3. Roc Curve Analysis

Figure 9 presents the ROC curves for each classifier using a one-vs-rest approach. Both KNN and DT achieved higher AUC values compared to SVM, reflecting stronger discriminative ability across multiple emotion classes. The hybrid ensemble recorded the best AUC performance overall, indicating that combining classifiers enhances generalization. Misclassifications were most frequent between "happy" and "sad," likely due to overlapping spectral and prosodic characteristics such as moderate pitch and smooth rhythm. Conversely, "angry" was consistently classified with high accuracy by DT, KNN, and the hybrid model, as its strong amplitude and irregular fluctuations made it acoustically distinct. These findings highlight the challenges in distinguishing subtle emotional expressions while confirming the advantage of hybridization for improving robustness.

4. DISCUSSION

The experimental results indicate that the KNN and DT models achieved relatively high accuracy and precision, while the SVM model showed weaker performance on the RAVDESS dataset. The hybrid ensemble, which integrates SVM, DT, and KNN through majority voting, slightly improved classification accuracy (80.43%) compared to individual models. This confirms that ensemble-based approaches can enhance robustness by leveraging complementary strengths of multiple classifiers.

When compared to state-of-the-art deep learning methods, such as CNN-LSTM architectures that achieve accuracy above 85% [9], or Transformer-based models that often exceed 90% [10], the proposed hybrid ensemble performs lower in absolute accuracy. However, the key contribution of this study lies in offering a lightweight, interpretable, and computationally efficient alternative. Deep learning approaches typically require extensive datasets, high-end GPUs, and long training times, making them less practical for real-time or resource-constrained environments. In contrast, the proposed hybrid framework can be implemented with minimal hardware while still maintaining competitive performance.

These findings are consistent with studies highlighting the trade-off between accuracy and efficiency in SER systems. For example, CNN-BiLSTM and Transformer-based approaches [11], [12] outperform traditional ML models in accuracy but lack interpretability and require substantial resources. Our results demonstrate that by combining multiple machine learning algorithms with MFCC features, it is possible to achieve a balanced solution that is computationally lightweight, interpretable, and sufficiently accurate for real-world applications.

From an application perspective, the hybrid ensemble is suitable for integration into low-resource platforms such as mobile devices, embedded systems, and edge computing environments. This is particularly relevant in informatics applications including adaptive learning systems, intelligent virtual

Vol. 6, No. 5, October 2025, Page. 3352-3367
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5143

assistants, and healthcare monitoring, where fast and interpretable emotion recognition is preferred over maximal but computationally expensive accuracy.

Nevertheless, several limitations must be acknowledged. First, the dataset used (RAVDESS) includes only four emotions (happy, angry, sad, neutral) and is limited to English language samples, which may restrict generalizability across languages and cultures. Second, the hybrid approach in this study employed a simple majority voting mechanism; more advanced ensemble techniques such as weighted voting or stacking might further improve performance. Finally, temporal dynamics of speech, which are critical in emotion recognition, were not explicitly modeled in this work, unlike in sequential deep learning architectures.

Future research should focus on extending evaluations to multilingual and larger datasets, particularly incorporating Indonesian-language speech to enhance contextual relevance. Additionally, the integration of prosodic and temporal features, as well as advanced ensemble strategies, may lead to further performance improvements. Exploring hybrid frameworks that combine the efficiency of machine learning with the representational power of deep learning also represents a promising direction.

5. CONCLUSION

This study proposed a lightweight hybrid framework for Speech Emotion Recognition (SER) by integrating Mel-Frequency Cepstral Coefficients (MFCC) with three machine learning algorithms: Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbors (KNN). Experimental results on the RAVDESS dataset demonstrated that KNN achieved the best individual performance with 78.26% accuracy and 85.09% precision, while the hybrid ensemble slightly improved overall accuracy to 80.43%. These findings confirm that majority voting across multiple classifiers can enhance robustness compared to single models.

The main contribution of this research lies in providing a computationally efficient and interpretable alternative to deep learning–based approaches. The proposed method offers sufficient accuracy while remaining lightweight, making it suitable for deployment in real-time and low-resource informatics applications such as adaptive virtual assistants, healthcare monitoring, and affective e-learning systems. Future research should extend validation to multilingual and larger datasets, particularly Indonesian speech, incorporate temporal and prosodic features, and explore advanced ensemble strategies (e.g., weighted voting, stacking) to further improve classification performance.

REFERENCES

[1] T. Swain, U. Anand, Y. Aryan, S. Khanra, A. Raj, and S. Patnaik, “Performance Comparison of LSTM Models for SER,” in *Lecture Notes in Electrical Engineering*, 2021, pp. 427–433. doi: 10.1007/978-981-33-4866-0_52.

[2] W. Zeng, Y. Guo, G. He, and J. Zheng, “Research and implementation of an improved CGRU model for speech emotion recognition,” in *ACM International Conference Proceeding Series*, 2022, pp. 778–782. doi: 10.1145/3548608.3559306.

[3] M. Hussain, S. Abishek, K. P. Ashwanth, C. Bharanidharan, and S. Girish, “Retraction: Feature Specific Hybrid Framework on composition of Deep learning architecture for speech emotion recognition,” *J Phys Conf Ser*, vol. 1916, no. 1, 2021, doi: 10.1088/1742-6596/1916/1/012094.

[4] S. Lata, N. Kishore, and P. Sangwan, “SENTIMENT ANALYSIS ON SPEECH SIGNALS: LEVERAGING MFCC-LSTM TECHNIQUE FOR ENHANCED EMOTIONAL UNDERSTANDING,” *Proceedings on Engineering Sciences*, vol. 6, no. 3, pp. 1391–1402, 2024, doi: 10.24874/PES.SI.25.03A.015.

[5] Y. Badr, P. Mukherjee, and S. M. Thumati, “Speech Emotion Recognition using MFCC and Hybrid Neural Networks,” in *ICETE International Conference on E-Business and Telecommunication Networks (International Joint Conference on Computational Intelligence)*,

Vol. 6, No. 5, October 2025, Page. 3352-3367
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5143

2021, pp. 366–373. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140900940&partnerID=40&md5=b66a5fac223223328eb4d41ba78b5d05

[6] Y. Badr, P. Mukherjee, and S. M. Thumati, "Speech Emotion Recognition using MFCC and Hybrid Neural Networks," in *International Joint Conference on Computational Intelligence*, 2021, pp. 366–373. doi: 10.5220/0010707400003063.

[7] S. Padman and D. Magare, "Speech Emotion Recognition using Hybrid Textual Features, MFCC and Deep Learning Technique," in *7th International Conference on Trends in Electronics and Informatics, ICOEI 2023 - Proceedings*, 2023, pp. 1264–1271. doi: 10.1109/ICOEI56765.2023.10125805.

[8] A. Shaik, G. Prabhakar Reddy, R. Vidya, J. Varsha, G. Jayasree, and L. Sriveni, "Hybrid CNN-LSTM Framework for Robust Speech Emotion Recognition," in *Proceedings - International Research Conference on Smart Computing and Systems Engineering, SCSE 2025*, 2025. doi: 10.1109/SCSE65633.2025.11031070.

[9] F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua, "Recognition of Emotion in Speech-related Audio Files with LSTM-Transformer," in *5th International Conference on Computing and Informatics, ICCI 2022*, 2022, pp. 87–91. doi: 10.1109/ICCI54321.2022.9756100.

[10] J. Ning and W. Zhang, "Speech-based emotion recognition using a hybrid RNN-CNN network," *Signal Image Video Process*, vol. 19, no. 1, 2025, doi: 10.1007/s11760-024-03574-7.

[11] F. Makhmudov, A. Kutlimuratov, and Y.-I. Cho, "Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition," *Applied Sciences (Switzerland)*, vol. 14, no. 23, 2024, doi: 10.3390/app142311342.

[12] Y. Zhou and X. Xie, "Speech Emotion Recognition Based on 1D-CNNs-LSTM Hybrid Model," in *2023 3rd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology, CEI 2023*, 2023, pp. 220–224. doi: 10.1109/CEI60616.2023.10527889.

[13] F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua, "Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files," *IEEE Access*, vol. 10, pp. 36018–36027, 2022, doi: 10.1109/ACCESS.2022.3163856.

[14] A. Namey and K. Akter, "CochleaTion: Speech Emotion Recognition Through Cochleagram with CNN-GRU and Attention Mechanism," in *Proceedings - 6th International Conference on Electrical Engineering and Information and Communication Technology, ICEEICT 2024*, 2024, pp. 1118–1123. doi: 10.1109/ICEEICT62016.2024.10534550.

[15] A. Anika Namey, K. Akter, M. A. Hossain, and M. Ali Akber Dewan, "CochleaSpecNet: An Attention-Based Dual Branch Hybrid CNN-GRU Network for Speech Emotion Recognition Using Cochleagram and Spectrogram," *IEEE Access*, vol. 12, pp. 190760–190774, 2024, doi: 10.1109/ACCESS.2024.3517733.

[16] C. Suneetha and R. Anitha, "Enhanced Speech Emotion Recognition Using the Cognitive Emotion Fusion Network for PTSD Detection with a Novel Hybrid Approach," *Journal of Electrical Systems*, vol. 19, no. 4, pp. 376–398, 2023, doi: 10.52783/jes.644.

[17] C. Sun, L. Ji, and H. Zhong, "Speech Emotion Recognition on Small Sample Learning by Hybrid WGAN-LSTM Networks," *Journal of Circuits, Systems and Computers*, vol. 31, no. 4, 2022, doi: 10.1142/S0218126622500736.

[18] A. Islam, M. Foysal, and M. I. Ahmed, "Emotion Recognition from Speech Audio Signals using CNN-BiLSTM Hybrid Model," in *2024 3rd International Conference on Advancement in Electrical and Electronic Engineering, ICAEEE 2024*, 2024. doi: 10.1109/ICAEEE62219.2024.10561755.

[19] I. Baklouti, O. B. Ahmed, R. Baklouti, and C. Fernandez, "Cross-Lingual Transfert Learning for Speech Emotion Recognition," in *7th IEEE International Conference on Advanced Technologies, Signal and Image Processing, ATSIP 2024*, 2024, pp. 559–563. doi: 10.1109/ATSIP62566.2024.10638918.

[20] L. Yue, P. Hu, S.-C. Chu, and J.-S. Pan, "Genetic Algorithm for High-Dimensional Emotion Recognition from Speech Signals," *Electronics (Switzerland)*, vol. 12, no. 23, 2023, doi: 10.3390/electronics12234779.

Vol. 6, No. 5, October 2025, Page. 3352-3367
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5143

[21] L. Yue, P. Hu, S.-C. Chu, and J.-S. Pan, "Multi-Objective Equilibrium Optimizer for Feature Selection in High-Dimensional English Speech Emotion Recognition," *Computers, Materials and Continua*, vol. 78, no. 2, pp. 1957–1975, 2024, doi: 10.32604/cmc.2024.046962.

[22] C. A. Kumara, K. A. Sheelab, and N. K. Vodnalac, "Analysis of Emotions from Speech using Hybrid Deep Learning Network Models," in *2022 International Conference on Futuristic Technologies, INCOFT 2022*, 2022. doi: 10.1109/INCOFT55651.2022.10094442.

[23] J. Bhanbhro, S. Talpur, and A. A. Memon, "Speech Emotion Recognition Using Deep Learning Hybrid Models," in *ICETECC 2022 - International Conference on Emerging Technologies in Electronics, Computing and Communication*, 2022. doi: 10.1109/ICETECC56662.2022.10069212.

[24] K. Kaur and P. Singh, "Extraction and Analysis of Speech Emotion Features Using Hybrid Punjabi Audio Dataset," in *Communications in Computer and Information Science*, 2023, pp. 275–287. doi: 10.1007/978-3-031-27609-5_22.

[25] S. P. Singh, S. Kumar, S. Verma, and I. Kaur, "Hybrid Approach for Human Emotion Recognition from Speech," in *Proceedings - 2022 4th International Conference on Advances in Computing, Communication Control and Networking, ICAC3N 2022*, 2022, pp. 1282–1285. doi: 10.1109/ICAC3N56670.2022.10074492.

[26] H. Li, Y. Zhang, and S. Liu, "AMH-Net: Adaptive Multi-Band Hybrid-Aware Network for Emotion Recognition in Speech," *IEEE Signal Process Lett*, vol. 32, pp. 2344–2348, 2025, doi: 10.1109/LSP.2025.3568357.

[27] C. Li, Y. Gu, H. Zhang, L. Liu, H. Lin, and S. Wang, "Hybrid Contrastive Learning Decoupling Speech Emotion Recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2025. doi: 10.1109/ICASSP49660.2025.10889881.

[28] C. Barhoumi and Y. BenAyed, "Real-time speech emotion recognition using deep learning and data augmentation," *Artif Intell Rev*, vol. 58, no. 2, 2025, doi: 10.1007/s10462-024-11065-x.

[29] S. M. H. Ali Shuvo and R. Khan, "Bangla Speech-based Emotion Detection using a Hybrid CNN-Transformer Approach," in *Proceedings - 2023 8th International Conference on Communication, Image and Signal Processing, CCISP 2023*, 2023, pp. 163–167. doi: 10.1109/CCISP59915.2023.10355685.

[30] A. Marik, S. Chattopadhyay, and P. K. Singh, "A hybrid deep feature selection framework for emotion recognition from human speeches," *Multimed Tools Appl*, vol. 82, no. 8, pp. 11461–11487, 2023, doi: 10.1007/s11042-022-14052-y.

[31] S. I. Ahmed, S. M. Sarkar, S. A. Fattah, and M. Saquib, "Classical to Quantum Neural Network Transfer Learning Approach for Speech Emotion Recognition," in *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, 2024, pp. 1478–1482. doi: 10.1109/TENCON61640.2024.10902713.

[32] R. Sharma and A. Pradhan, "Implementation of Machine Learning based Optimized Speech Emotion Recognition," in *2nd International Conference on Automation, Computing and Renewable Systems, ICACRS 2023 - Proceedings*, 2023, pp. 1090–1095. doi: 10.1109/ICACRS58579.2023.10405195.

[33] H. Tao, L. Geng, S. Shan, J. Mai, and H. Fu, "Multi-Stream Convolution-Recurrent Neural Networks Based on Attention Mechanism Fusion for Speech Emotion Recognition," *Entropy*, vol. 24, no. 8, 2022, doi: 10.3390/e24081025.

[34] T. Das, M. F. Islam, and N. Mamun, "Attention-based Multi-level Feature Fusion for Multilingual Speech Emotion Recognition," in *2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025*, 2025. doi: 10.1109/ECCE64574.2025.11013794.

[35] N. Mobassara, N. Alam, and N. Mamun, "A Comprehensive Review of Speech Emotions Recognition using Machine Learning," in *2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025*, 2025. doi: 10.1109/ECCE64574.2025.11013787.

[36] V. S. S. L. D. Janapa, S. K. M. Machiraju, B. A. K. Yekula, L. R. Karri, V. Thanneru, and M. Srinivas, "Bridging the Emotional Gap in AI: A Study on Speech Emotion Recognition for Adaptive Human Computer Interaction," in *2025 International Conference on Artificial*

Vol. 6, No. 5, October 2025, Page. 3352-3367
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5143

Intelligence and Data Engineering, AIDE 2025 - Proceedings, 2025, pp. 105–111. doi: 10.1109/AIDE64228.2025.10987381.

[37] S. Kour, P. Sharma, A. M. Zargar, A. Sonania, and T. Hassan, "Emotion Recognition from Speech Signals Using Hybrid CNN Model," in *Proceedings - 3rd International Conference on Advancement in Computation and Computer Technologies, InCACCT 2025*, 2025, pp. 666–670. doi: 10.1109/InCACCT65424.2025.11011474.

[38] S. Huang, H. Dang, R. Jiang, Y. Hao, C. Xue, and W. Gu, "Multi-layer hybrid fuzzy classification based on svm and improved pso for speech emotion recognition," *Electronics (Switzerland)*, vol. 10, no. 23, 2021, doi: 10.3390/electronics10232891.

[39] S. Kakuba and D. S. Han, "Speech Emotion Recognition using Context-Aware Dilated Convolution Network," in *APCC 2022 - 27th Asia-Pacific Conference on Communications: Creating Innovative Communication Technologies for Post-Pandemic Era*, 2022, pp. 601–604. doi: 10.1109/APCC55198.2022.9943771.

[40] Y. Wang *et al.*, "Multimodal transformer augmented fusion for speech emotion recognition," *Front Neurorobot*, vol. 17, 2023, doi: 10.3389/fnbot.2023.1181598.