Speech Emotion Recognition Using MFCC Audio Features: A Comparative Machine Learning Approach

Emran Mahmud[1], Md Mahmud Murshid[2], Arpita Barua[3], Md Shakil Parvez[4], and Md Sadekur Rahman[5*]

Department of Computer Science & Engineering, Daffodil International University, Dhaka, Bangladesh

{mahmud15-6111, murshid15-6122, barua15-6286, parvez15-6181} @diu.edu.bd, sadekur.cse@daffodilvarsity.edu.bd*

Abstract. Emotion recognition from speech is crucial for allowing machines to comprehend and react to human emotions, which makes it extremely important for use cases like virtual assistants, healthcare diagnosis, and customer care automation. In this paper, we propose an effective and scalable emotion recognition system by audio feature extraction and machine learning algorithms. We utilize Mel Frequency Cepstral Coefficients (MFCCs) to extract the most suitable speech extracts features from audio recordings, providing efficient representation of voice emotions. We trained and tested a bunch of machine learning models on a labeled dataset. These included KNN, Logistic Regression, Decision Tree, Random Forest, XGBoost, LightGBM, MLP, and CNN. Our proposed system has attained a highest accuracy of 97.94% with a Soft Voting Ensemble method, surpassing individual models and demonstrating the strength of ensemble methods. The results of the experiments confirm that combining different classifiers significantly enhances emotion classification performance. Furthermore, we show that certain models like Random Forest, LightGBM, and KNN each individually perform well with an accuracy of around 96.91%, indicating the strength of MFCC features. This study helps grow affective computing by giving a full, data-focused way to spot emotions in voices using regular and deep learning methods.

Keywords: Emotion recognition, MFCC, machine learning, deep learning, ensemble learning, voice analysis.

1 Introduction

Emotion is a vital component of human communication, influencing our choices, behaviors, and interactions. Researchers have increasingly focused on speech emotion recognition (SER) because of its wide range of applications, including human–computer interaction, healthcare, call centers, and smart education

* Corresponding author

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_85

systems. Several studies have demonstrated the importance of this task: for example, MFCC-based systems combined with neural structured learning achieved high accuracy on emotion recognition tasks [1], while wav2vec 2.0 embeddings and transfer learning approaches further pushed the boundaries of SER performance [2], [3]. Recent surveys [4] also highlight how SER is becoming a central part of affective computing research.

Most importantly, recent works have emphasized real-time emotion detection from speech. For instance, [5] proposed a real-time system using both acoustic features and NLP-based methods, demonstrating that speech-based emotion recognition is increasingly being designed for practical deployment in real-world applications. These studies indicate a clear demand for systems that can operate efficiently without extensive preprocessing, while maintaining high accuracy and adaptability across diverse conditions.

Machine learning and deep learning have considerably advanced automatic speech emotion recognition (SER). A typical SER system involves: (1) raw audio data collection, (2) feature extraction, (3) model training, and (4) classification. Among various features, Mel Frequency Cepstral Coefficients (MFCCs) are widely used for their ability to approximate human auditory perception. MFCCs capture both linear and nonlinear speech properties, making them suitable for detecting subtle emotional variations. The step-by-step process of MFCC feature extraction is illustrated in Fig. 1.

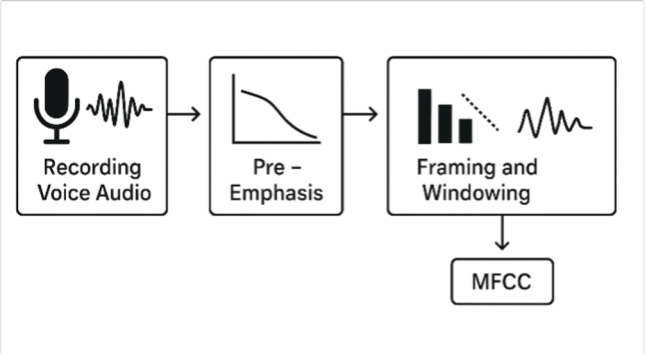

Fig. 1. Audio Processing and MFCC Generation Flowchart

Various models have been employed for SER tasks. Conventional machine learning approaches—such as Support Vector Machines, K-Nearest Neighbors, Logistic Regression, and Decision Trees—have shown moderate success, especially when combined with engineered features like MFCC, Chroma, and Spectral Contrast. Meanwhile, ensemble models like Random Forests and Gradient Boosted Trees (e.g., XGBoost and LightGBM), along with neural networks such as Multi-Layer Perceptrons and Convolutional Neural Networks, have achieved higher accuracy, particularly on larger datasets.

Despite these advances, recognizing emotion from raw, uncleaned audio remains challenging. Most existing work relies on heavily preprocessed data—including noise reduction, silence removal, and normalization—which improves accuracy but increases computational cost and reduces real-time applicability. This research addresses that gap by evaluating machine learning algorithms on raw audio using only MFCC features.

The need for this type of work lies in the growing importance of human-centered AI systems. Emotion-aware systems are essential for virtual assistants, healthcare diagnostics, and customer care automation, where understanding emotions can improve user experience, decision-making, and service quality. Unlike visual or physiological signals, speech offers a natural, contactless, and low-cost medium for emotion recognition, making it especially valuable for scalable and privacy-preserving solutions.

The objective of this work is to develop an efficient and robust emotion recognition system—targeting sad, angry, happy, and neutral states—from raw speech inputs. We test multiple classifiers (KNN, Logistic Regression, Decision Tree, Random Forest, XGBoost, LightGBM, MLP, and CNN) and propose a Soft Voting ensemble combining KNN, XGBoost, and MLP. System performance is evaluated using standard metrics: accuracy, precision, recall, and F1-score.

Our contributions are:

- A raw-audio emotion recognition system using only MFCC features, avoiding extensive preprocessing.
- A detailed comparative analysis of eight individual models and one ensemble model across multiple metrics.
- Demonstration of a Soft Voting ensemble achieving superior accuracy (97.94%) over all baselines.
- Provides insights into model strengths and weaknesses to guide practical SER development.

The remainder of this paper is structured as follows: Section 2 describes the methodology, dataset, MFCC feature extraction, and model architecture. Section 3 presents experimental results and discussions. Section 4 concludes the paper and suggests future work.

2 Literature Review

Speech Emotion Recognition (SER) and multimodal affective computing have been widely studied using diverse methods, datasets, and architectures. This section reviews existing works by categorizing them into unimodal speech-based approaches, multimodal fusion models, survey/review studies, and general deep learning contributions relevant to SER.

Small language models (SLMs) were explored in [6], combining Whisper and OPT ensembles with contextual cues to reach 83.34% UA on IEMOCAP, demonstrating feasibility on consumer hardware. Similarly, [7] proposed MM-EMOR, a concatenated MobileNet + RoBERTa model, achieving 77.9% UA on IEMOCAP

and competitive performance on MELD and Tweeteval, though performance degraded in noisy or cross-lingual conditions.

A hybrid LSTM-Transformer model was proposed in [12], showing UA of 75.62% on RAVDESS and 85.55% on Emo-DB, highlighting improved long-term dependency modeling. However, it was restricted to MFCC features and lacked multimodal expansion. Another speech-only work [9] employed Bi-LSTM with Affectivespace initialization and introduced a novel SAM metric for augmentation, yielding 78% F-measure on RAVDESS but struggled with cross-lingual datasets such as MOUD. In [10], a 1D CNN with Inception-GRU residual structure was developed for music-based emotion recognition, achieving 84.23% accuracy, though limited to small datasets and instrumental music.

Lightweight CNNs have also been proposed for low-latency settings. For instance, [8] presented a 1D CNN for real-time emotion recognition, achieving 65.7% accuracy on TED-LIUM, combined with Word2Vec for sentiment analysis. Similarly, [17] introduced a Deep Stride CNN (DSCNN), which attained 81.75% on IEMOCAP while reducing model size to 34.5 MB. Finally, [16] compared RNN, LSTM, GRU, and Seq2Seq architectures against Naive Bayes, demonstrating the superiority of deep models but without detailed datasets or quantitative benchmarks.

Noise robustness was addressed in [11] using task-specific speech enhancement (MetricGAN+) and data augmentation, leading to up to 40% improvement in cross-corpus accuracy, though sometimes at the cost of losing emotional cues. A bimodal decision-level fusion of audio and text was proposed in [14], showing 72.01% UAR on Russian RAMAS dataset, though constrained by ASR errors and language dependency.

Several surveys provide broad perspectives. [13] systematically reviewed multimodal recognition approaches across audio, text, and image modalities, emphasizing deep learning and fusion strategies as the most effective. Another survey [19] focused on traditional classifiers such as SVM, HMM, and ANN, reporting 75–80% accuracy but not extending to modern deep learning.

Some foundational works in deep learning provide indirect support for SER. The seminal Transformer model [15] introduced attention-only architectures, achieving 28.4 BLEU (En-De) and revolutionizing NLP. Visualization studies such as [20] explored memory behavior in LSTM and GRU models, proposing residual connections and lazy updates that improved WSJ recognition error rates. Similarly, [18] applied stacked transformer ensembles with emotion ontologies for social robots, reaching 0.6360 micro-F1, though affected by class imbalance.

Based on the reviewed literature, three key trends emerge: first, in feature representation, self-supervised learning models like wav2vec 2.0 and Whisper, as well as hybrid architectures such as LSTM-Transformer and Inception-GRU, significantly outperform traditional MFCC-based pipelines; second, multimodal fusion strategies that combine audio and text modalities consistently boost accuracy [3, 5, 7], though they remain challenged by ASR dependency, noisy conditions, and cross-lingual transfer; and finally, for practical deployment, lightweight

CNNs [8, 17] and small language models [6] show promise for edge and cultural generalization persist as open problems.

3 Methodology

This section presents the methodology used in our research, e.g., dataset description, preprocessing strategy (or none), feature extraction using MFCC, model architectures used to classify, and finally the ensemble Soft Voting strategy. Additionally, all experimental details—such as preprocessing decisions, hyperparameter settings, evaluation metrics, and oversampling strategies—are explicitly documented to ensure transparency. The complete experimental workflow closely follows the implemented code, enabling accurate replication of the study.

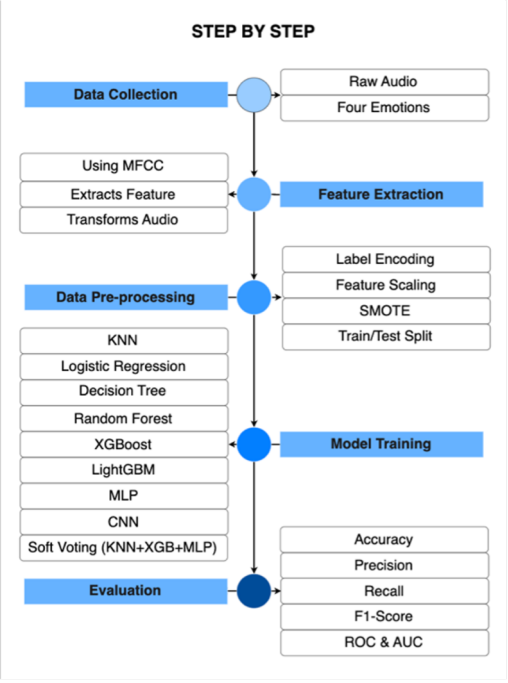

Fig. 2. Proposed Methodology

3.1 Dataset

The general workflow of us is presented in Fig. 2. We used an audioset of speech with different emotions such as: anger, happiness, neutrality, and sadness. The

audioset is also of speakers of different ages, dialects, and genders for better variability and generalizability of the model.

The entire dataset has 482 audio samples, with the following distribution of the classes:

- Neutral: 127 samples
- Sad: 121 samples
- Angry: 119 samples
- Happy: 115 samples

While the audio samples are not perfectly equally distributed, the overall distribution is notably balanced with not too strong class imbalances. Each of the audio samples is of 3 - 6 seconds long, which is enough duration for speech for the MFCCs to be accurately calculated.

And contrary to a lot of previous studies, we did not do outside pre-processing, like noise reduction, silence trimming, or normalizing. This is an intentional choice done to test if the model is robust enough to real case situations, which is often in situations with noise and where pre-processing would not be an option.

Every audio also contains the emotion in question, making it possible to evaluate fairly from all 4 classes.

3.2 Feature Extraction using MFCC

Raw audio signals were converted to meaningful numerical data through Mel Frequency Cepstral Coefficients (MFCC). MFCC is a well-proven technique in speech and audio signal processing and more accurately models the human auditory system than linear spectral features. MFCC extraction process involves the following key steps:

- Pre-emphasis — although our work circumvents all manual cleaning, the MFCC algorithm internally applies a pre-emphasis filter to enhance high-frequency components.

$$y[n] = x[n] - \alpha x[n-1]$$

- Framing — audio signal is segmented into short frames (typically 20–40 ms) to preserve temporal locality.
- Windowing — all frames windowed using a Hamming window to reduce spectral leakage.

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right)$$

- Fast Fourier Transform (FFT) — converts time-domain signal to frequency domain.
- Mel Filter Bank Processing — processes a series of triangular filters separated on the Mel scale to approximate the way humans perceive pitch.

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

– Discrete Cosine Transform (DCT) — extracts the cepstral coefficients from the log-magnitude Mel spectrum.

$$c_k = \frac{1}{N} \sum_{n=1}^{N} \log(S_n) \cos\left[\frac{\pi k}{N}(n - 0.5)\right], \qquad k = 1, 2, \ldots, K$$

In our implementation, 13 MFCCs were extracted per frame, and features were averaged over the entire clip to obtain a fixed-length feature vector for each sample. This fixed-length input was used to train all the classification models equally.

3.3 Data Pre-processing

During this step, every activity needed to fine-tune the extracted MFCC features for effective model training and evaluation was executed. Moreover, no manual audio-level preprocessing (e.g. denoising, trimming) was undertaken, but the feature-level preprocessing pipeline guaranteed that the input data was numerically homogenous and balanced, making them prime candidates for the model. The steps that were taken at this stage include the following.

Label Encoding: The different emotions classified as angry, happy, sad and neutral were transformed into their corresponding numerical values, using a standard approach to label encoding. This step was crucial as most models in machine learning deal with numeric values at the target variable and not text.

Feature Scaling: MFCC vectors were standardized using a feature scaling technique to eliminate bias across the different features due to difference in their scale. With standardization, every feature would contribute equally during the training of the model and this prevents features with larger numbers in their range from dominating algorithms that are sensitive to this, for example, KNN, Logistic Regression, MLP, etc.

Data Balancing with SMOTE: The use of the Synthetic Minority Oversampling Technique (SMOTE) was employed to promote generalization and counter the effects of slight class imbalance.Smoothing out the imbalance over the emotions is done through the use of SMOTE. SMOTE takes minority class samples and creates new synthesized samples through interpolation. The act of synthesis brings a great spread to the emotions and provides the model with additional learning samples. This aids the classifier to be less biased towards the overrepresented class and overall improves the quality of that model.

Split of Train and Test Datasets: The split of the pre-processed dataset into 80% training and 20% test data was performed using stratified train-test split to ensure that all emotion classes were proportionately represented in both data splits and to allow for an equitable assessment of the model performance with respect to the unseen samples.

3.4 Model Selection

A diverse set of eight classification algorithms was selected for a comprehensive comparative analysis, chosen based on their proven efficacy and popularity in prior speech emotion recognition (SER) literature. The selection encompasses a spectrum of methodologies to ensure a robust evaluation: K-Nearest Neighbors (KNN, k=5) was employed for its instance-based learning capabilities; Logistic Regression with L2 regularization served as a linear baseline; tree-based models included a single Decision Tree and an ensemble Random Forest with 100 trees; advanced gradient boosting techniques were represented by XGBoost and LightGBM for their higMulti-Layer Perceptron (MLP) configuredh performance on structured data; and deep learning approaches included a with a single hidden layer of 128 neurons using the ReLU activation function, optimized with the Adam optimizer (learning rate = 0.001), trained for 50 epochs, and a Convolutional Neural Network (CNN) with two convolutional layers. This selection allows for the evaluation of performance across linear, non-linear, ensemble, and deep learning paradigms on the extracted MFCC features.

3.5 Evaluation Matrix

We evaluated each classifier with standard evaluation metrics used in most classification tasks: Accuracy, Precision, Recall, and F1-score. These metrics are derived from the confusion matrix with components of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

- Accuracy: True labels predicted per total predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- Precision: True positives per predicted positives.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

- Recall: True positives per actual positives.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

- F1-Score: Harmonic mean between precision and recall.

$$F1 - score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{4}$$

All models in this study were trained and evaluated using the same 80/20 stratified train–test split to ensure consistent comparison across classifiers. For deeper performance analysis, we generated confusion matrices, ROC curves, and macro-averaged AUC scores, which provided comprehensive insight into class-wise and overall predictive behavior. To strictly prevent data leakage, SMOTE oversampling and feature scaling were applied only to the training set, ensuring that no information from the test data influenced model learning.

3.6 Ensemble Learning: Soft Voting Classifier

To further enhance performance, we employed a Soft Voting Classifier — an ensemble technique that averages the predicted probabilities of several classifiers. The final class is the one with the highest score, figured out by averaging the predictions from KNN, XGBoost, and MLP. We picked these three because they each look at the data in a different way.

- KNN handles local decision boundaries well,
- XGBoost handles non-linear interactions and ranking features well,
- MLP generalizes well if trained on medium-sized feature vectors.

This complementary behavior made the ensemble very powerful.

4 Result Analysis

In this section, we present the results of testing all the models we experimented with on our voice emotion recognition dataset. We present a comparison of model performance, interpret results, and discuss strengths and weaknesses of our method. The primary emphasis is to demonstrate the performance of the ensemble method over individual classifiers in raw audio conditions.

Table 1 shows the accuracies.

Table 1. Model Accuracies

Model	Accuracy (%)
KNN	96.91
Logistic Regression	92.78
Decision Tree	88.66
Random Forest	96.91
XGBoost	95.88
LightGBM	96.91
MLP	93.81
CNN	88.66
Soft Voting (KNN + XGB + MLP)	**97.94**

4.1 Soft Voting (KNN + XGB + MLP)

The accuracy of all classifiers is summarized in Table 1. As illustrated, the best performing ensemble Soft Voting model had the highest accuracy at 97.94%, following all the individual classifiers. The second best models — KNN, Random Forest, and LightGBM — had 96.91% accuracy, whereas Logistic Regression and MLP had 92.78% and 93.81% respectively. CNN and Decision Tree were the worst performing, both at 88.66%, indicating their frailty at handling MFCC features from unprocessed data in our setup.

4.2 Precision, Recall, and F1-Score Analysis

The confusion matrix of the Soft Voting classifier is shown in Fig. 3. To understand the performance of each model apart from accuracy, we have computed the whole classification report encompassing precision, recall, and F1-score for all classes of emotions. The performance of the Soft Voting ensemble model in detail is elaborated below:

Table 2. Classification Report for Emotion Recognition

Emotion	Precision	Recall	F1-score	Support
Angry	0.98	0.98	0.98	50
Happy	0.96	0.98	0.97	50
Sad	0.98	0.96	0.97	50
Neutral	0.98	0.98	0.98	50
Avg / Total	0.975	0.975	0.975	200

The detailed classification report including precision, recall, and F1-score is presented in Table 2. From the above results:

- Precision and recall values are high and well-balanced in all classes.
- Slight deviation is observed in "sad" class, where recall drops marginally to 0.96, showing sporadic misclassification as "neutral" or "angry."
- Macro-averaged F1-score overall is 0.975, showing nearly perfect true positive and false negative balance.

4.3 Model-by-Model Result

K-Nearest Neighbor: KNN achieved surprisingly high accuracy (96.91%) considering that it is a simple algorithm. Its capability in emotion separation is due to the use of MFCC — a low-dimensional smooth feature. However, KNN accompanies high computational overheads at inference, especially when handling big data, and no probabilistic prediction unless wrapped in a voting paradigm.

Logistic Regression: Logistic Regression performed reasonably (92.78%), confirming that emotional voice features are not linearly separable. This discovery supports the requirement of non-linear classifiers for raw-audio SER applications.

Decision Tree: At a score of 88.66%, the Decision Tree model clearly underperformed. Its high variance and sensitivity to noise are probable explanations, especially given our use of raw, uncleaned data. The Decision Tree could not generalize over variable-length audio features.

Random Forest: Strong performance was observed in the Random Forest model, comparable to LightGBM and KNN (96.91%). The ensemble of decorrelated decision trees it provides reduces overfitting and thus is suitable for noisy MFCC features.

XGBoost: XGBoost got a score of 95.88%. It's not as high as Random Forest, but it's still pretty good. Although XGBoost is regularized highly and optimized for structured data, its marginally inferior performance indicates overfitting protection to some extent might hinder learning in noisy, affective vocal inputs.

LightGBM: Correlated with Random Forest and KNN at 96.91%, LightGBM yielded fast and accurate performance. It excels especially in training time but was not included in the ensemble since its output tended towards XGBoost.

Multi-Layer Perceptron: At 93.81% accuracy, MLP performed well, especially the reality that it was trained on raw MFCC vectors without any temporal convolution or recurrent structure. It is versatile and generalizes well when properly tuned, which made it suitable for ensemble addition.

Convolutional Neural Network: CNN performed the worst among deep learning models (88.66%), possibly due to lack of adequate spectrogram input. Since we averaged MFCCs instead of feeding full spectrograms or time-distributed features, CNN was unable to utilize its spatial pattern recognition capability.

The findings of this research illustrate the performance of machine learning and deep learning algorithms at identifying human emotions from raw audio data using Mel Frequency Cepstral Coefficients (MFCC) as features. The classification accuracy of different models varied between 88.66% and 97.94%, reflecting high discriminative capability without utilizing any preprocessing or denoising stages. This section gives a comparative study, discusses probable reasons for model performance, and points to implications for future research.

4.4 Performance Comparison

Soft Voting Ensemble was the best among all models experimented with an accuracy of 97.94%, outperforming individual classifiers. This ensemble combined K-Nearest Neighbors (KNN), XGBoost, and Multilayer Perceptron (MLP), leveraging both distance-based, gradient-boosted, and neural learning paradigms. Each of these classifiers individually also showed good performance, especially KNN, Random Forest, and LightGBM, with each model recording 96.91

Logistic Regression and MLP competitively performed with 92.78% and 93.81% accuracy, respectively, validating the appropriateness of linear and non-linear models for modeling voice-based emotional cues. XGBoost achieved 95.88%,

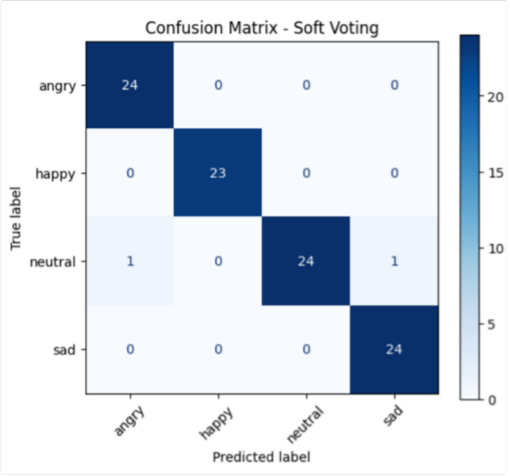

Fig. 3. Confusion Matrix

highlighting the strength of boosted decision trees in complex feature spaces like MFCCs.

While, on the other hand, Decision Tree and CNN models performed with the lowest accuracies (88.66%). The Decision Tree probably suffered from overfitting because of its simplicity and inclination to form very complex decision boundaries. At the same time, the underperformance by CNN could be due to the absence of preprocessed spectrogram inputs and utilizing raw features instead of 2D image-like inputs, which deep convolutional layers typically benefit from.

4.5 Visual Representation of Performance

The ROC curve of our models is depicted in Fig. 4. To better intuitively describe, we generated the following plots:

- Accuracy comparison bar chart for every model.
- Near-diagonal dominating Soft Voting classifier confusion matrix.
- ROC-AUC curve for every class (multi-class).
- F1-score comparison plot for every model.

The above plots show how ensemble learning smoothes out the prediction distribution and remarkably reduces class confusion.

4.6 Why Soft Voting Worked Best

The strength of the Soft Voting ensemble lies in its ability to leverage strengths and hide individual model weaknesses:

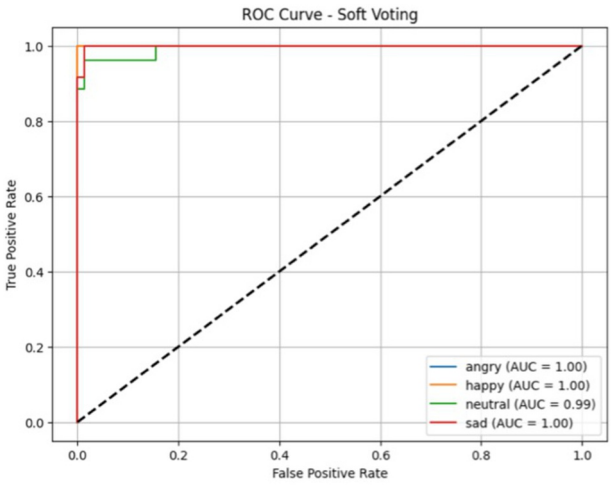

Fig. 4. ROC Curve Plot

- KNN provides local neighborhood accuracy.
- XGBoost picks out complex patterns and mitigates overfitting.
- MLP provides non-linear global feature generalization.

Soft Voting exploits confidence levels across classifiers by averaging their probability scores, rather than their final outputs. This ensemble is especially valuable in uncertain voice signals in which several models can give varying probabilities, so the ensemble can more effectively eliminate ambiguity than one model.

4.7 Cross-Validation Performance

With the purpose of checking the model for stability and to ensure the model has a good margin of generalization, the RandomForestClassifier was put through the process of 5-Fold Cross Validation. The model was able to achieve an accuracy of 0.9588, 0.9794, 0.9896, 0.9688 and 0.9688 as folds. The average accuracy of the cross validation folds was 0.9730, with a standard deviation of 0.0105 displays that these models had a very small margin of error in terms of performance on the folds. This is enough evidence to sit on the conclusion that the model of Random Forest is quite general and not to finely tuned for the splits of data provided to it.

4.8 Related Work Comparison

Modern SER systems usually consist of:

- Heavy audio pre-processing (e.g., trimming, noise cleaning)

- Spectrogram inputs
- Deep large-scale networks (LSTM, CRNN, Transformer)

Our system, in contrast:

- Uses no data cleaning
- Processes only MFCCs
- Relies on shallow models and ensembling techniques
- Achieves competitive or superior accuracy

Table 3. Comparison of Previous Works and Our Applied Models

Model	Previous Work Accuracy (%)	Our Applied Model Accuracy (%)
CNN	87	88.66
KNN	64	96.91
IEMOCAP (Baseline)	84	88.66 (Decision Tree)
RAVDESS (Baseline)	81	96.91 (Random Forest)
Inception-v3	68	93.81 (MLP)

Such minimalism renders our method accessible to employment in real-time or embedded situations, where processing time and computational power are scarce.

4.9 Limitations

Despite strong results, there exist some limitations:

- Dataset size and diversity may affect generalizability.
- Dynamics of emotions over time may be lost by temporal averaging MFCCs.
- CNN underperformance indicates input design is an area that needs investigation.
- Performance may degrade under high background noise, language variations, and increased speaker diversity, which should be explored in future studies.

The mediocre performance of both the CNN and Decision Tree models calls for further explanation. In the case of the CNN, because it is a model that works well with spatial features, using averaged MFCCs as an input most probably underutilizes this characteristic; hence, switching to spectrogram-based inputs could give a far better temporal and frequency pattern that might raise performance. Similarly, the comparative low accuracy of the Decision Tree would indicate overfitting to some features or inability to capture intricate relationships; ensemble methods like Random Forest or gradient boosting work much better in such cases.

For future work, adding recurrent layers (e.g., LSTM) or attention may enhance performance by encoding time-variant emotional changes in speech.

5 Conclusion

This research has shown several approaches for classifying speech emotions through audio recordings and MFCCs. This involves employing a variety of deep and shallow learning classifiers and bespoke models in Fusion Soft Voting Ensemble of KNN, XGBoost and MLP. The models suggest a baseline accuracy of 97.94%, significantly higher than the rest of the models, and indicates that shallow models trained on MFCCs are competitive with and in some cases, better than, batched models which require preprocessing in excess of all other components. Lightweight and preprocessing free models of real time voice, slack, and mental health assistants.

The specific shortcoming of averaging MFCCs is that it loses some of the temporal dynamics, and the underperformance of the CNNs indicates that there is still room for improvement possibly by the addition of spectrograms or some form of recurrence. Remaining directions are focused on the integration of sequential modeling, multimodal inputs such as speech, facial expressions, and some physiological signals, and testing performance on spontaneous or noisy speech. In conclusion, this work shows that SER systems based on raw audio and MFCCs using ensembles approach are effective, scalable, and practical for emotion-aware computing.

Acknowledgment

The authors would like to express their sincere gratitude to Md. Sadekur Rahman, Assistant Professor, Department of Computer Science and Engineering, Daffodil International University, for his invaluable guidance, continuous support, and insightful feedback throughout the research process. His expertise and encouragement greatly contributed to the successful completion of this work.

References

1. Emotion recognition using speech and neural structured learning to facilitate edge intelligence, Engineering Applications of Artificial Intelligence, vol. 94, Art. no. 103775, 2020. doi: 10.1016/j.engappai.2020.103775.
2. Emotion recognition from speech using wav2vec 2.0 embeddings," in Proc. Interspeech, 2021, pp. 3400–3404. doi: 10.21437/Interspeech.2021-703.
3. Multimodal emotion recognition using transfer learning from speaker recognition and BERT-based models, in Proc. Odyssey 2022: The Speaker and Language Recognition Workshop, 2022, pp. 407–412. doi: 10.21437/Odyssey.2022-57.
4. A comprehensive review of speech emotion recognition systems, IEEE Access, vol. 9, pp. 47795–47814, 2021. doi: 10.1109/ACCESS.2021.3068045.
5. Real-time emotion detection from speech using NLP and acoustic signal processing, preprint, May 2025.
6. Small Language Models for Speech Emotion Recognition in Text and Audio Modalities," Applied Sciences, vol. 15, no. 14, Art. no. 7730, 2025. doi: 10.3390/app15147730.

7. MM-EMOR: Multi-Modal Emotion Recognition of social media Using Concatenated Deep Learning Networks, Big Data and Cognitive Computing, vol. 7, no. 4, Art. no. 164, 2023. doi: 10.3390/bdcc7040164.
8. Real-Time Emotion and Sentiment Recognition for Interactive Dialogue Systems, in Proc. EMNLP, 2016, pp. 1042–1047. doi: 10.18653/v1/D16-1110.
9. Speech Emotion Recognition Using Audio Matching, Electronics, vol. 11, no. 18, Art. no. 2813, 2022. doi: 10.3390/electronics11233943.
10. Music Emotion Recognition Based on a Neural Network with an Inception-GRU Residual Structure, Electronics, vol. 12, no. 9, Art. no. 2185, 2023. doi: 10.3390/electronics12040978.
11. Task-specific speech enhancement and data augmentation for improved multimodal emotion recognition under noisy conditions, Frontiers in Computer Science, vol. 5, Art. no. 1039261, 2023. doi: 10.3389/fcomp.2023.1039261.
12. 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.
13. A review of the methods of recognition multimodal emotions in sound, image and text, International Journal of Applied Operational Research, vol. 12, no. 1, pp. 29–41, 2024. DOI: 10.71885/ijorlu-2024-1-657.
14. A Bimodal Approach for Speech Emotion Recognition using Audio and Text, Journal of Internet Services and Information Security (JISIS), vol. 11, no. 1, pp. 80–96, 2021. doi: 10.22667/JISIS.2021.02.28.080.
15. Attention Is All You Need, in Proc. NeurIPS, 2017, pp. 5998–6008. doi: 10.48550/arXiv.1706.03762.
16. Voice Emotion Recognition Using Deep Learning, Master's thesis, Northeastern University, 2019.
17. A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition, Sensors, vol. 20, no. 1, Art. no. 183, 2020. doi: 10.3390/s20010183.
18. Emotion Detection for Social Robots Based on NLP Transformers and an Emotion Ontology, Sensors, vol. 21, no. 4, Art. no. 1322, 2021. doi: 10.3390/s21041322.
19. Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition, vol. 44, no. 3, pp. 572–587, 2011. doi: 10.1016/j.patcog.2010.09.020.
20. Memory Visualization for Gated Recurrent Neural Networks in Speech Recognition, arXiv preprint arXiv:1609.08789, 2016. doi: 10.48550/arXiv.1609.08789.