Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2069−2077
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8878

Speech Emotion Classification Using MFCC Feature Extraction and Bagging-Based Ensemble Learning

Ivan Haristyawan[1,*], Eka Arriyanti[2], Wahyuni[2]

[1] Program Studi Sistem Informasi, STMIK Widya Cipta Dharma, Samarinda, Indonesia
[2] Program Studi Teknik Informatika, STMIK Widya Cipta Dharma, Samarinda, Indonesia
Email: [1,*]ivan_haristyawan@yahoo.com, [2]ekaarry@wicida.ac.id, [3]wahyuni@wicida.ac.id
Corresponding Author Email: ivan_haristyawan@yahoo.com
Submitted: **08/12/2025**; Accepted: **31/12/2025**; Published**: 31/12/2025**

Abstract−Speech emotion classification, also known as Speech Emotion Recognition (SER), has become increasingly important with the growing prevalence of human–machine interaction, particularly in the domains of healthcare, online education, and customer service. This study aims to develop a robust speech emotion classification system by employing Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction and a Decision Tree–based Bagging algorithm for classification. The proposed approach is designed to address the challenges of low classification accuracy, especially under speaker-independent conditions and limited availability of labeled emotional speech data. The research workflow includes speech signal preprocessing, MFCC feature extraction, dataset partitioning through bootstrapping, ensemble model training, and performance evaluation using accuracy, precision, recall, and F1-score metrics. Experimental results on a balanced dataset comprising five emotion classes (anger, disgust, fear, happy, and sad) demonstrate that the proposed model achieves an overall accuracy of 61.04%. While the fear and happy emotions are classified effectively with recall values of 0.75, the anger class exhibits the lowest performance with an F1-score of 0.49. Confusion matrix analysis further reveals substantial acoustic overlap among several emotion categories, particularly the frequent misclassification of sad as disgust or anger. In conclusion, the integration of MFCC features with the Bagging algorithm improves model stability and robustness; however, further optimization of acoustic features and hyperparameters is required to enhance overall classification accuracy.

Keywords: Speech Emotion Classification; MFCC; Bagging Algorithm; Decision Tree; Ensemble Learning

1. INTRODUCTION

Speech Emotion Recognition (SER) has gained increasing attention along with the rapid growth of human–machine interaction, particularly in the fields of healthcare, online education, and customer service [1][2][3]. Emotions conveyed through speech contain paralinguistic information that can be captured using acoustic features such as the Mel Frequency Cepstral Coefficient (MFCC), which has been widely proven effective in representing emotional characteristics in speech signals [2][4][5]. With the advancement of machine learning technologies, the development of MFCC-based SER systems has become one of the main research focuses to enhance classification accuracy and computational efficiency [1][2][4].

One of the primary challenges in SER is the relatively low classification accuracy, especially under speaker-independent scenarios and limited availability of labeled emotional speech data [1][2][4]. Recent studies report that emotion classification accuracy on the RAVDESS dataset reaches only 78.8% when using 14 emotion classes, although it can increase up to 98.5% when the number of classes is reduced [6]. In addition, the subjective variability of emotional expression and data scarcity hinder the generalization capability of learning models [1][2]. Other studies have demonstrated that the combination of MFCC features with ensemble learning techniques can improve classification accuracy up to 97.04% on the RAVDESS dataset; however, performance degradation still occurs under class imbalance and noise conditions [7]. These limitations indicate the need for more robust and adaptive solutions to achieve consistent SER performance across diverse data conditions.

This study proposes the application of the Bagging (Bootstrap Aggregating) algorithm for MFCC-based speech emotion classification. Bagging operates by constructing multiple classification models from different subsets of the training data and then aggregating their predictions to improve stability and accuracy. By integrating Bagging with MFCC features, the proposed system is expected to reduce model variance, enhance robustness against noise, and improve performance on imbalanced datasets. This approach also enables the exploitation of the collective strength of multiple base learners, thereby producing more reliable and stable classification results [5][7].

Previous studies have validated the effectiveness of combining MFCC features with ensemble and deep learning methods for SER. Chowdhury et al. (2025) demonstrated that CNN–BiLSTM-based ensemble models using MFCC features outperformed single models across five benchmark datasets [4]. Aishwarya et al. (2024) reported that ensemble techniques such as Voting Classifier and CatBoost with MFCC features achieved an accuracy of up to 97.04% on the RAVDESS dataset [7]. Patnaik (2022) introduced c-MFCC with deep sequential models, achieving up to 98.5% accuracy for six emotional classes [6]. Furthermore, Pagidirayi and Bhuma (2022) showed that the combination of the Random Subspace Method (bagging) and kNN with MFCC features was effective in classifying eight primary emotions [5]. All of these studies are indexed in Scopus, emphasizing both the relevance and urgency of MFCC-based ensemble approaches in SER research.

The main objective of this study is to design and implement an MFCC-based speech emotion classification system using the Bagging algorithm to improve accuracy, stability, and model robustness against data variability and noise. This study also aims to evaluate the system performance across multiple emotional speech datasets and to

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2069−2077
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8878

compare the results with conventional classification methods. Through this effort, the proposed work is expected to contribute to the development of more reliable and practical SER systems for real-world applications.

2. RESEARCH METHODOLOGY

The research process was conducted through several interrelated stages, starting from the initial processing of speech data to model evaluation.

2.1 Research Stages

The research stages illustrate the workflow of speech emotion classification using MFCC features and the Bagging algorithm. The process begins with speech data preprocessing to remove noise and silence segments. Subsequently, MFCC feature extraction is performed to transform speech signals into numerical representations. The training data are then divided into several subsets using the bootstrapping technique, and each subset is used to independently train a Decision Tree model. The predictions generated by all models are combined through majority voting to obtain the final decision. The final stage consists of ensemble model training and performance evaluation using accuracy and precision metrics. Figure 1 presents the overall research framework employed in this study.

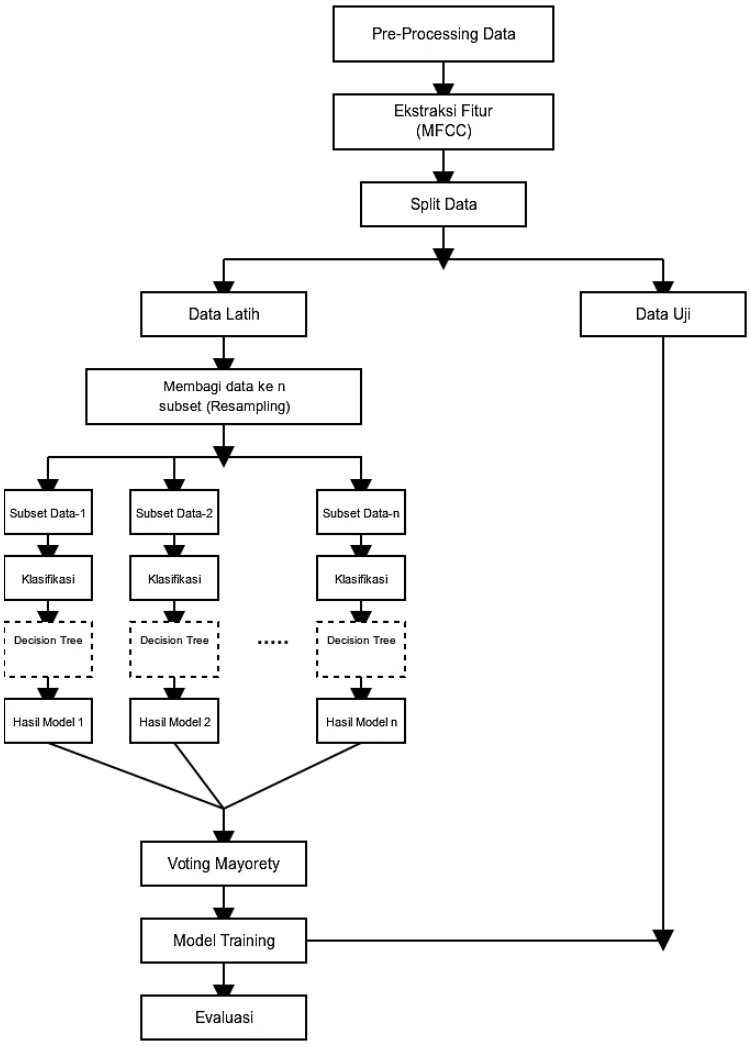

Figure 1. Bagging Scheme with MFCC

Based on Figure 1, the research starts with a data preprocessing stage aimed at preparing and improving the quality of the speech signals prior to further analysis. This stage includes normalization, noise removal, and elimination of silent segments so that the audio data become clean and suitable for processing. This step is critical because the quality of the input data strongly influences both the feature extraction results and the performance of the classification model.

Next, feature extraction is performed using Mel-Frequency Cepstral Coefficients (MFCC). MFCC extracts the main characteristics of speech signals by simulating the human auditory perception mechanism, producing numerical coefficients that represent frequency patterns and distinctive features of the speech signal. The extracted features are then used as inputs for the classification process.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2069−2077
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8878

The extracted data are divided into two parts, namely training data and testing data. The training data are used to build the classification model, while the testing data are used to evaluate model performance. The training data are then processed using resampling techniques to generate multiple subsets, each of which is used to train a different Decision Tree model.

Each Decision Tree model produces its own prediction, and the outputs of all models are combined using the Majority Voting method to obtain a more stable and accurate final decision. The final stage consists of model training and performance evaluation using metrics such as accuracy and precision. Through this workflow, the study produces a reliable speech classification system by integrating effective preprocessing, MFCC feature extraction, and an ensemble-based Decision Tree approach.

2.2 Ensemble Learning

Ensemble learning is a machine learning technique that combines the predictions of multiple models, known as base learners, to improve prediction accuracy and robustness. By aggregating the outputs of several base models, ensemble learning typically achieves higher predictive performance than a single model. This technique can be applied to both classification and regression problems. The ensemble modeling process generally involves two main stages: first, training multiple base learners using the available dataset; second, combining their predictions using a specific aggregation method, such as majority voting, averaging, or weighted averaging, to produce a single final prediction[8][9].

2.2.1 Max Voting

Max-voting, commonly used in classification problems, is one of the simplest ensemble methods for combining predictions from multiple machine learning models. In this approach, each base learner produces a prediction for each sample, and the class that receives the highest number of votes from all models is selected as the final predicted class. This process reflects the principle of majority rule, where the final decision is determined by the dominant consensus among the models. In cases where multiple classes receive the same number of votes, several strategies can be adopted, such as random selection, confidence-based weighting, or prioritizing the model with the highest individual performance. As an analogy, in a five-point Likert-scale survey, if the majority of respondents select level four, then the final decision is level four, which corresponds to identifying the statistical mode of the responses[10][11].

2.2.2 Bagging Algorithm

The Bagging (Bootstrap Aggregating) algorithm was first introduced by Breiman in 1996 and is a fundamental ensemble learning technique. Breiman explained that bagging is particularly effective for unstable learning algorithms, namely algorithms whose predictions change significantly due to small variations in the training data. Examples of such algorithms include Neural Networks and Decision Trees[12]. In addition, the Bootstrap Aggregating method has been proven to be efficient for small-sized datasets[13].

Secara umum, bagging berfungsi untuk mengurangi varians pada algoritma dengan varians tinggi dan In general, bagging functions to reduce the variance of high-variance algorithms and helps mitigate the risk of overfitting[14]. This method can be applied to both classification and regression problems[15].

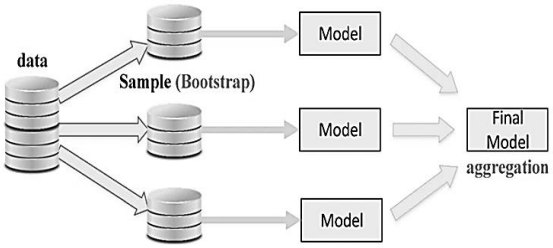

Figure 2. General Diagram of the Bagging Algorithm

Based on Figure 2 (Bagging algorithm pseudo-code), the process begins by generating multiple different versions of the original training dataset (S) using the bootstrap sampling technique, namely random sampling with replacement. Each bootstrap sample (S_t) is used to train a classification model (h_t), resulting in an ensemble of diverse models. Predictions for a new data instance (x) are then determined using the majority voting method, where the final output (h(x)) is the most frequently predicted class among all classifiers..

Meanwhile, in Figure 2 (General Diagram of Bagging Classification), the process illustrates that the training data are divided into multiple subsets using the bootstrapping technique, producing N samples (X_1, X_2, …, X_n). Each subset is used to train an individual classification model, and the predictions from all models are aggregated through voting or output averaging to generate a more accurate final prediction.

2.2.3 Bootstrapping

In modern statistical analysis, bootstrapping has become one of the most widely used approaches for estimating parameter uncertainty without requiring specific distributional assumptions[16][17]. This method belongs to the class

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2069−2077
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8878

of non-parametric resampling techniques that allow researchers to assess the reliability of an estimator by repeatedly sampling from the original dataset[18]. The bootstrapping procedure is performed by generating a large number of new samples from the original dataset through sampling with replacement. As a result, a single observation may appear multiple times in different bootstrap samples or may not appear at all. From this process, n bootstrap samples are generated, each having the same size as the original dataset. The larger the value of n, the closer the bootstrap distribution approaches the ideal population distribution[19].

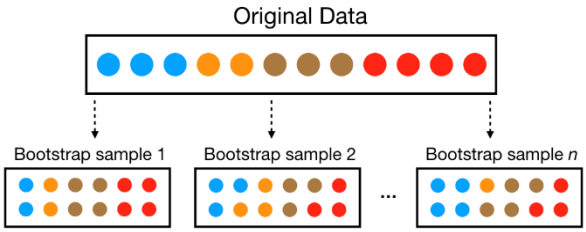

Figure 3. Bootstrap Sampling

As shown in Figure 3, several observations from subset S_1 also appear in subsets S_2 and S_4, illustrating the repetition property caused by sampling with replacement. Suppose there are n bootstrap samples generated from the original dataset, with $\hat{\theta}_i$ denoting the estimated parameter value from the i-th sample (i = 1, 2, 3, …, n). If $\hat{\theta}$ represents the parameter estimate from the original sample, then the standard error (SE) and the mean of the estimates ($\bar{\theta}$) can be calculated using the following equations[20][21]:

$$SE(\hat{\theta}) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta})^2} \quad (1)$$

$$\bar{\theta} = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\hat{\theta}_i \quad (2)$$

The value of $\bar{\theta}$ represents the estimated mean of the overall n bootstrap samples. Conceptually, the core idea of the bootstrap method is to repeatedly resample from the original dataset, where each observation has an equal probability of being selected in each iteration[22]. The results of this resampling process form a collection of bootstrap samples that are used to approximate the true distribution of the parameter of interest.

2.3 Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCC) is an audio feature extraction method designed to mimic the characteristics of human auditory perception[6]. This process transforms a speech signal into a numerical representation through several stages, including spectral analysis, mapping onto the Mel frequency scale, and the application of the Discrete Cosine Transform (DCT)[23]. The final output is a low-dimensional feature vector that captures the dominant characteristics of the spoken signal. The main stages of MFCC extraction include[24][25][26][27]:

a. *Pre-Emphasis*
This stage aims to amplify the high-frequency components of the speech signal

$$y[n] = x[n] - \alpha.x[n-1] \quad (3)$$

b. *Framing*
The signal is divided into short frames with a duration of approximately 20–40 ms in order to maintain quasi-stationary signal characteristics.

c. *Windowing*
To reduce discontinuities at the frame boundaries, a Hamming window function is applied:

$$w[n] = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N-1}\right) \quad (4)$$

d. *Fast Fourier Transform* (FFT)
This step converts the signal from the time domain to the frequency domain using:

$$X[k] = \sum_{n=0}^{N-1} x[n].e^{j2\pi kn/N} \quad (5)$$

e. *Mel Filter Bank*
The FFT output is passed through a bank of filters based on the Mel frequency scale:

$$m(f) = 2595.log_{10}\left(1 + \frac{f}{700}\right) \quad (6)$$

f. Log Energi
The energy of each Mel filter is computed and logarithmically scaled as:

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2069−2077
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8878

$$E_i = log(\sum_{k=1}^{N} |X[k]|^2 . H_i[k]) \quad (7)$$

g. *Discrete Cosine Transform* (DCT)
The final stage aims to decorrelate the filter bank energies and generate the MFCC coefficients:

$$c_n = \sum_{i=1}^{K} log(E_i).cos\left[\frac{\pi n(i-0.5)}{K}\right] \quad (8)$$

2.4 *Decision Trees*

Decision Trees are well-known non-parametric supervised learning algorithms that are highly effective for predictive modeling and classification tasks. This method was first introduced by Leo Breiman and has since become one of the most widely used approaches due to its intuitive interpretation of the decision-making proces[28]. The construction of a Decision Tree is performed through recursive partitioning, which iteratively splits the dataset into smaller subsets based on selected independent attributes. The splitting process continues until a stopping criterion is met, such as when all samples in a node belong to the same class or when the maximum tree depth is reached[29]. The optimal attribute is selected based on a splitting criterion that produces the highest node purity.

For classification tasks, several commonly used splitting measures include Entropy, Information Gain, the Gini Index, and the Gain Ratio[30]. Entropy is employed to quantify the degree of uncertainty or randomness of an attribute and is mathematically defined as:

$$Entropy = \sum_{i=1}^{n} -p_i log_2(p_i) \quad (9)$$

For a binary-class attribute, the entropy value ranges between 0 and 1, whereas for an attribute with n classes, the value can reach up to $\log_2(n)$. If all samples within a node belong to the same class (i.e., the node is homogeneous), the entropy becomes 0, indicating the absence of uncertainty. To determine the optimal attribute for data partitioning, Information Gain is used, which represents the reduction in entropy after a split is performed. It is expressed as:

$$Information\ Gain = Entropy(S_{Before\ Split}) - Entropy(S_{After\ Split}) \quad (10)$$

A higher Information Gain value indicates that the corresponding attribute is more effective in separating the data. This procedure is applied at each node, where the feature with the highest gain is selected to form the root node and subsequent branches. In addition to entropy-based measures, the Gini Index is also widely used to assess node impurity and is computed as:

$$Index\ Gini = D1 - \sum_{i=1}^{n} p_i^2 \quad (11)$$

In this formulation, P_1 denotes the probability of a training sample belonging to the i-th class. A lower Gini Index value indicates higher node purity, and therefore, a more optimal data split.

3. RESULTS AND DISCUSSION

3.1 Data Distribution by Label

Based on Figure 4, the dataset consists of five main emotion classes, namely sad, happy, fear, disgust, and anger. The histogram indicates that the number of samples in each emotion category is relatively balanced, with each class containing nearly 1,000 audio samples.

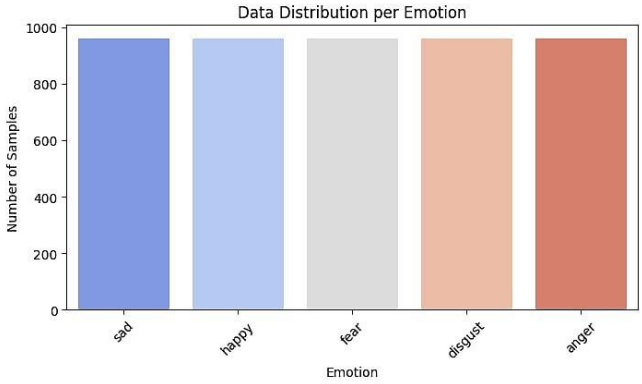

Figure 4. Distribution of Data by Label

The dataset comprises five primary emotion classes (anger, disgust, fear, happy, and sad) with a total of 4,800 samples. The distribution of data across the five classes is 960 samples per emotion, indicating a balanced dataset. This balance is crucial to ensure that the Bagging + Decision Tree model can be trained robustly and to prevent classification bias toward any particular emotion class.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2069−2077
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8878

3.1.1 Audio Data Preprocessing

In the signal plot prior to preprocessing, significant gaps or silent segments are observed at the beginning of the recordings (from 0 to 1 second), along with inconsistent amplitude variations. In contrast, the signal plot after preprocessing shows that these silent segments have been successfully removed using silence removal techniques, and the signal amplitude appears more normalized.

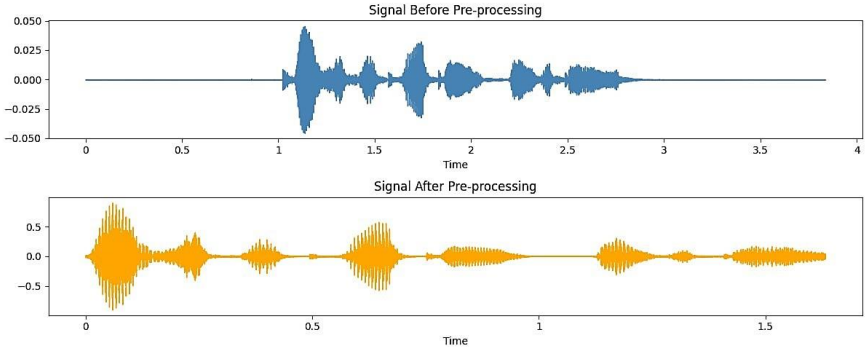

Figure 5. Audio Preprocessing

This stage is consistent with the research methodology, which states that noise and silent segment removal aims to ensure that the speech data are clean and ready for further processing. These results confirm that the preprocessing stage successfully focused the data on speech segments containing relevant emotional information, which ultimately contributes to the accuracy of MFCC feature extraction.

3.2 MFCC Feature Extraction

The feature extraction process transforms the audio signal into a compact numerical representation in the form of MFCC feature vectors, which are highly effective in representing the spectral characteristics of human speech. Each row in the table below represents the average values of the 13 MFCC coefficients extracted from each audio file.

Table 1. MFCC Feature Extraction Vector Values

Filename	Label	0	1	…	12
03-01-03-01-01-01-10_aug2.wav	Sad	-338.2336	77.1707	…	-13.8983
…	…	…	…	…	…
03-01-02-01-01-01-08.wav	Happy	-342.8881	111.8887	…	- 3.2593

The comparison of MFCC vector values between two samples, namely Sad and Happy emotions, reveals significant differences in their acoustic representations.

a. Coefficient 0 (Signal Energy): The first coefficient (C_0) represents the average logarithmic signal energy. The Sad emotion exhibits a C_0 value of −338.2336, whereas the Happy emotion shows a lower value of −342.8881. Although both values are negative, indicating relatively low signal energy, the lower value observed for Happy suggests greater energy variation or slightly more attenuated average energy in the recording context.
b. Coefficient 1 (Spectral Representation): Coefficient 1 and subsequent coefficients describe the spectral shape of the speech signal. A marked difference is observed in C_1, where Sad has a value of 77.1707, while Happy exhibits a substantially higher value of 111.8887. The higher C_1 value for Happy indicates that this emotion has a greater concentration of energy in higher frequency ranges, corresponding to a higher pitch and more expressive vocal characteristics. In contrast, the lower C_1 value for Sad reflects a flatter spectrum or dominance of lower-frequency components.
c. Coefficient 12 (Spectral Detail): Differences are also evident in C_{12}. The Sad emotion shows a larger negative value of −13.8983, whereas Happy has a value closer to zero at −3.2593. Higher-order coefficients such as C_{12} represent fine details of the spectral envelope. These disparities indicate that the spectral patterns of Sad and Happy emotions are fundamentally different and can be effectively separated by machine learning models.

3.3 Model Evaluation Results

The speech emotion classification model was built using a Bagging Classifier with Decision Tree as the base learner. The dataset was divided into training and testing sets with an 80:20 ratio using a stratified split to ensure proportional representation of each emotion class. As a result, 3,840 samples were used for training and 960 samples for testing.

The training process produced an ensemble model consisting of 50 estimators, where each estimator was trained on a bootstrap subset of the training data generated through sampling with replacement. Each subset had the same size as the full training dataset (3,840 samples), although the sample composition differed due to random resampling. This approach allowed each learner to receive unique data variations, thereby reducing overfitting commonly associated with single Decision Trees and improving generalization performance on unseen data.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2069−2077
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8878

3.3.1 *Clasification Report*

The system performance was evaluated using accuracy, precision, recall, and F1-score based on testing results from 960 samples out of a total of 4,800 data points. The Bagging Classifier with Decision Tree as the base learner achieved an overall accuracy as presented in Table 2.

Table 2. MFCC-Based Classification Performance

Emosi	Precision	Recall	F1-Score	Support
anger	0.48	0.51	0.49	192
disgust	0.55	0.52	0.53	192
fear	0.78	0.75	0.76	192
happy	0.62	0.75	0.68	192
sad	0.64	0.53	0.58	192
Akurasi Rata-rata	0.61	0.61	0.61	960

Based on Table 2, the MFCC-based Bagging Classifier exhibits varying performance across emotion categories. The fear class achieved the best performance with a precision of 0.78 and a recall of 0.75, indicating strong recognition capability for fear-related vocal patterns. The happy class also demonstrated relatively high performance with a recall of 0.75 and an F1-score of 0.68, showing that most happy samples were correctly identified. In contrast, the anger and disgust classes yielded lower precision and recall values around 0.5, indicating difficulty in distinguishing these acoustic expressions. The sad class achieved moderate performance with an F1-score of 0.58. Overall, the average accuracy of 61% indicates moderate classification capability and suggests potential improvement through parameter optimization or the incorporation of richer acoustic features.

3.3.2 *Confusion Matrix*

The confusion matrix illustrates the relationship between true labels and predicted labels, providing a detailed view of the model's classification behavior across emotion categories, as shown in Figure 6.

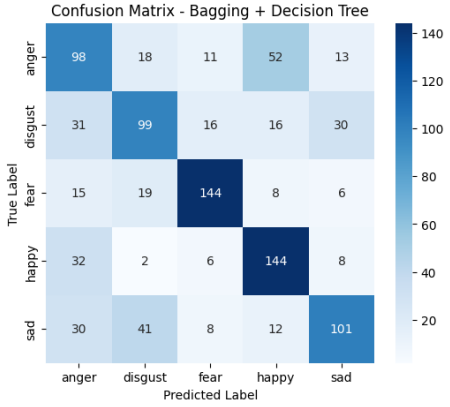

Figure 6. Confusion Matrix

The matrix indicates that the fear and happy classes achieved the highest numbers of correct predictions, with 144 correctly classified samples each. The anger class was also classified reasonably well with 98 correct predictions. In contrast, the disgust and sad classes recorded lower correct prediction counts of 99 and 101 samples, respectively. Misclassification analysis is critical for identifying model weaknesses. A notable number of sad samples were incorrectly predicted as disgust (41 samples) and as anger (30 samples). This phenomenon indicates a strong acoustic similarity between sad and disgust emotions in the training data, which makes class separation challenging. Additionally, a considerable number of anger samples were misclassified as happy (52 samples), suggesting overlapping intensity-related acoustic features between these two emotions. Overall, despite these challenges in separating acoustically similar emotion pairs, most correct predictions lie along the diagonal of the matrix, supporting the reported model accuracy of 61.04%.

The balanced distribution of data across the five emotion classes (anger, disgust, fear, happy, and sad), with 960 samples for each class, provides ideal conditions for model training and evaluation. This balance plays a critical role in preventing class bias and ensures that performance metrics genuinely reflect the model's generalization ability across all emotion categories. With a total of 4,800 audio samples, intra-class variability is sufficiently represented, enabling the model to learn relevant acoustic patterns for each emotion.

3.4 Discussion

The preprocessing stage makes a significant contribution to signal quality prior to feature extraction. The removal of silent segments and amplitude normalization reduces distortions that do not carry emotional information, making the

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2069−2077
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8878

processed signals more representative of emotional characteristics. This step is particularly critical because MFCC features are highly sensitive to spectral variations. Cleaned signals enable the extraction process to capture emotional information more consistently and reliably, directly influencing classification performance.

The MFCC feature extraction results demonstrate clear numerical differences among emotion classes, particularly within the lower-order and higher-order coefficients. Variations in the zeroth and first coefficients reflect differences in energy and spectral distribution between emotions such as happy and sad. The happy emotion tends to exhibit more dominant high-frequency components, while sad shows a lower and more stable spectral profile. Higher-order coefficients, such as C_{12}, further reveal variations in spectral detail, confirming that MFCC remains a relevant primary feature representation for speech emotion recognition.

The application of the Bagging algorithm with Decision Tree as the base learner improves model stability through the bootstrap aggregation mechanism. A total of 50 estimators were trained using subsets generated via sampling with replacement, allowing each decision tree to learn from different data distributions. This approach effectively reduces model variance and minimizes the overfitting risk commonly found in single Decision Tree models. The use of a stratified split further ensures proportional representation of each emotion class in both training and testing datasets, enhancing the validity of performance evaluation.

The evaluation results show that the system achieved an overall accuracy of 61.04%, indicating moderate classification performance. The fear class achieved the best performance with a precision of 0.78 and a recall of 0.75, showing that fear-related acoustic characteristics are relatively easier to distinguish using MFCC features. The happy class also performed well with a recall of 0.75 and an F1-score of 0.68. In contrast, the anger and disgust classes achieved lower precision and recall values around 0.5, indicating substantial overlap in their spectral characteristics. The sad class showed intermediate performance with an F1-score of 0.58. This performance variation suggests that not all emotions can be optimally separated using only static MFCC features.

The confusion matrix analysis provides deeper insight into misclassification patterns. The fear and happy classes show the highest correct prediction rates, consistent with the classification report. Conversely, substantial misclassification occurs in the sad class, which is frequently predicted as disgust and anger. This indicates strong similarity in prosodic and energy-related vocal characteristics among these emotions. Furthermore, anger is often misclassified as happy, likely due to similar high-intensity and dynamic vocal traits. These findings highlight the limitations of MFCC in distinguishing emotions with similar acoustic properties when temporal information is not explicitly modeled.

Compared with previous studies, the performance of the proposed model remains below that of deep learning-based approaches. Several studies report accuracies exceeding 90% by combining MFCC with CNN–BiLSTM architectures[4] and boosting-based ensemble methods as well as deep sequential models[7]. In contrast, the 61.04% accuracy obtained in this study reflects the limitations of spectral feature-based and tree-based ensemble approaches in capturing the temporal dynamics of speech signals. Nevertheless, the proposed model offers advantages in terms of computational efficiency, ease of implementation, and higher interpretability compared to deep learning models, which are typically considered black-box systems.

Overall, the results indicate that the combination of MFCC and Bagging Decision Tree can construct a stable speech emotion classification system, but substantial room for performance improvement remains. Future work may incorporate additional prosodic features such as pitch, formant, and energy contours, or utilize deep learning-based features as front-end extractors. Further improvements may also be achieved through hyperparameter optimization, data augmentation strategies, and the exploration of more advanced ensemble architectures to significantly enhance generalization performance.

4. CONCLUSION

In conclusion, this study successfully implemented a speech emotion classification system for five emotional categories (anger, disgust, fear, happy, and sad) by integrating Mel-Frequency Cepstral Coefficients (MFCC) feature extraction with a Bagging ensemble algorithm based on Decision Tree classifiers. Although a balanced dataset was employed, with 960 samples per emotion, the developed model achieved an overall accuracy of 61.04%. These results indicate that the combination of MFCC and Bagging is effective in constructing a stable classification framework; however, substantial performance variations were observed across different emotion classes. The fear and happy emotions were recognized with the highest effectiveness (recall of 0.75), whereas anger was identified as the most challenging class with an F1-score of 0.49. Therefore, future work is recommended to explore more discriminative acoustic features and to conduct deeper hyperparameter optimization in order to improve the overall classification accuracy and address the misclassification issues observed in specific emotional categories.

REFERENCES

[1] S. Madanian *et al.*, “Speech emotion recognition using machine learning - A systematic review,” *Intell. Syst. Appl.*, vol. 20, p. 200266, 2023, doi: 10.1016/j.iswa.2023.200266.

[2] S. Mishra, P. Warule, and S. Deb, “Speech emotion recognition using MFCC-based entropy feature,” *Signal, Image Video Process.*, vol. 18, pp. 153–161, 2023, doi: 10.1007/s11760-023-02716-7.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2069−2077
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8878

[3] J. H. Chowdhury, S. Ramanna, and K. Kotecha, "Speech emotion recognition with light weight deep neural ensemble model using hand crafted features," *Sci. Rep.*, vol. 15, 2025, doi: 10.1038/s41598-025-95734-z.
[4] J. H. Chowdhury, S. Ramanna, and K. Kotecha, "Speech emotion recognition with light weight deep neural ensemble model using hand crafted features," *Sci. Rep.*, vol. 15, no. 1, pp. 1–14, 2025, doi: 10.1038/s41598-025-95734-z.
[5] A. K. Pagidirayi and A. Bhuma, "Speech Emotion Recognition Using Machine Learning Techniques," *Rev. d'Intelligence Artif.*, vol. 36, no. 2, pp. 271–278, 2022, doi: 10.18280/ria.360211.
[6] S. Patnaik, "Speech emotion recognition by using complex MFCC and deep sequential model," *Multimed. Tools Appl.*, vol. 82, pp. 11897–11922, 2022, doi: 10.1007/s11042-022-13725-y.
[7] N. Aishwarya, K. Kaur, and K. Seemakurthy, "A computationally efficient speech emotion recognition system employing machine learning classifiers and ensemble learning," *Int. J. Speech Technol.*, vol. 27, pp. 239–254, 2024, doi: 10.1007/s10772-024-10095-8.
[8] S. Chen and W. Zheng, "RRMSE-enhanced weighted voting regressor for improved ensemble regression," *PLoS One*, vol. 20, 2025, doi: 10.1371/journal.pone.0319515.
[9] A. Rojarath and W. Songpan, "Cost-sensitive probability for weighted voting in an ensemble model for multi-class classification problems," *Appl. Intell.*, vol. 51, pp. 4908–4932, 2021, doi: 10.1007/s10489-020-02106-3.
[10] P. Natha, S. P. Tera, R. Chinthaginjala, S. Rab, V. Narasimhulu, and T. H. Kim, "Boosting skin cancer diagnosis accuracy with ensemble approach," *Sci. Rep.*, vol. 15, 2025, doi: 10.1038/s41598-024-84864-5.
[11] A. Assiri, S. Nazir, and S. Velastín, "Breast Tumor Classification Using an Ensemble Machine Learning Method," *J. Imaging*, vol. 6, 2020, doi: 10.3390/jimaging6060039.
[12] X. Li et al., "Beds: Bagging Ensemble Deep Segmentation For Nucleus Segmentation With Testing Stage Stain Augmentation," 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), Nice, France, 2021, pp. 659-662, doi: 10.1109/ISBI48211.2021.9433869.
[13] N. H. A. Malek, W. Yaacob, Y. B. Wah, S. A. Md Nasir, N. Shaadan, and S. Indratno, "Comparison of Ensemble Hybrid Sampling with Bagging and Boosting Machine Learning Approach for Imbalanced Data," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 29, no. 1, pp. 598–608, 2022, doi: 10.11591/ijeecs.v29.i1.pp598-608.
[14] P. Patro, T. Goel, S. A. Varaprasad, M. I. Tanveer, and R. Murugan, "Lightweight 3D Convolutional Neural Network for Schizophrenia Diagnosis Using MRI Images and Ensemble Bagging Classifier," *ArXiv*, vol. abs/2211.0, 2022, doi: 10.48550/arXiv.2211.02868.
[15] L. N. Mabumbi *et al.*, "New Approach Based on the Ensemble Learning Estimator to Maximize Accuracy," *J. Adv. Math. Comput. Sci.*, 2025, doi: 10.9734/jamcs/2025/v40i31976.
[16] S. F. Mokhtar, Z. M. Yusof, and H. Sapiri, "Confidence intervals by bootstrapping approach: a significance review," *Malaysian J. Fundam. Appl. Sci.*, vol. 19, no. 1, pp. 30–42, 2023, doi: https://doi.org/10.11113/mjfas.v19n1.2660.
[17] B. Efron and T. Hastie, *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, 2021. doi: 10.1017/9781108660966.
[18] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in Python*. Springer, 2023. doi: 10.1007/978-3-031-38747-0.
[19] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Chapman and Hall/CRC, 2024. doi: 10.1201/9781003317821.
[20] E. Rigdon, M. Sarstedt, and O. Moisescu, "Quantifying model selection uncertainty via bootstrapping and Akaike weights," *Int. J. Consum. Stud.*, vol. 47, no. 4, pp. 1596–1608, 2023, doi: https://doi.org/10.1111/ijcs.12906.
[21] B. S. Maitner *et al.*, "Bootstrapping outperforms community-weighted approaches for estimating the shapes of phenotypic distributions," *Methods Ecol. Evol.*, vol. 14, no. 10, pp. 2592–2610, 2023, doi: https://doi.org/10.1111/2041-210X.14160.
[22] G. Rousselet, C. R. Pernet, and R. R. Wilcox, "An introduction to the bootstrap: a versatile method to make inferences by using data-driven simulations," *Meta-Psychology*, vol. 7, pp. 1–24, 2023, doi: https://doi.org/10.15626/MP.2019.2058.
[23] Y. Liu, H. Zhang, and L. Wang, "Enhanced speech emotion recognition using deep fusion of MFCC and spectral features," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1124–1135, 2022, doi: 10.1109/TAFFC.2021.3112105.
[24] S. Juyal and P. Gupta, "Emotion Recognition from Speech Using Deep Neural Network," *Int. J. Adv. Comput. Sci. Appl.*, 2021, doi: 10.14569/IJACSA.2021.0120561.
[25] N. Iqbal, "MFCC and Machine Learning Based Speech Emotion Recognition Over TESS and IEMOCAP Datasets," *Int. J. Eng. Res. Technol.*, vol. 10, no. 12, 2021.
[26] Ravi and S. Taran, "Emotion Recognition in Speech Using MFCC and Energy Based Ratio Features," *2024 11th Int. Conf. Signal Process. Integr. Networks*, pp. 367–371, 2024, doi: 10.1109/spin60856.2024.10511355.
[27] D. Yuan and S. Zhang, "A Single Channel Speech Enhancement Algorithm for Long Distance Scene," in *2024 17th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2024, pp. 1–5. doi: https://doi.org/10.1109/CISP-BMEI64163.2024.10906108.
[28] X. Zhao and X. Nie, "Splitting Choice and Computational Complexity Analysis of Decision Trees," *Entropy*, vol. 23, 2021, doi: 10.3390/e23101241.
[29] Z. Saurav, M. M. Mitu, N. S. Ritu, M. A. Hasan, S. Arefin and D. M. Farid, "A New Method for Learning Decision Tree Classifier," 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE), Chittagong, Bangladesh, 2023, pp. 1-6, doi: 10.1109/ECCE57851.2023.10101557
[30] S. A. Fayaz, M. Zaman, and M. A. Butt, "Performance Evaluation of GINI Index and Information Gain Criteria on Geographical Data: An Empirical Study Based on JAVA and Python," in *International Conference on Innovative Computing and Communications*, 2021. doi: 10.1007/978-981-16-3071-2_22.