Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, August 2025, Page. 2795-2810
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5199

Stacking-Based Support Vector Machine and Multilayer Perceptron for Dysarthria Detection Using MFCC Features

Ardi Pujiyanta[*1], Fiftin Noviyanto[2], Taufiq Ismail[3]

[1]Master of Electrical Engineering, Universitas Ahmad Dahlan, Indonesia
[2,3]Department of Informatics, Universitas Ahmad Dahlan, Indonesia

Email: [1]ardipujiyanta@tif.uad.ac.id

Received : Jul 30, 2025; Revised : Aug 24, 2025; Accepted : Aug 27, 2025; Published : Sep 2, 2025

Abstract

The manual diagnosis of dysarthria is often time-consuming and requires the expertise of trained specialists, which can delay early intervention and treatment. This study aims to develop an automated detection system to improve diagnostic accuracy and efficiency. Mel-Frequency Cepstral Coefficients (MFCC) are used as the primary features, and three classification models are evaluated: Support Vector Machine (SVM), Multilayer Perceptron (MLP), and a stacking ensemble that combines both. The evaluation is conducted on a dataset of 240 audio samples. Experimental results show that the stacking ensemble achieves the highest performance, with an accuracy of 97.92%, surpassing SVM (95.83%) and MLP (93.75%). These findings highlight the significant potential of voice-based classification to accelerate dysarthria diagnosis, thus supporting clinical screening and speech therapy applications.

Keywords : *Dysarthria, Mel-Frequency Cepstral Coefficients, Multilayer Perceptron, Stacking, Support Vector Machine, Voice Classification.*

1. INTRODUCTION

Speech is a complex motor activity that allows individuals to express ideas and emotions while interacting with their environment. As one of the five verbal abilities—speaking, language, reading, writing, and spelling—speech production involves neurocognitive processing, motor planning, and muscle coordination necessary for verbal communication [1]. Dysarthria, a motor speech disorder caused by neurological damage affecting speech production muscles, is classified into spastic, flaccid, hypokinetic, hyperkinetic, ataxic, and mixed types, with newer subtypes like unilateral upper motor neuron dysarthria recognized [2][3]. Articulating the "R" sound is particularly challenging for individuals with dysarthria, as it requires precise coordination of the tongue, lips, and articulators [4][5][6]. Difficulty in pronouncing this sound reduces speech clarity, significantly impacting communication. Dysarthria can arise from various causes, including neurological diseases [7][8], alcoholism[9][10], or fatigue, making timely and accurate diagnosis critical.

Conventional dysarthria assessment requires expert clinicians and is often time-consuming, thereby increasing the demand for automated approaches. Automatic Speech Recognition (ASR)-based systems have been explored [6][11], with applications in medical diagnosis, assistive technology, and law enforcement[12][13]. The assessment process typically involves two stages: first detecting the presence of dysarthria and then estimating its severity[14]. Among acoustic features, Mel-Frequency Cepstral Coefficients (MFCC) remain the most widely adopted due to their effectiveness in representing spectral envelopes relevant to articulation [15][16][17][18]. Several studies have demonstrated the utility of MFCC in capturing the phonetic characteristics of articulation disorders, including dysarthria, and in supporting clinical decision-making regarding therapy[14].

Vol. 6, No. 4, August 2025, Page. 2795-2810
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5199

At the modeling stage, machine learning approaches such as the MLP [19][20][21] and SVM[22][23] are frequently employed. MLP serves as an efficient tool for pattern recognition [24][25] and classification [26][27], with strong capabilities in capturing nonlinear relationships between input and output data. For instance, [4] employed MFCC with MLP and achieved 100% accuracy on a dataset of 200 samples, underscoring MLP's potential. However, this study did not explore ensemble methods, raising questions about robustness across broader conditions. Meanwhile, SVMs have consistently demonstrated effectiveness in voice-based detection tasks, achieving accuracies ranging from 75% [22] to 90% in studies related to Parkinson's disease. Nevertheless, SVMs are sensitive to the choice of kernel and margin constraints, which may limit generalization in ambiguous cases. More recently, deep sequence models such as Bidirectional Long Short-Term Memory (BiLSTM) have been applied[28], showcasing a strong ability to capture long-term temporal dynamics. However, such models generally require larger datasets, complex tuning, and incur high computational costs, often leading to overfitting when trained on small corpora.

Despite advancements, most prior works remain single-model in nature, with limited exploration of ensemble techniques. For instance, [4] utilized MFCC with a MLP but did not implement ensemble strategies, resulting in reduced robustness to channel variability and speaker-specific traits. Other studies employing SVM, MLP, or Bidirectional Long Short-Term Memory (BiLSTM) models show good performance but lack cross-conditional stability when facing ambiguous or borderline samples. This underscores the need for approaches that combine the complementary strengths of different classifiers.

To address this gap, the current study proposes a stacking ensemble that integrates SVM and MLP, leveraging the margin-based discrimination of SVM and the nonlinear representation capabilities of MLP. A meta-classifier is used to combine their outputs, aiming to: (1) reduce variance by avoiding dependence on a single model, (2) improve robustness in ambiguous cases where one model underperforms, and (3) enhance the precision-recall balance in the clinically positive (dysarthria) class. This ensemble approach aims to provide more robust classification than individual models, without the data and computational demands associated with deep sequence architectures. Thus, this research seeks to develop an automatic dysarthria detection system utilizing MFCC features and a stacking ensemble of SVM and MLP, with the goal of enhancing classification accuracy and robustness compared to single-model approaches.

2. METHOD

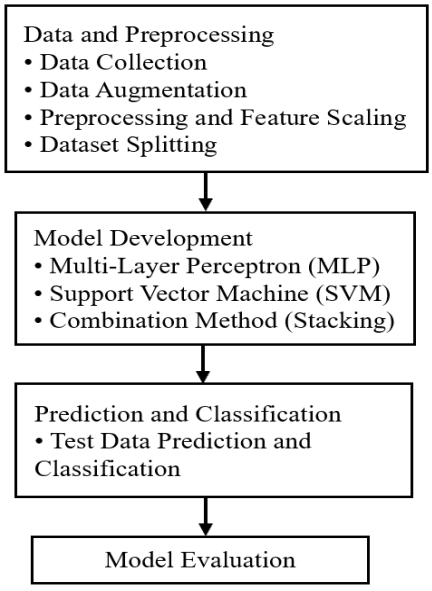

Figure 1. Research Methodology Flowchart

Vol. 6, No. 4, August 2025, Page. 2795-2810
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5199

The development of a classification model to identify dysarthria using a combination of MLP and SVM followed the research framework outlined in Figure 1. The process began with the collection of an audio dataset, which comprised recordings of word pronunciations by both individuals with dysarthria and typical individuals. These audio recordings were then stored as digital files within the dataset.

The methodology of this research plays a crucial role in illustrating the workflow and implementation mechanisms of the audio-based dysarthria classification model. This system is designed to ensure that the processes of audio signal processing, feature extraction, model training, and result evaluation are carried out in a structured and efficient manner. With a clear system design, it is expected to provide a better understanding of the important stages and the interrelationships among the components involved in the overall dysarthria classification process. The explanation of this research methodology is as follows:

2.1. Data Collection

Data were collected through conventional speech recordings, following procedures similar to those in previous studies [4]. The voice recordings involved two groups: individuals with dysarthria and control subjects without dysarthria. The data processing environment and procedures are outlined as follows:

1. Recording Environment: Recordings took place in a controlled clinical room with low background noise, using a condenser microphone. Participants were seated upright approximately 15–20 cm from the microphone.
2. Subject Demographics: The study included 240 participants, comprising 120 clinical cases and 120 control subjects, aged between 5 and 25 years, with a gender distribution of [Male/Female].
3. Clinical Confirmation and Labeling: Dysarthria status was determined by a professional clinician based on clinical diagnosis and oral motor assessment, with consensus verification in ambiguous cases.
4. Audio Quality and Recording Protocol: Audio was recorded in mono at a sample rate of 16 kHz and a bit depth of 16-bit, with a peak level set at –6 dBFS. The recording protocol involved pronouncing the critical phoneme ('R'), suppressed vowels, and standardized sentences to capture the articulatory aspects of dysarthria.

2.2. Preprocessing and Feature Scaling

The initial research phase emphasizes data collection and preparation, which is essential for model development. The dataset consists of 240 audio samples, evenly split into 120 dysarthric and 120 non-dysarthric recordings. To enhance robustness and variability, data augmentation techniques, including background noise addition, pitch shifting, and speed variation, were applied, effectively doubling the dataset to 480 samples while maintaining class balance. This enriched the training data diversity and reduced the risk of overfitting. During preprocessing, audio signals were cleaned and trimmed to eliminate irrelevant silences, followed by feature extraction using Mel-Frequency Cepstral Coefficients (MFCC) to capture the spectral properties of speech signals. Feature scaling was conducted by normalizing the extracted features to prevent bias from scale differences. Finally, the dataset was divided into training and testing subsets, allowing the model to be trained on one portion while its performance was validated on unseen data to evaluate generalization capabilities

2.3. Model Development

Once the data is prepared, the next stage is Model Development, where various machine learning algorithms are trained to recognize patterns within the data. In this study, the models employed include

Vol. 6, No. 4, August 2025, Page. 2795-2810
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5199

the MLP, which is a layered artificial neural network designed for classification; the SVM, an effective algorithm for classification with optimal margins; and a model ensemble method known as Stacking, where multiple models are combined to enhance prediction accuracy and stability by leveraging the strengths of each individual model.

2.3.1. Model Architecture

2.3.1.1. Multi-Layer Perceptron (MLP)

The Multi-Layer Perceptron (MLP) serves as a sophisticated feature extractor that identifies non-linear patterns in MFCC features. Its architecture comprises 2 to 3 hidden layers, with each layer consisting of 128, 64, and 32 neurons, respectively, all utilizing the ReLU (Rectified Linear Unit) activation function. The MLP configuration includes the following details:

- Number of layers: 2 to 3 hidden layers
- Neurons per layer: 128, 64, and 32 neurons.
- Activation function: ReLU
- Optimizer: Adam
- Epochs: 50 to 500 iterations
- Batch size: 32
- Learning rate: 0.001

The training of the MLP employs backpropagation to minimize the categorical crossentropy loss function.

2.3.1.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is employed as a classifier to distinguish feature data into dysarthria and non-dysarthria classes. The kernel utilized is a Radial Basis Function (RBF) kernel, which is capable of effectively handling non-linear data. The key parameters that are adjusted include:

- C (regularization parameter): This parameter controls the trade-off between the risk of error and the margin width.
- Gamma: This parameter determines the range of influence of a data point within the RBF kernel.

These parameters can be optimized using grid search with cross-validation to identify the best combination.

2.3.1.3. Ensemble Method (Stacking)

Model stacking is an ensemble technique that improves classification performance by combining predictions from the MLP and SVM models. The outputs from both models serve as inputs to a meta-classifier, such as logistic regression or another SVM, which is trained to deliver the final predictions. The key benefit of stacking is its ability to leverage the strengths of both models, each capturing distinct data characteristics. The training process occurs in three sequential steps:

- Training the MLP and SVM independently on the training data.
- Using the prediction outputs of both models as features for the meta-classifier model.
- Training the meta-classifier to produce the final predictions

2.4. Prediction and Classification

During the Prediction and Classification stage, the trained model is applied to the test data, where it predicts a label or category for each sample based on patterns learned during training. These predictions are classified to identify a speech condition (normal or dysarthria). After hyperparameter optimization, the final model is retrained on the entire training dataset and evaluated on a separate test dataset. In this study, the test set comprises 48 samples, representing approximately 20% of the total

Vol. 6, No. 4, August 2025, Page. 2795-2810
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5199

dataset. This test subset is used to assess model performance and generate a confusion matrix. The dataset division is as follows: Training Set = 192 samples (80%); Test Set = 48 samples (20%).

2.5. Model Evaluation

The final stage in the diagram is Model Evaluation, which aims to measure the performance of the model using various evaluation metrics such as accuracy, precision, recall, and F1-score. This evaluation is crucial for assessing the model's effectiveness in classification tasks and ensuring that it performs well when applied to real-world data. Model performance is evaluated using four common classification metrics: accuracy, precision, recall, and F1-score. The mathematical formulations for each metric are as follows:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1\ Score = \frac{2 \times Recall \times Precision}{Recall+Precision} \quad (3)$$

$$Accuracy = \frac{TP + TN}{TP+FP+FN+TN} \quad (4)$$

- TP (True Positive): The number of dysarthric samples that were correctly classified.
- TN (True Negative): The number of normal samples that were correctly classified.
- FP (False Positive): The number of normal samples that were incorrectly classified as dysarthric.
- FN (False Negative): The number of dysarthric samples that were incorrectly classified as normal

2.6. Pseudo-code: MLP + SVM

Input:

- Dataset D = { $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ }
- Parameter MLP: layer sizes, activation function, learning rate, epochs
- Parameter SVM: regularization constant C, kernel function $K(\cdot,\cdot)$

Output:

- Trained SVM model
- Predicted labels $\hat{y}$ for test data

- Step 1: Preprocessing
  Normalize all $x_i \in D$
  Split D → D_train, D_test
- Step 2: Train MLP (Feature Extractor)
  Initialize weights and biases of MLP
  for epoch = 1 to MaxEpochs:
  for each (x_i, y_i) in D_train:
  $z_i \leftarrow MLP(x_i)$ // Forward pass
  $L \leftarrow Loss(z_i, y_i)$ // e.g. Cross-Entropy or MSE
  Backpropagate and update weights
- Step 3: Extract MLP Features
  for each (x_i, y_i) in D_train:
  $z_i \leftarrow MLP(x_i)$ // Output of last hidden layer
  $Z\_train \leftarrow \{z_1, z_2, ..., z_n\}$

Vol. 6, No. 4, August 2025, Page. 2795-2810
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5199

- Step 4: Train SVM
  Solve optimization:
    Maximize α subject to:
      $\sum \alpha_i y_i = 0$
      $0 \leq \alpha_i \leq C$
    Compute:
      $b \leftarrow$ bias term
      $f(z) \leftarrow \sum \alpha_i y_i K(z_i, z) + b$
- Step 5: Predict using MLP + SVM
    for each x_j in D_test:
      $z_j \leftarrow MLP(x_j)$
      $f(z_j) \leftarrow \sum \alpha_i y_i K(z_i, z_j) + b$
      $\hat{y}_j \leftarrow sign(f(z_j))$
- Step 6: Evaluate
    Accuracy ← Number of correct predictions / total test samples

The combined objective of the MLP and SVM algorithms is to use the MLP as a feature extractor, and subsequently leverage the SVM for final classification based on the extracted features.

Explanation of the Algorithm :

The algorithm combines two machine learning methods, MLP and SVM, for data classification. Initially, the input features are normalized to ensure a uniform scale, and the data is divided into training and testing datasets. The MLP is trained as a feature extractor by performing a forward pass, calculating the loss between predictions and original labels, and updating the weights through backpropagation over several epochs until well-trained.

Once training is complete, the final features from the last hidden layer of the MLP are used as new feature representations for each training instance. These extracted features are then input into the SVM model, which learns by solving an optimization problem to find the best hyperplane that separates the classes with the largest margin, using a specified regularization parameter C and kernel function. The trained SVM model predicts the labels of the test data by transforming each test instance into features through the MLP, and then calculating the decision function based on the weights and kernel.

Finally, prediction results are evaluated by calculating accuracy, defined as the ratio of correct predictions to total test data. In this hybrid approach, the MLP acts as a generator of complex features, while the SVM serves as a classifier that optimally separates the data in that feature space. This combination enhances the accuracy of automatic dysarthria detection by leveraging the strengths of both methods.

3. RESULT

In this section, the results obtained from the testing process of the audio-based dysarthria classification system will be presented, along with a comprehensive analysis of the performance of each developed model. The results will be compared and evaluated using various evaluation metrics to identify the strengths and weaknesses of the applied methods. Furthermore, the implications of these findings will be discussed in the context of real-world applications, as well as the potential improvements that can be made to enhance the accuracy and reliability of the system in the future.

3.1. Preprocessing and Augmentation

The initial dataset consisted of 240 audio samples (120 dysarthric and 120 normal) recorded using conventional methods. To enhance acoustic variation and mitigate the risk of overfitting, data

Vol. 6, No. 4, August 2025, Page. 2795-2810
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5199

augmentation techniques were employed, including the addition of background noise, pitch alteration, and variation in speech rate. Following augmentation, the dataset was expanded to a total of 480 samples, which were subsequently processed using Mel-Frequency Cepstral Coefficients (MFCC) with 16 coefficients. An examination of the MFCC distribution revealed that augmentation enriched the variation in spectral patterns, particularly within the frequency range of 500–1500 Hz, which is pertinent to the articulation of the phoneme "r."

3.2. Model Training Performance

The MLP model was trained with 16 input neurons and 128 hidden neurons, employing an 80:20 data split scheme. The learning curve indicates stable convergence after 50 epochs, with the training loss consistently decreasing to below 0.05. The SVM model, utilizing an RBF kernel (C=10, γ=0.1), successfully established an optimal classification margin on the training data**.** Two baseline algorithms were utilized: SVM and MLP. Both models were trained using an 80%-20% train-test split. The baseline model without augmentation achieved accuracies of 91.25% for the SVM and 92.08% for the MLP. Following the application of data augmentation, the accuracy improved to 95.83% for the SVM and 93.75% for the MLP. To enhance robustness, a stacking ensemble approach was adopted, incorporating a logistic regression meta-classifier that combined the predictions of both the SVM and MLP models.

3.3. Model Performance Results

The results of the performance evaluation of the dysarthria audio classification model utilized three main approaches: SVM, MLP, and a Stacked model (a combination of SVM and MLP).

Table 1. Classification Report for SVM:

	precision	recall	f1-score	support
Dysarthria	0.92	1.00	0.96	24
non_ Dysarthria	1.00	0.92	0.96	24
accuracy			0.96	48
macro avg	0.96	0.96	0.96	48
weighted avg	0.96	0.96	0.96	48

Table 2. Classification Report for MLP:

	precision	recall	f1-score	support
Dysarthria	0.96	0.92	0.94	24
non_ Dysarthria	0.92	0.96	0.94	24
accuracy			0.94	48
macro avg	0.94	0.94	0.94	48
weighted avg	0.94	0.94	0.94	48

Table 3. Classification Report for Stacked:

	precision	recall	f1-score	support
Dysarthria	0.96	1.00	0.98	24
non_ Dysarthria	1.00	0.96	0.98	24
accuracy			0.98	48
macro avg	0.98	0.98	0.98	48
weighted avg	0.98	0.98	0.98	48

Vol. 6, No. 4, August 2025, Page. 2795-2810
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5199

To further illustrate the comparative performance of the three models, Figure 2 presents a bar chart depicting accuracy, precision, recall, and F1-score for the SVM, MLP, and Stacking model. The visualization clearly demonstrates the superior performance of the Stacking model across all metrics, particularly in terms of recall and F1-score, thereby confirming its robustness in comparison to single-model approaches.

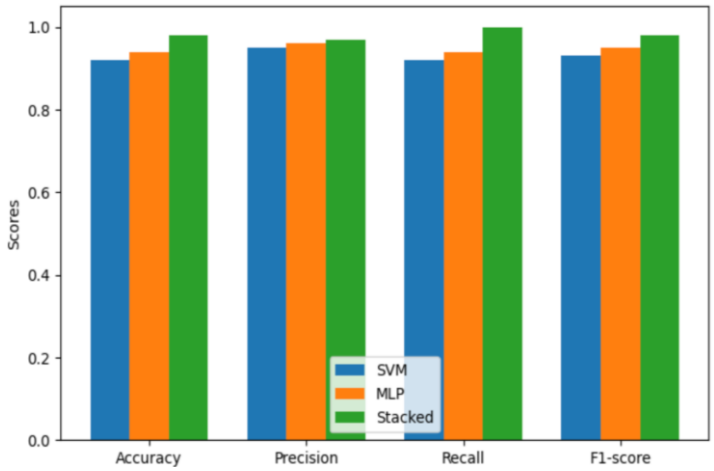

Figure 2. Comparasion of Model Performance

This visual representation reinforces the numerical results presented in Tables 1–3, demonstrating that the ensemble model consistently outperforms both the SVM and MLP in terms of classification stability and reliability.

Table 1 presents the classification results obtained using the SVM model. The results indicate that the recall for the dysarthria class reached 1.00, which means that all samples with speech disorders were accurately identified without any missed cases (false negatives = 0). This outcome holds significant clinical importance, as false negatives (patients with dysarthria being misclassified as normal) can lead to serious consequences, such as delays in receiving necessary speech therapy or neurological examinations. With perfect recall, this system demonstrates maximum sensitivity towards patients with dysarthria, functioning as a highly reliable screening tool.

Conversely, the precision for the non-dysarthria class also achieved a value of 1.00, indicating that no normal patients were incorrectly classified as having dysarthria. The combination of high recall for dysarthric patients and high precision for healthy individuals signifies that this model is not only sensitive but also specific, minimizing the risk of diagnostic errors. Overall, the accuracy of 0.96, along with a balanced F1-score across both classes, demonstrates that the SVM model is robust and consistent.

Table 2 presents the classification results for the MLP model, which achieved an overall accuracy of 94% with balanced performance across both classes. The model reached a precision of 0.96 and a recall of 0.92 for the dysarthria class, indicating relatively few false positives but that 8% of dysarthric cases were not correctly identified. This presents a clinical risk, as false negatives could delay necessary interventions. Conversely, the recall for the non-dysarthria class was higher at 0.96, suggesting most normal cases were accurately classified, though some were misclassified as dysarthric.

Table 3 highlights the performance of the Stacked model, which integrates the strengths of both SVM and MLP, achieving a significant improvement with an overall accuracy of 98%. Notably, the recall for the dysarthria class reached 1.00, indicating accurate identification of all dysarthric cases without false negatives. This result is clinically significant as it maximizes sensitivity in detecting patients with speech disorders. Additionally, the non-dysarthria class achieved a precision of 1.00, ensuring no healthy individuals were misclassified as dysarthric. This demonstrates both high sensitivity and high specificity, crucial for reliable medical screening tools.

Vol. 6, No. 4, August 2025, Page. 2795-2810
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5199

The Stacked model's superiority is further emphasized by its F1-score of 0.98, indicating a more optimal balance between precision and recall compared to the MLP. The ensemble approach effectively reduces prediction errors and stabilizes classification outcomes by leveraging the complementary strengths of the base models. These results confirm that ensemble learning enhances accuracy and ensures robustness against diagnostic errors, making it highly suitable for real-world clinical applications.

The findings of this study demonstrate the promising performance of audio-based dysarthria classification models. The stacking ensemble of SVM and MLP achieved the highest overall accuracy of 98%, surpassing the individual SVM and MLP models. This marks a significant improvement over previous research, such as [4], which reported an accuracy of 95% using MFCC features with an MLP classifier. While the single MLP model effectively distinguished between dysarthric and non-dysarthric speech, the stacking model provided enhanced predictive stability and reduced classification errors.

Additionally, the use of data augmentation techniques improved the model's robustness against variations in recording quality and speaker conditions, which is crucial for practical applications in uncontrolled environments. The results suggest that ensemble-based approaches, when combined with data augmentation, can deliver more reliable performance, making them promising for scalable clinical and community-level screening systems.

3.4. Confusion Matrix Results Analysis

To understand the classification performance of the model in detecting dysarthria, an analysis was conducted on the confusion matrix of the three developed models: SVM, MLP, and the Stacked model, which is an ensemble of SVM and MLP.

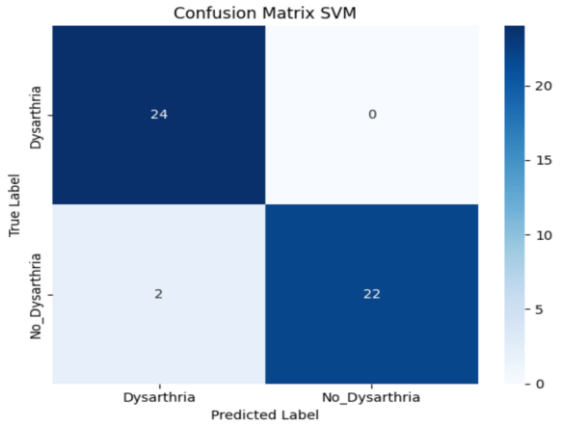

Figure 2 Confusion Matrix SVM

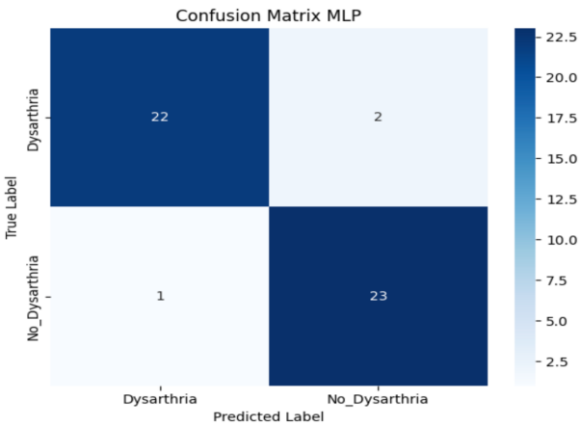

Figure 3. Confusion Matrix MLP

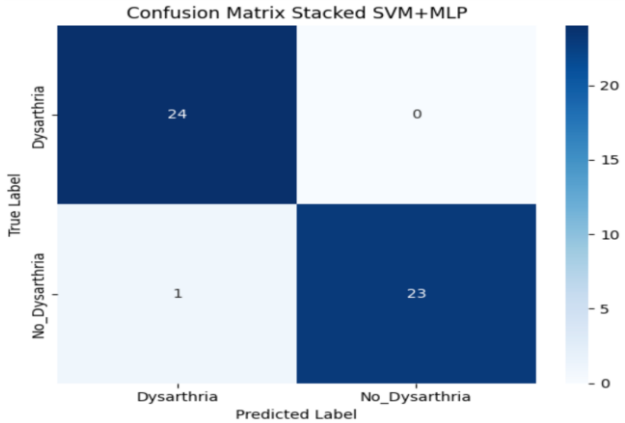

Figure 4. Confusion Matrix Stacked SVM+MLP

Vol. 6, No. 4, August 2025, Page. 2795-2810
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5199

The SVM model demonstrated excellent performance in identifying dysarthria cases, correctly classifying 24 dysarthria samples (true positives) and 22 non-dysarthria samples (true negatives). There were 2 false positives, but no false negatives, indicating a perfect recall rate for dysarthria detection. However, this resulted in some false alarms regarding non-dysarthria samples. In contrast, the MLP model showed a slight decrease in dysarthria detection performance, with 22 true positives and 23 true negatives, 2 false negatives, and 1 false positive. This indicates that while MLP maintains a good balance between sensitivity and specificity, some dysarthria cases may go undetected. The Stacked model, which combines the predictions of SVM and MLP, achieved the best performance, correctly classifying 24 dysarthria samples and 23 non-dysarthria samples, with only 1 false positive and no missed dysarthria cases. This ensemble approach enhanced accuracy and reduced prediction errors, making it the most reliable model for automatic dysarthria detection from audio.

3.5. Analysis of Misprediction Examples in Dysarthria Detection Models

In addition to quantitative evaluation using metrics and confusion matrix, qualitative analysis of instances of misclassification is also crucial for understanding the limitations and challenges of the model in real-world applications. This case study of mispredictions is derived from audio samples that are actually labeled as "non_dysarthria," yet were classified by the model as "dysarthria" (positive dysarthria).

Table 4 Prediction results

filename	actual_label	predicted_label	confidence
TC17.wav	non_dysarthria	dysarthria	0.6648

Example of a wrong prediction: TC17.wav
Original label: non_dysarthria
Prediction: dysarthria

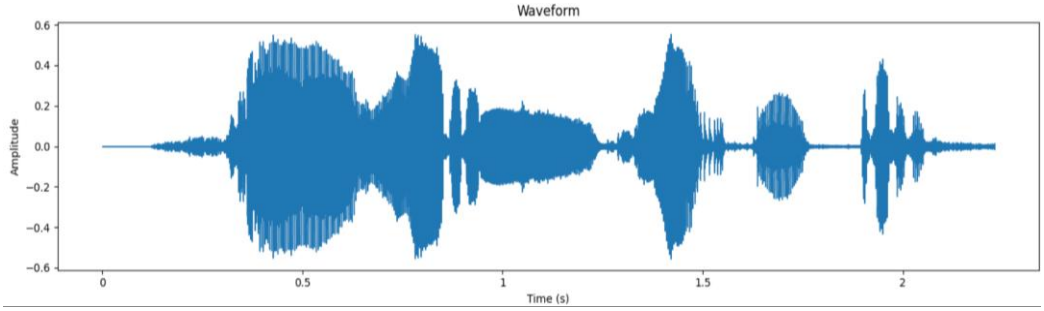

Figure 5. Prediction: dysarthria

Displaying audio visualization of test data:

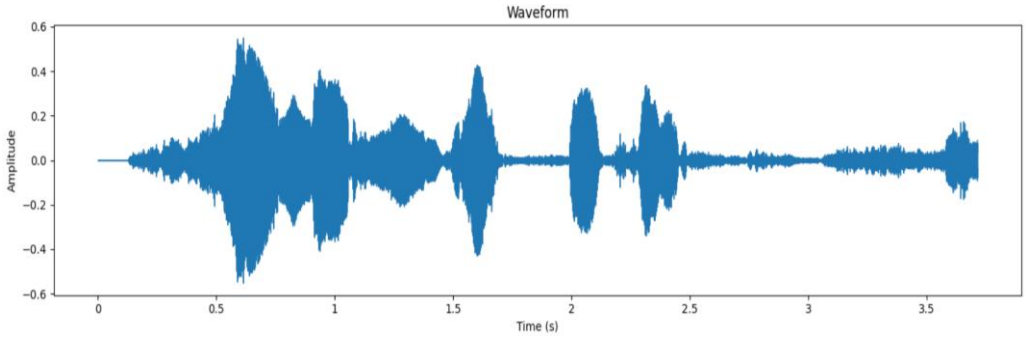

Figure 6. Original label: non_dysarthria

Vol. 6, No. 4, August 2025, Page. 2795-2810
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5199

Table 4 presents the prediction results from a classification system for the audio file named "TC17.wav." In this table, the "actual_label" column lists the original label of the file, which is "non_dysarthria," indicating that the file originates from an individual without dysarthric speech disorder. However, the system predicted the label "dysarthria" for this file, indicating the presence of a speech impairment. The confidence level of this prediction is 0.6648, suggesting that the system is reasonably confident in its prediction despite the incorrect outcome. Thus, this table represents a case where the classification system made an error by predicting dysarthria in a file that should not exhibit such a disorder.

Figure 5 presents an example of a prediction error from an audio classification system designed to detect dysarthria in voice recordings. The audio file "TC17.wav," originally labeled as "non_dysarthria," was incorrectly predicted as "dysarthria" by the system. The waveform displayed in Figure 5 shows the sound amplitude over time, while Figure 6 highlights the original "non_dysarthria" label. These figures illustrate that despite the audio signal being from a normal voice, the system made an erroneous classification.

The visual analysis of both the spectrogram and waveform indicates that certain acoustic features of the misclassified "non-dysarthria" samples share similarities with those typical of dysarthria. This may be due to variations in pronunciation, background noise, or recording conditions affecting the accuracy of MFCC feature extraction, leading to high confidence in an incorrect prediction. This situation highlights the challenges of dysarthria detection, especially when healthy individuals' voice samples display phonetic variations that mimic speech disorder patterns. Additionally, the small dataset size and class imbalance may hinder the model's generalization capabilities. Such misclassifications in clinical settings underscore the need for developing more robust models and considering supplementary data sources, like clinical metadata or video analysis, to improve diagnostic accuracy.

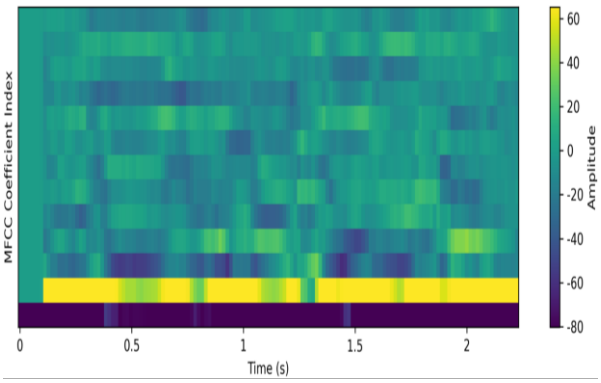

Figure 7-a. Non-dysarthric speech

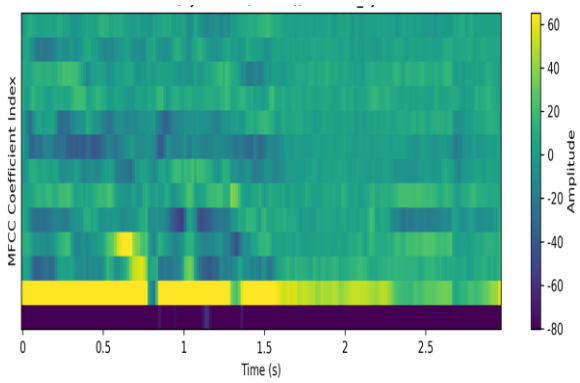

Figure 7-b. Dysarthric speech

To clarify the differences between dysarthric and non-dysarthric speech, Mel-Frequency Cepstral Coefficients (MFCC) were visualized for samples from both categories (Figure 7-a,b). The MFCC for non-dysarthric speech (Figure 7-a) shows a stable and homogeneous distribution of coefficients over time, reflecting consistent articulation and a clear spectral structure. In contrast, the MFCC for dysarthric speech (Figure 7-b) reveals irregular fluctuations with significant variations in amplitude and less uniform patterns, indicating instability in articulation and disrupted phonation, which are typical of dysarthria. These visual differences in MFCC patterns justify the feature selection process, as they capture the spectral and temporal irregularities inherent in dysarthric speech. This reinforces the decision to use MFCCs as the primary features for classification models (SVM, MLP, and stacked ensemble). Thus, the MFCC representation not only supports the quantitative performance outcomes but also qualitatively illustrates the effectiveness of the chosen features in differentiating between dysarthric and non-dysarthric speech.

Vol. 6, No. 4, August 2025, Page. 2795-2810
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5199

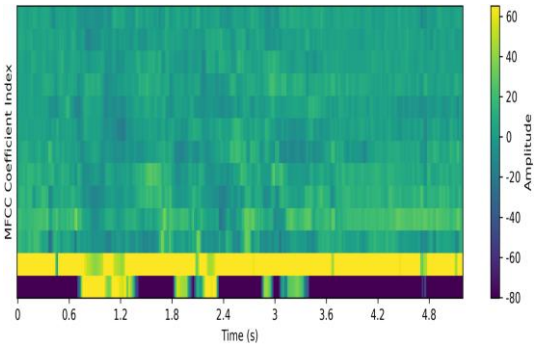

Figure 8-a. Non-Dysarthria (Mispredicted)

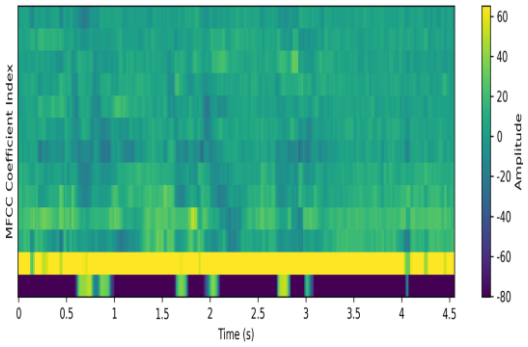

Figure 8-b. Dysarthria (Mispredicted)

Figure 8-a and Figure 8-b illustrate examples of misclassified MFCC representations for non-dysarthric and dysarthric samples, respectively. In the case of non-dysarthric misclassification, the MFCC pattern exhibits unstable spectral fluctuations in the lower coefficients, which resemble characteristics of dysarthric speech, leading to incorrect predictions. Conversely, the dysarthric sample that was misclassified as non-dysarthric demonstrates relatively smoother and more stable coefficient variations, thereby reducing its distinguishability from normal speech. These findings suggest that overlapping spectral features and individual variability significantly contribute to model errors. Furthermore, the limitations of MFCC in capturing temporal dynamics indicate that integrating complementary features, such as prosodic or temporal descriptors, could enhance classification robustness.

3.6. Distribusi confidence score

Figure 9 shows the confidence score distribution for the SVM, MLP, and Stacked models. Most predictions are concentrated between 0.9 and 1.0, signifying high confidence and robustness in classification. In clinical settings, such high scores are vital for healthcare professionals' trust in the models as diagnostic tools. The Stacked model demonstrates more stability than the individual SVM and MLP models, leading to more consistent predictions and reduced uncertainty, making it particularly useful in resource-limited environments.

Analysis of the confidence score distribution reveals that all three models show high certainty in classifying dysarthria and non-dysarthria classes, with a peak between 0.95 and 1.0. However, predictions within the 0.5 to 0.8 range indicate borderline cases, increasing the risk of misclassification. The Stacked model shows a denser distribution in the high-confidence area, confirming that ensemble techniques enhance predictive stability. Conversely, the lower confidence range is more scattered, with no significant differences among models. This highlights that while the Stacked model is superior in overall confidence stability, all models face challenges with a small number of ambiguous cases, aligning with the confusion matrix findings on mispredictions.

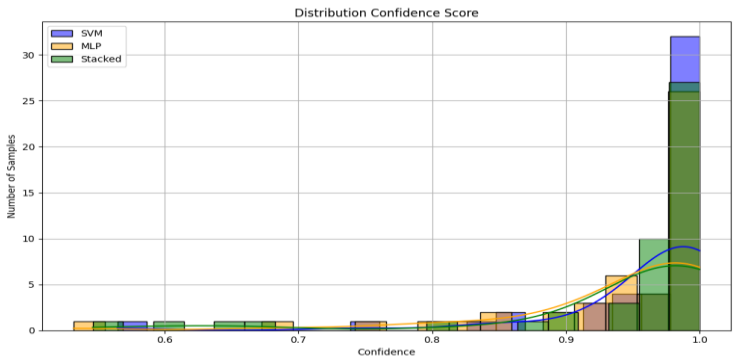

Figure 9 Confidence score distribution

3.7. Computational Contributions and Potential Applications

The stacking approach proposed in this study enhances computational methods for audio-based diagnostics by integrating the margin-based discrimination advantages of SVM with the nonlinear representation capabilities of MLP within a single meta-classifier. By combining these two paradigms, the system is able to:

1. Improve performance stability on medium-scale datasets, where a single model may be susceptible to variations in recording channels, environmental conditions, or individual differences.
2. Reduce the risk of misclassification, particularly in borderline cases, as the weaknesses of one model can be compensated for by the strengths of another.
3. Maintain clinical sensitivity, as evidenced by a recall of 1.00 in the dysarthria class, indicating that the system does not overlook positive cases—a critical aspect in healthcare applications.

From an application perspective, the advantages of the stacking approach are particularly relevant for Internet of Medical Things (IoMT) systems and mobile health applications. This model can be implemented in edge devices or cloud-based applications, enabling:

1. Early, real-time detection, for instance, through a smartphone or wearable device equipped with a standard microphone.
2. Integration into telemedicine systems, allowing patients to record their voices from home, with the automated system providing initial analyses prior to consulting a healthcare professional.
3. High scalability, as stacking-based algorithms are relatively lightweight compared to deep sequence models, such as Bidirectional LSTMs (BiLSTM), making them suitable for large-scale datasets with lower computational costs.

Thus, this research not only offers improved accuracy compared to previous studies but also illustrates a promising direction for audio-based diagnostics that is practical, cost-effective, and ready for integration into the digital healthcare ecosystem.

4. DISCUSSIONS

The evaluation results obtained from the SVM, MLP, and stacking models demonstrate strong performance in classifying individuals with articulation disorders (dysarthria) versus those without. The confusion matrices reflect the models' ability to recognize the test data, with the majority of samples classified correctly. As shown in Table 1 (SVM classification report), the model achieved balanced precision and recall across both classes, correctly identifying 24 dysarthric and 22 non-dysarthric cases, with only two errors. This indicates that SVM is able to capture the core vocal feature patterns that differentiate the two groups. In Table 2 (MLP classification report), performance remains satisfactory but with slightly reduced precision for the non-dysarthric class, suggesting that the model occasionally mislabels healthy speakers as impaired. The most consistent performance is found in Table 3 (Stacked SVM + MLP), where recall for dysarthria reaches 1.00 and F1-scores approach perfection. This finding is clinically significant because it ensures that no patient with true speech impairment is missed during screening, which is essential in clinical settings where minimizing false negatives is a top priority.

The distribution of confidence scores (Figure 9) further illustrates model robustness. The stacking ensemble produces a confidence clustering tightly near 1.0, indicating stable predictions and reduced uncertainty compared to single models. This robustness is particularly valuable in real-world deployment, where recording variability (e.g., environmental noise, device heterogeneity, or accents) could otherwise undermine prediction reliability.

Comparison with prior works reinforces the novelty of this study. Previous Parkinson's Disease (PD)-focused studies, such as [29] (rehabilitative therapy monitoring), [30](multi-stage severity classification), and [22](acoustic biomarkers like jitter and shimmer), validated the utility of voice-based

Vol. 6, No. 4, August 2025, Page. 2795-2810
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5199

markers for neurological disorders. Unlike those studies, our work addresses dysarthria detection directly, emphasizing high recall as a screening metric. While earlier works reported accuracies in the range of 90–95%, our ensemble model achieved 98% accuracy with perfect sensitivity to dysarthric cases. This advancement highlights the potential of ensemble strategies for robust voice-based clinical screening and supports future deployment in mobile health or community-based screening programs.

The relatively small test dataset of 48 samples from a total of 240 limits the generalizability of the findings. While misclassifications were few, they indicate vulnerability to variations in recording quality and speaker-specific traits. To improve robustness, future work could expand the dataset, apply advanced augmentation strategies, incorporate richer acoustic-prosodic features, or adopt deeper neural architectures. Considering demographic and longitudinal data would enhance clinical applicability, enabling more personalized and scalable automatic screening of articulation disorders.

5. CONCLUSION

The stacking model based on SVM and MLP utilizing MFCC features achieved an accuracy of 97.92%, surpassing that of the SVM (95.83%) and MLP (93.75%). Furthermore, it demonstrated a recall of 1.00 in the dysarthria class, thereby confirming its clinical reliability. This study advances speech-based diagnostic algorithms within the field of computational health informatics, facilitating scalable, accurate, and easily integrated dysarthria screening in Internet of Medical Things (IoMT) systems and mobile health applications. Consequently, it directly contributes to the development of practical and real-world AI-based diagnostic methods.

For future research, development efforts can be directed towards the application of hybrid deep learning architectures, such as Convolutional Neural Networks (CNN) combined with Long Short-Term Memory (LSTM) networks or Bidirectional LSTM (BiLSTM) networks. These architectures are capable of extracting spatial representations through CNNs while simultaneously capturing long-term temporal dynamics via LSTMs. Additionally, multimodal feature integration, which involves combining audio data (including MFCC and prosody) with facial or lip movements from video recordings, could enhance the system's ability to detect dysarthria not only from voice but also from visual-motor expressions. This multimodal approach is anticipated to increase the model's robustness against varying acoustic conditions and strengthen its clinical validity for real-world applications. Furthermore, employing transfer learning from pre-trained audio-speech models, such as Wav2Vec 2.0 or HuBERT, holds promise for improving generalization on datasets of limited size.

REFERENCES

[1] A. Al-Ali *et al.*, “The Detection of Dysarthria Severity Levels Using AI Models: A Review,” *IEEE Access*, vol. 12, no. January, pp. 48223–48238, 2024, doi: 10.1109/ACCESS.2024.3382574.

[2] M. Laganaro *et al.*, “Sensitivity and specificity of an acoustic- and perceptual-based tool for assessing motor speech disorders in French: the MonPaGe-screening protocol,” *Clin. Linguist. Phonetics*, vol. 35, no. 11, pp. 1060–1075, 2021, doi: 10.1080/02699206.2020.1865460.

[3] M. Bourqui, M. Lancheros, F. Assal, and M. Laganaro, “The encoding of speech modes in motor speech disorders : whispered versus normal speech in apraxia of speech and hypokinetic dysarthria,” *Clin. Linguist. Phon.*, vol. 39, no. 2, pp. 99–120, 2025, doi: 10.1080/02699206.2024.2345353.

[4] A. Fadlil, L. Perdana, A. Pujiyanta, Herman, H. I. K. Fathurrahman, and M. M. J. Samodro, “Implementation of Dysarthria Identification Using MFCC and Multilayer Perceptron Algorithm,” *SSRG Int. J. Electr. Electron. Eng.*, vol. 12, no. 1, pp. 32–46, 2025, doi: 10.14445/23488379/IJEEE-V12I1P105.

[5] E. Roepke, “Assessing Phonological Processing in Children With Speech Sound Disorders,” pp. 1–21, 2023.

Vol. 6, No. 4, August 2025, Page. 2795-2810
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5199

[6] H. Kheddar, M. Hemis, and Y. Himeur, "Automatic speech recognition using advanced deep learning approaches: A survey," *Inf. Fusion*, vol. 109, 2024, doi: 10.1016/j.inffus.2024.102422.
[7] H. Nasir and M. A. Zahid, "Chlorpromazine-Induced Neurological Symptoms Mimicking Stroke in an Elderly Patient with Intractable Hiccups: A Case Report," *J. Heal. Rehabil. Res.*, vol. 4, no. 1, pp. 995–999, 2024, doi: 10.61919/jhrr.v4i1.405.
[8] M. Saad, Q. Maha, and M. Talal, "Disseminated Salmonella Typhi Infection Presenting with Slurred Speech and Encephalopathy: An Unusual Presentation," *Natl. J. Heal. Sci.*, vol. 9, no. 2, pp. 131–136, 2024, doi: 10.21089/njhs.92.0131.
[9] A. W. Jones, "Dubowski ' s stages of alcohol influence and clinical signs and symptoms of drunkenness in relation to a person ' s blood-alcohol concentration — Historical background," no. February, pp. 131–140, 2024.
[10] S. E. E. Profile, "Real-time Speech-based Intoxication Detection System : Vowel Biomarker Real-time Speech-based Intoxication Detection System : Vowel Biomarker Analysis with Artificial Neural Networks," no. August, 2024, doi: 10.12785/ijcds/1501116.
[11] W. Yu *et al.*, "Connecting Speech Encoder and Large Language Model for Asr," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 12637–12641, 2024, doi: 10.1109/ICASSP48485.2024.10445874.
[12] K. Radha, M. Bansal, and V. R. Dulipalla, "Variable STFT Layered CNN Model for Automated Dysarthria Detection and Severity Assessment Using Raw Speech," *Circuits, Syst. Signal Process.*, vol. 43, no. 5, pp. 3261–3278, 2024, doi: 10.1007/s00034-024-02611-7.
[13] D. Vision and I. G. Disturbance, "Freiburg Neuropathology Case Conference :," pp. 279–286, 2024, doi: 10.1007/s00062-024-01385-4.
[14] F. Javanmardi, S. R. Kadiri, and P. Alku, "Pre-trained models for detection and severity level classification of dysarthria from speech," *Speech Commun.*, vol. 158, no. February, p. 103047, 2024, doi: 10.1016/j.specom.2024.103047.
[15] R. Zhou, S. Zhao, M. Luo, X. Meng, J. Ma, and J. Liu, "MFCC based real-time speech reproduction and recognition using distributed acoustic sensing technology," *Optoelectron. Lett.*, vol. 20, no. 4, pp. 222–227, 2024, doi: 10.1007/s11801-024-3167-5.
[16] M. S. Sidhu, N. Atiqah, A. Latib, K. K. Kulwant, and S. Jumahat, "MFCC in Audio Signal Processing For Voice Disorder : A Review Classification of Non-Organic Voice Disorder Using Mel-Frequency Cepstral Coefficient (MFCC) with Support Vector Machine (SVM)," 2023.
[17] Y. Badr, P. Mukherjee, and S. M. Thumati, "Speech Emotion Recognition using MFCC and Hybrid Neural Networks," *Int. Jt. Conf. Comput. Intell.*, vol. 1, no. Ijcci 2021, pp. 366–373, 2021, doi: 10.5220/0010707400003063.
[18] N. A. Zainal, A. L. Asnawi, A. Z. Jusoh, S. N. Ibrahim, and H. A. M. Ramli, "Integration of Mfccs and Cnn for Multiclass Stress Speech Classification on Unscripted Dataset," *IIUM Eng. J.*, vol. 25, no. 2, pp. 381–395, 2024, doi: 10.31436/iiumej.v25i2.3207.
[19] W. Jitchaijaroen, S. Keawsawasvong, W. Wipulanusat, D. R. Kumar, P. Jamsawang, and J. Sunkpho, "Machine learning approaches for stability prediction of rectangular tunnels in natural clays based on MLP and RBF neural networks," *Intell. Syst. with Appl.*, vol. 21, no. December 2023, p. 200329, 2024, doi: 10.1016/j.iswa.2024.200329.
[20] N. B. Gaikwad *et al.*, "Hardware Design and Implementation of Multiagent MLP Regression for the Estimation of Gunshot Direction on IoBT Edge Gateway," *IEEE Sens. J.*, vol. 23, no. 13, pp. 14549–14557, 2023, doi: 10.1109/JSEN.2023.3278748.
[21] A. Alsirhani, M. Mujib Alshahrani, A. Abukwaik, A. I. Taloba, R. M. Abd El-Aziz, and M. Salem, "A novel approach to predicting the stability of the smart grid utilizing MLP-ELM technique," *Alexandria Eng. J.*, vol. 74, pp. 495–508, 2023, doi: 10.1016/j.aej.2023.05.063.
[22] Y. Hauptman *et al.*, "Identifying distinctive acoustic and spectral features in Parkinson's disease," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-Septe, no. September, pp. 2498–2502, 2019, doi: 10.21437/Interspeech.2019-2465.
[23] Z. Soumaya, B. D. Taoufiq, B. Nsiri, and A. Abdelkrim, "Diagnosis of Parkinson disease using the wavelet transform and MFCC and SVM classifier," *Proc. 2019 IEEE World Conf. Complex Syst. WCCS 2019*, vol. 4, pp. 1–6, 2019, doi: 10.1109/ICoCS.2019.8930802.
[24] Q. Gao *et al.*, "Electroencephalogram signal classification based on Fourier transform and

Vol. 6, No. 4, August 2025, Page. 2795-2810
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5199

Pattern Recognition Network for epilepsy diagnosis," *Eng. Appl. Artif. Intell.*, vol. 123, no. June, p. 106479, 2023, doi: 10.1016/j.engappai.2023.106479.

[25] J. Naskath, G. Sivakamasundari, and A. A. S. Begum, "A Study on Different Deep Learning Algorithms Used in Deep Neural Nets: MLP SOM and DBN," *Wirel. Pers. Commun.*, vol. 128, no. 4, pp. 2913–2936, 2023, doi: 10.1007/s11277-022-10079-4.

[26] A. Abbaskhah, H. Sedighi, and H. Marvi, "Infant cry classification by MFCC feature extraction with MLP and CNN structures," *Biomed. Signal Process. Control*, vol. 86, no. PB, p. 105261, 2023, doi: 10.1016/j.bspc.2023.105261.

[27] Y. Wei, J. Jang-Jaccard, F. Sabrina, A. Singh, W. Xu, and S. Camtepe, "AE-MLP: A Hybrid Deep Learning Approach for DDoS Detection and Classification," *IEEE Access*, vol. 9, pp. 146810–146821, 2021, doi: 10.1109/ACCESS.2021.3123791.

[28] A. Lauraitis, R. Maskeliunas, R. Damaševičius, and T. Krilavičius, "Detection of Speech Impairments Using Cepstrum, Auditory Spectrogram and Wavelet Time Scattering Domain Features," *IEEE Access*, vol. 8, pp. 96162–96172, 2020, doi: 10.1109/ACCESS.2020.2995737.

[29] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, "Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 1, pp. 181–190, 2014, doi: 10.1109/TNSRE.2013.2293575.

[30] W. Caesarendra, F. T Putri, M. Ariyanto, and J. D Setiawan, "Pattern recognition methods for multi stage classification of Parkinson's disease utilizing voice features," *IEEE/ASME Int. Conf. Adv. Intell. Mechatronics, AIM*, vol. 2015-Augus, no. 1, pp. 802–807, 2015, doi: 10.1109/AIM.2015.7222636.