SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V12I9P112

Volume 12 Issue 9, 119-127, September 2025
© 2025 Seventh Sense Research Group®

Original Article

Improved Hybrid Tuning Mel Frequency Cepstral Coefficients with Ant Colony Optimization, and Long Short Term Memory on Speech Hoarseness Detection

Noraziahtulhidayu Kamarudin[1], SAR Al Haddad[2]

[1]*Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia.*
[2]*Department of Computer and Communication Engineering, Faculty of Engineering, Universiti Putra Malaysia.*

[1]*Corresponding Author : noraziah@uthm.edu.my*

Received: 08 July 2025 Revised: 10 August 2025 Accepted: 09 September 2025 Published: 30 September 2025

Abstract - *Hoarseness speech detection through machine learning has been discussed quite extensively. However, not many people are trying to apply with different datasets and identify the type of algorithm that would be able to produce high accuracy, with the appropriate precision, recall, and F1-score. Two types of datasets are used in this study, including the Kaggle Speech dataset and the Saarbrucken Voice Dataset (SVD). The disadvantages of the Mel Frequency Cepstral Coefficient that affect the accuracy rate are overcome by using feature selection techniques, pitch features, and the selection of appropriate coefficients. From this technique, the accuracy rate has increased, especially using the selection of different coefficient parameters and the feature selection technique. Through this study, the increase in accuracy and increased performance metrics show the advantages of machine learning techniques in identifying hoarse and normal voices, especially in cancer patients.*

Keywords - *Speech hoarseness, Normal, Hoarse speech, Ant colony optimization, Long short-term memory, Feature selection, Feature vector.*

1. Introduction

Hoarseness often manifests as a change in voice quality, including breathiness, roughness, or strain, which introduces variability in the spectral structure of the speech signal. Mel-Frequency Cepstral Coefficients (MFCCs) are consistently highlighted as a crucial and widely utilized feature extraction technique in the field of voice analysis, particularly for voice pathology detection and speech recognition. However, the sources also implicitly and explicitly point toward areas where MFCCs might have limitations or require further consideration.

The final consideration is that the algorithm's implementation could not be strong enough [1], and this study started to conduct a more detailed study to produce more significant and fulfilling results. Patients' prognosis and quality of life can be greatly enhanced by improving laryngeal cancer diagnosis and therapy. There is potential for Artificial Intelligence (AI) technology to be a useful diagnostic tool for laryngeal cancer. However, obtaining accuracy and efficiency in AI-based diagnosis offers obstacles since laryngeal cancer lesions are hidden and heterogeneous [2]. The usage of machine learning algorithms may well improve the finding and justification of laryngeal cancer that contributes to speech hoarseness (Kim H et. al(2020) [3]; Marrero-Gonzalez et. al (2025)) [4]. The study [5] reveals that detection accuracy is significantly influenced by the MFCC frame length, with a longer frame length of 500 ms yielding the best results in their experiments. The paper details the standard MFCC extraction process and notes that MFCCs are also part of larger feature sets. This study directly addresses a lack of systematic investigation into the effect of a basic MFCC attribute, like frame length, in voice pathology detection.

As in Table 1, previous research shows that longer frame lengths can enhance the laryngeal pathology detection by providing more data using frame-to-frame analysis. Overall, spectral shapes, incorporated with feature selection of speech signals, may be helpful with feature extraction to improve the final results of classification and find the best technique to be used in detecting speech hoarseness. The MFCCs may capture similar sound patterns for different types of voice disorders, making it difficult to classify them accurately. MFCCs have weaknesses in low-frequency analysis, which can be important for capturing certain aspects of hoarseness. Below are some preliminary analyses by three researchers that utilizing MFCC with a comparison of a few techniques of classification focusing on speech hoarseness. The results are quite convincing, but in-depth studies are needed to present an optimizing robust algorithm in identifying speech hoarseness.

Table 1. Analysis of feature extraction and classification for speech hoarseness

Feature	Leite et al. (2022) [6]	Islam et al. (2022) [7]	Narendra & Alku (2020) [8]
Dataset Used	Analyzed 435 samples (/e/ vowel) from dysphonic (384) and non-dysphonic (51) individuals. Categorization based on laryngeal examination and perceptual judgment.	Saarbrücken Voice Database (SVD). Binary: 150 control, 65 pathological (/a/ vowel). Multiclass: subset of pathological samples.	Not explicitly stated in the provided text for review.
Feature Extraction	Extracted 34 acoustic measures. Used variance threshold for selection, resulting in 15 features (PA, HNR, CPP, CPPs, SFR, PM, ENTR, RPDE).	Raw EGG and speech signals were used directly as input to CNNs (avoiding explicit extraction).	Explored glottal source features (QCP, ZFF, direct acoustic) and openSMILE features. Compared to MFCC and PLP.
Classification Methods	Compared 10 supervised ML classifiers (RF, NB, SVM, MLPC, DT, GBC, KNN, SGDC, ETC, LR) with k-fold cross-validation and Bayesian optimization.	Proposed a dual cascaded CNN system (CNN-1 for binary, CNN-2 for multiclass) with 5-fold cross-validation.	Used SVM and deep learning networks (CNN+MLP, CNN+LSTM) for classification.
Key Findings	NB and SGDC performed best on 15 acoustic features (SGDC: Accuracy 0.91, Kappa 0.57; NB: Accuracy 0.76, Kappa 0.45). Variance threshold found useful for feature selection.	Binary: Speech signals better than EGG. Multiclass: EGG generally has better F1 for laryngitis and polyps (accuracy 88.67%). Aimed for low computational burden.	Glottal source features comparable or better than MFCC/PLP with SVM. Raw glottal flow improved the accuracy in deep learning models.
Signal Type(s) Used	Acoustic (sustained /e/ vowel)	Electroglottographic (EGG) and speech signals (sustained /a/ vowel).	Acoustic (implied for glottal flow and comparison to MFCC/PLP), likely glottal source signals, and possibly others for openSMILE.
Voice Disorder Focus	General dysphonia detection.	Binary (pathological vs. healthy) and specific disorders (dysphonia, laryngitis, vocal fold polyps).	General voice classification (likely including disordered voices based on the context in Islam et al.).

The above Table 1 summarized those studies that explore different approaches to voice disorder detection, utilizing various signal types (acoustic, EGG, glottal source), feature extraction techniques (explicit acoustic features, direct signal input, glottal source parameters), and classification methods (traditional machine learning and deep learning). The findings highlight the potential of different feature sets and classifiers for achieving accurate and efficient voice disorder detection.

From Leite et. al [6], the classification process that took place; compared 10 supervised machine learning classifiers: Random Forest (RF), Naive Bayes (NB), Support Vector Machines (SVM), Multilayer Perceptron Classifier (MLPC), Decision Tree (DT), Gradient Boosting Classifier (GBC), K-Nearest Neighbor (KNN), Stochastic Gradient Descent Classifier (SGDC), Extra-Tree Classifier (ETC), and Logistic Regression (LR). Used k-fold cross-validation for training and testing. Bayesian optimization was used to efficiently tune the parameters for each model, ensuring the best configuration was found before training. And for the results acquired for the Naive Bayes (NB) and Stochastic Gradient Descent Classifier (SGDC) performed best on the reduced dataset of 15 acoustic measures. SGDC achieved an accuracy of 0.91 and a Kappa of 0.57, while NB achieved an accuracy of 0.76 and a Kappa of 0.45. Only NB and SGDC met the eligibility criteria (accuracy, sensitivity, specificity, and F1-Score > 0.70, and Kappa > 0.40) with variance thresholds of 0.020, 0.025, and 0.030. The study concluded that the variance threshold is useful for automatic feature selection and reduction.

The researcher in [8] explored glottal source features (using QCP, ZFF, and directly from acoustic signals) and openSMILE features and used SVM and deep learning networks (CNN+MLP, CNN+LSTM) for classification. The glottal source features are comparable to or better than MFCC and PLP with SVM, and glottal flow as raw input improved accuracy in deep learning models. The study on MFCC frame length focuses on task-specific optimization [5, 9] and clearly demonstrates that the default parameters of MFCC extraction might not be optimal for all tasks. The best frame length for voice pathology detection (500 ms in their study) differed significantly from the typical shorter frame lengths used in

Automatic Speech Recognition (ASR). This highlights the deficiency in the MFCC that it is a universal, one-size-fits-all approach to MFCC parameterization. While MFCCs effectively capture spectral envelope information related to the vocal tract, the research proposing a hybrid model [10] suggests that relying solely on MFCCs might not be sufficient for achieving the highest accuracy in complex tasks like discerning various voice disorders. The integration of features like fundamental frequency (related to vocal fold vibration) and spectral centroid (related to spectral energy distribution) provides a more comprehensive view of vocal quality, indicating a potential lack in MFCCs when used in isolation for nuanced pathological voice analysis. The review paper [11] points out that the interaction between MFCC features can lead to redundancy, potentially increasing computational cost without necessarily adding significant discriminative information. This implies a deficiency in the raw MFCC output regarding inherent feature selection or optimization for efficiency, necessitating additional feature selection techniques.

2. Methodology

Throughout these findings, the algorithm used for this research has tried to overcome the shortcomings that exist in MFCC by integrating different coefficient settings, feature selection techniques, including ACO and CNN or SVM algorithms, to prove which enhancement of the algorithm. Can produce robust results compared to previous researchers.

Some researchers utilize feature selection methods to optimize the accuracy for speech hoarseness detection. This study has already selected different classification algorithms and 2 sets of datasets (the Saarbrucken voice dataset and the Kaggle Dataset on hoarseness and normal dataset), to identify the variety of performance analysis, including accuracies, F1-score, and recall.

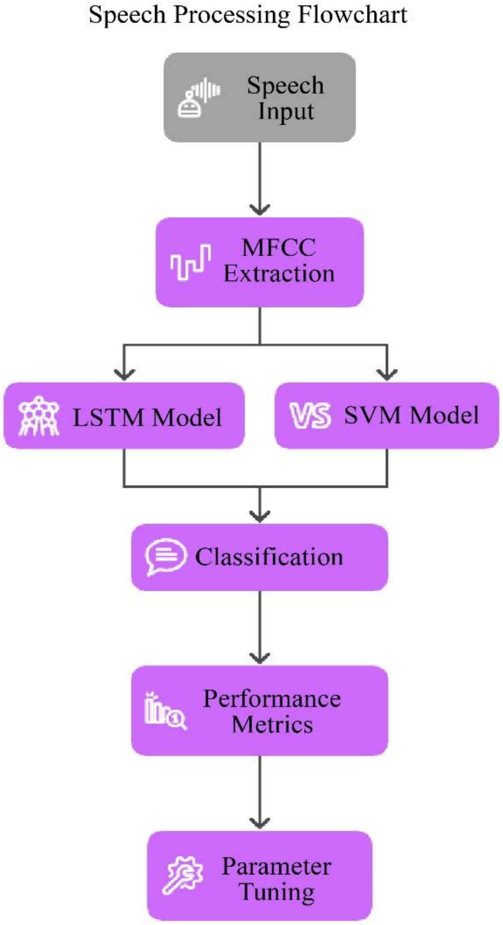

Fig. 1 Speech processing flow chart

Figure 1 depicts several types of algorithms involved from the beginning of recognising sound types involving hoarse sounds and normal sounds. In the feature extraction phase, Mel Frequency Cepstral Coefficients (MFCC) were used. During this study, two datasets were used, namely the Saarbrucken Voice Dataset (SVD) and the Kaggle Patient Dataset. This dataset is used on each of the specified algorithms. MFCC is an algorithm that has been identified as the best algorithm for feature extraction in the field of voice recognition. The MFCC process involves the speech data in the range of 20-40ms, and the MFCC algorithm process involves framing, windowing, Discrete Fourier Transform (DFT) or Fast Fourier Transform (FFT) that is applied for each frame to convert the signal from the time domain to the frequency domain. The deficient addressed in MFCC has been tried to overcome by integrating different feature techniques in improving the final result with a few types of classification. The format frequencies, pitch variation, and feature selection with parameter tuning on selected coefficients have been used in this study to get improved and robust results of speech hoarseness detection. The frequency spectrum is processed using Mel's filter bank. The bank consists of a triangular filter that is more closely spaced at lower frequencies, reflecting the nonlinear frequency perception of the human auditory system and the magnitude of the filtered spectrum is converted to a logarithmic scale. Discrete Cosine Transformation (DCT) [12-14] is used on the log-mail spectrum to further compress the information and produce the MFCC coefficient, which will be used in conjunction with the classification algorithm, namely Long Short Term Memory (LSTM) and Support Vector Machine (SVM) [15]. For this classification stage, this voice dataset has been tested with the Ant Colony Optimization (ACO) algorithm to identify the advantages of using this feature selection technique. At the initial stage of the algorithm, the ants are placed at the starting point (e.g., the nest) and start looking for the best route at random. Each ant builds a path by choosing the next node to visit probabilistically, considering the level of pheromones and other heuristic information. Once the route is complete, the ants deposit pheromones on the route they take. The number of pheromones deposited depends on the quality of the pathway (e.g., shorter pathways receive more pheromones). Over time, the pheromones in the pathway evaporate, encouraging the ant to explore other options. This process is repetitive, with the ants following the traces of stronger pheromones, which leads to better identification of pathways over time.

For the classification process, two types of algorithms are used, namely Long Short Term Memory Algorithm (LSTM), the same technique used [16, 17], and Support Vector Machine (SVM). LSTMs receive sequential data, such as word sequences, time series data, or audio. The input data, along with the hidden state and the previous state of the cell, is fed into the LSTM unit. The gate selectively controls the flow of information, storing, updating and retrieving information from memory cells. LSTM units generate new hidden states based on current inputs, previous hidden states, and cell states. The hidden state is then used as the output for the current time step, and the LSTM unit moves to process the next input. The researchers utilize an LSTM network, a type of recurrent neural network, to analyze the combined feature sets and effectively classify voice pathologies [18]. For Support Vector Machines (SVMs), SVMs are trained on labeled data, meaning they learn from instances where the desired output is already known. Although known primarily for classification, SVM can also be used for regression tasks by modifying objective functions. In a multidimensional characteristic space, the boundaries of results between classes are represented by hyperplanes. SVM focuses on finding hyperplanes that maximize margins, which provide wider separation between classes and are less sensitive to noise [19, 20]. The experiments also involved two types of datasets, which are known as the Saarbrucken Voice Dataset(SVD) and the Kaggle Patient Speech Dataset. The coefficients selected involved 1 to 13 coefficients, and another involved 13 to 20 coefficients.

3. Results and Discussion

Based on the sources provided, the results pertain to different methods and datasets used for detecting speech hoarseness. The performance is evaluated using metrics such as Accuracy, Precision, Recall, and F1-Score. Different combinations of features and classifiers are tested, as in Table 2 below:

Table 2. Algorithms used for feature extraction, feature selection and pattern classification

a)	MFCC (Mel-Frequency Cepstral Coefficients) combined with Pitch and SVM (Support Vector Machine).
b)	MFCC (Mel-Frequency Cepstral Coefficients) combined with Pitch and SVM (Support Vector Machine).
c)	MFCC is only combined with SVM.
d)	MFCC coefficients (specifically 1-13 or 13-39) combined with SVM.
e)	MFCC combined with ACO (Ant Colony Optimization) for feature selection and SVM.
f)	MFCC combined with LSTM (Long Short-Term Memory) and SVM.

Table 3. Results for MFCC and SVM-based methods (without ACO or LSTM)

Method	Accuracy (%)	F1-Score (%)	Recall (%)	Precision (%)
MFCC + Pitch + SVM (SVD)	66.67	66	66	66
MFCC + SVM only (SVD)	66.67	62	62	62
MFCC + SVM (Kaggle Dataset)	87.50	47	50	44

By using the SVD dataset, with MFCC and Pitch for the feature extraction, SVM achieved 66.67% accuracy, 66% F1-Score, 66% Recall, and 66% Precision. MFCC and SVM also achieved 66.67% Accuracy, with 62% F1-score, 62% Recall, and 62% Precision. Specifically using MFCC coefficients 1-13 with SVM, the SVD dataset showed 66.67% accuracy, 0.62 precision, 0.62 recall, and 0.62 F1-Score. Using MFCC coefficients 13-39 with SVM, the SVD dataset achieved

100.00% Accuracy, 1.00 Precision, 1.00 Recall, and 1.00 F1-Score. This indicates a perfect score across all metrics for this specific configuration on the SVD dataset, as depicted in Table 3.

With the Kaggle Dataset analysis utilizing MFCC and SVM, achieved 87.50% Accuracy, with a 47% F1-Score, 50% Recall, and 44% Precision. For MFCC coefficients selected 1 to 13 with SVM, the Kaggle dataset showed 87.50% Accuracy, 0.44 Precision, 0.50 Recall, and 0.47 F1-Score. With 13-39 coefficients, 13-39 with MFCC and SVM, the Kaggle dataset showed 87.50% Accuracy, but 0.00 Precision, 0.00 Recall, and 0.00 F1-score. Despite high accuracy, the zero values for Precision, Recall, and F1-Score suggest potential issues, such as an imbalanced dataset where the model predicts the majority class correctly most of the time but fails to identify instances of the minority class.

For results for MFCC and ACO with SVM (Kaggle dataset), these models used 10 iterations of ACO for feature selection and 5-fold cross-validation for SVM evaluation; whereby Model 1 (using ACO-selected features from MFCC 1-13) achieved 86.96% Accuracy. However, Precision, Recall, and F1-Score were reported as 0.00. Model 2 (using ACO-selected features from MFCC 13-39) achieved 86.96% Accuracy, but Precision, Recall, and F1-score were not reported.

For the Kaggle dataset utilizing MFCC, LSTM and SVM with 13 coefficients, training logs show Mini-batch Accuracy reaching 87.50% and validation accuracy consistently around 85.16% across different iterations. Validation loss stabilizes around 0.42-0.43. For the SVD dataset using 13 coefficients, training logs show Mini-batch Accuracy reaching 70.83% and validation accuracy reaching 66.67%. Validation loss stabilizes around 0.69. On the SVD dataset, using MFCC coefficients 13-39 with SVM achieved perfect scores (100% Accuracy, 1.00 Precision, 1.00 Recall, 1.00 F1-Score), which is a very strong result. However, other methods and coefficient ranges on SVD yielded lower performance, around 66% Accuracy. On the Kaggle dataset, SVM-based methods generally showed higher Accuracy (around 87%) compared to SVD, but often had very low or zero values for Precision, Recall, and F1-score, especially when using MFCC 13-39 or with ACO feature selection. This discrepancy between high accuracy and low other metrics on the Kaggle dataset suggests the model might be biased towards predicting one class, performing poorly at identifying instances of the other class. The LSTM model on the Kaggle dataset showed a validation accuracy of about 85.16%, while on the SVD dataset, it showed about 66.67%.

The results highlight that the choice of dataset, the range of MFCC coefficients used, and the specific method (SVM, SVM and Pitch, SVM+ACO, LSTM+SVM) significantly impact the performance metrics. The stark difference in results between using MFCC 1-13 and 13-39 on the Kaggle and SVD datasets is particularly notable, suggesting that the higher-order coefficients (13-39) might not be suitable for this dataset with a simple SVM classifier. Therefore, suggested that the dataset used may affect the final classification with the selection of coefficients of the features. Both models used 10 iterations of ACO and a 5-fold cross-validation for final SVM evaluation.

In summary, the results table and logs present performance figures for various machine learning models and feature sets applied to two different datasets for speech hoarseness detection, as shown in Figures 2 to 5. The metrics indicate varying levels of success depending on the specific combination used. Despite high accuracy, the extremely low/zero Precision, Recall, and F1-scores on the Kaggle dataset for certain configurations point to potential issues in the model's ability to correctly classify all instances, possibly due to dataset characteristics like class imbalance, as depicted in Table 4.

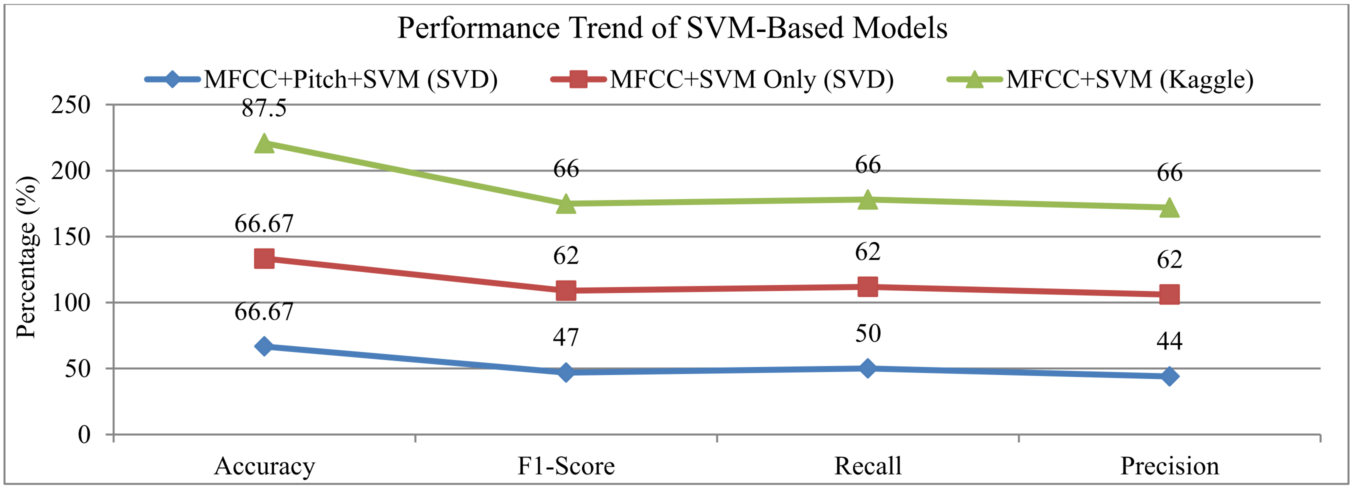

Fig. 2 Performance for SVM-based models and SVD

Dataset	MFCC Coefficients	Accuracy	Precision	Recall	F1-Score
SVD	1–13	66.67%	0.62	0.62	0.62
Kaggle	1–13	87.50%	0.44	0.50	0.47
SVD	13–39	100.00%	1.00	1.00	1.00
Kaggle	13–39	87.50%	0.00	0.00	0.00

Fig. 3 Comparison between SVD and Kaggle dataset perfomance

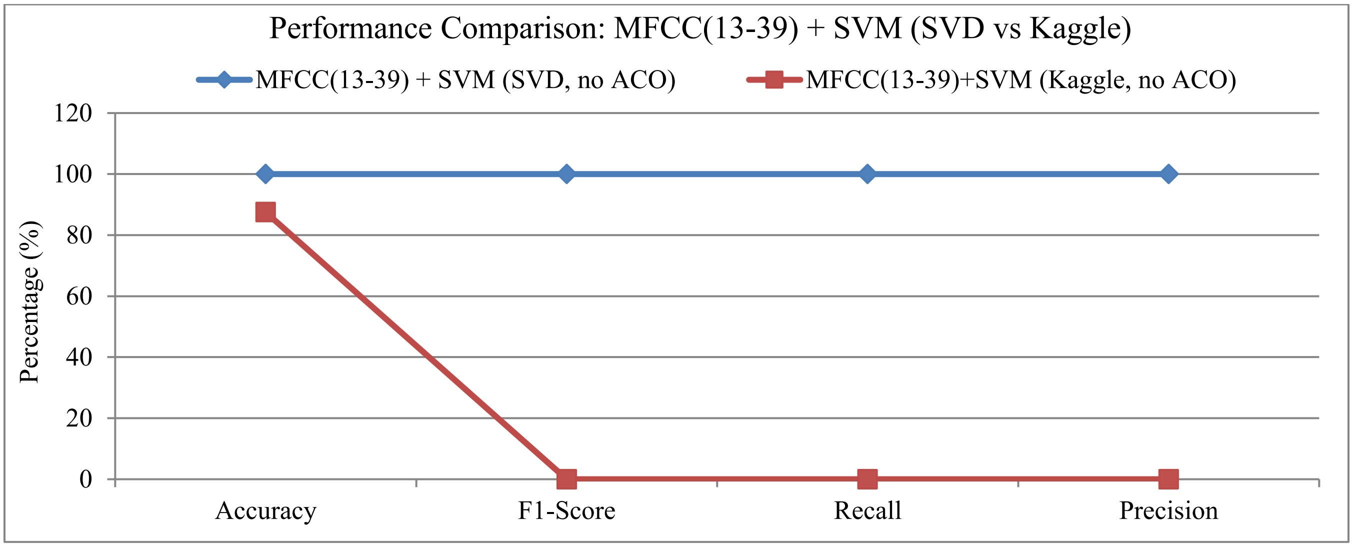

Fig. 4 Performance comparison for 13 to 39 coefficients without ACO selected features (SVD vs Kaggle dataset)

Table 4. Performance comparison for different coefficients with ACO selected features (SVD vs Kaggle dataset)

Model	MFCC Coefficients	ACO-Selected Features	Accuracy (%)	Precision	Recall	F1-Score
MFCC + ACO + SVM (Model 1)	1–13	[9 11 12 2 5 4 3 6 1 13]	86.96	0.00	0.00	0.00
MFCC + ACO + SVM (Model 2)	13–39	[11 8 26 14 7 2 18 5 3 19]	86.96	—	—	—

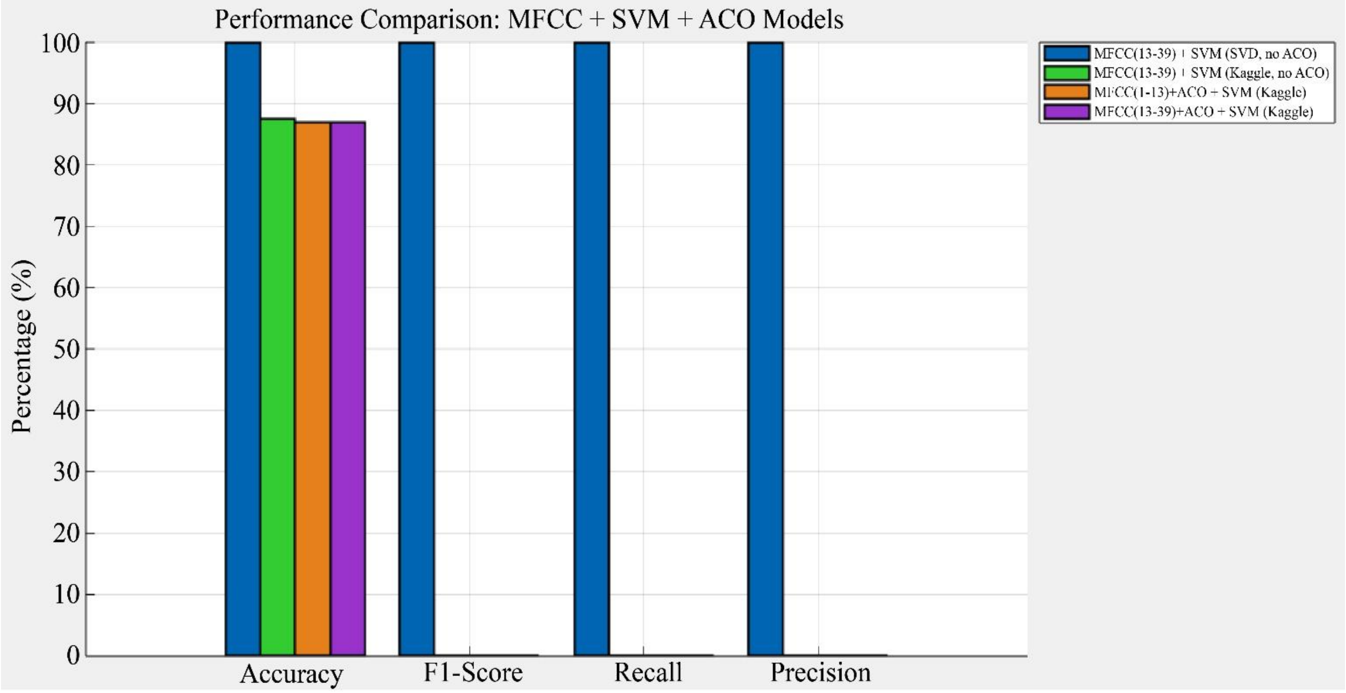

Fig. 5 MFCC, MFCC-ACO and SVM (for Kaggle and SVD dataset)

Table 5. MFCC, LSTM and SVM (Kaggle dataset) 13 coefficients

Epoch	Iteration	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Validation Accuracy	Mini-batch Loss	Validation Loss	Learning Rate
1	1	00:00:16	90.62%	85.16%	0.6360	0.6122	0.0010
7	50	00:00:21	84.38%	85.16%	0.4554	0.4270	0.0010
13	100	00:00:23	89.06%	85.16%	0.3369	0.4266	0.0010
19	150	00:00:25	87.50%	85.16%	0.3792	0.4260	0.0010
20	160	00:00:25	84.38%	85.16%	0.4326	0.4279	0.0010

Table 6. MFCC + LSTM + SVM (SVD dataset) 13 coefficients

Epoch	Iteration	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Validation Accuracy	Mini-batch Loss	Validation Loss	Learning Rate
1	1	00:00:08	47.92%	41.67%	0.6980	0.6931	0.0010
20	20	00:00:11	70.83%	66.67%	0.6623	0.6905	0.0010

The performance of the models varied significantly between the Saarbrucken Voice Dataset and the Kaggle Patient Speech Dataset, as shown in Tables 4 and 5. For the initial MFCC (1-13 coefficients) with SVM, the Kaggle dataset yielded much higher accuracy (87.50%) compared to the Saarbrucken dataset (66.67%). This suggests that these datasets might have different characteristics or levels of complexity regarding the features relevant for the classification task. However, when using MFCC (13-39 coefficients) with SVM without ACO, the Saarbrucken dataset achieved perfect classification (100% accuracy, precision, recall, F1-score), while the Kaggle dataset showed a high accuracy (87.50%). Focusing on different ranges of MFCC coefficients (1-13 vs. 13-39) substantially impacted performance.

For the Saarbrucken dataset, using the 13-39 coefficients resulted in a perfect score, a dramatic improvement over the 1-13 coefficients. This indicates that the higher-order MFCC coefficients might capture more discriminative information for this specific dataset. Conversely, for the Kaggle dataset without ACO, the 13-39 coefficients resulted in a degenerate classifier (0% precision, recall, F1-score), despite a relatively high accuracy. This suggests a potential issue with class imbalance or the features themselves in this higher-order range for this dataset. Applying the Ant Colony Optimization (ACO) algorithm for feature selection on the Kaggle dataset (using both 1-13 and 13-39 coefficients) resulted in similar accuracy (around 86.96%) across different ACO iterations. Interestingly, despite the consistent accuracy reported by ACO, the precision, recall, and F1-score remained at 0.00% after evaluating the SVM with the selected features.

This strongly suggests that while ACO might have identified subsets of features that lead to good overall accuracy, these subsets might be failing to correctly classify at least one of the classes, leading to the zero values in other metrics. The selected feature indices also differ between the two coefficient ranges. The initial epochs of the MFCC+LSTM+SVM model on a single CPU show promising training and validation accuracies (around 90% and 85% respectively, in the first epoch). The loss values also decrease as training progresses. This suggests that combining temporal modeling (LSTM) with spectral features (MFCC) and a classifier (SVM) could be a viable approach. However, the training was still in progress, and further epochs would be needed to assess its ultimate performance. The meaningful differences in performance between the two datasets show that understanding the specific characteristics of the data matters. All these variations could stem from factors such as noise levels, recording conditions, the distribution of classes, and the nature of the speech tasks.

The datasets themselves warrant further analysis to determine more of the perceptions. This is to understand why certain features and models perform better on one versus the other. The contrasting results with different MFCC coefficient ranges suggest that the relevant information for the classification task might be concentrated in different frequency bands for the two datasets. The higher-order coefficients seem crucial for the Saarbrucken dataset, while for the Kaggle dataset (without ACO), they lead to a problematic classifier. The Kaggle results with 13-39 coefficients (without ACO) and the ACO-selected features demonstrate a crucial point: high accuracy alone is not a sufficient metric for evaluating classifier performance, especially in cases of potential class imbalance. The 0% precision, recall, and F1-score indicate a severe issue in correctly identifying instances of at least one class. The consistent accuracy of nearly 87% merits deeper study with 0% precision, recall, and F1-score after ACO feature selection on the Kaggle dataset. Per its current settings, the ACO algorithm may consistently select features for a biased classifier. Alternatively, intrinsic challenges inside the Kaggle dataset itself might make it difficult to achieve good precision and recall with these feature sets and the SVM classifier. Combining different types of models for capturing spectral and temporal dependencies in speech might be a fruitful direction, as the initial results of the MFCC+LSTM+SVM model suggested in the analysis. The complete capacity of this model is not known yet. Therefore, training must be continued, and hyperparameters must be fine-tuned.

4. Conclusion

In conclusion, this study offers insightful information about the difficulties and possibilities of classifying speech data. Careful data exploration and algorithm selection are crucial, as evidenced by the strong dataset dependency and the contrasting performance with various feature sets and selection techniques. The hybrid LSTM-based model's early success points to a promising direction for further study. Perhaps more balanced feature subsets can be selected for the future direction of this research, and different parameters for the ACO algorithm should be tested. Better precision, along with recall, may be achieved through this experimentation. These findings also offer insightful information about the difficulties and possibilities in classifying speech data. Careful data exploration and algorithm selection are crucial, as evidenced by the strong dataset dependency and the contrasting performance with various feature sets and selection techniques. The hybrid LSTM-based model's early success points to a promising direction for further study.

Funding

The research was supported by the Ministry of Higher Education through the Fundamental Research Grant Scheme(FRGS-EC/1/2024/ICT06/UTHM/03/1).

Acknowledgments

The authors also would like to extend much appreciation for the resources provided by Saarbrucken Voice Dataset (SVD) and Kaggle Laryngeal dataset and the research is supported by the Ministry of Higher Education through the Fundamental Research Grant Scheme(FRGS-EC/1/2024/ICT06/UTHM/03/1).

References

[1] Ariel Roitman et al., “Harnessing Machine Learning in Diagnosing Complex Hoarseness Cases,” *American Journal of Otolaryngology*, vol. 46, no. 1, pp. 1-6, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[2] Xin Nie et al., “Laryngeal Cancer Diagnosis Based on Improved YOLOv8 Algorithm,” *Machine Learning: Science and Technology*, vol. 3, no. 1, pp. 1-14, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[3] HyunBum Kim et al., “Convolutional Neural Network Classifies Pathological Voice Change in Laryngeal Cancer with High Accuracy,” *Journal of Clinical Medicine*, vol. 9, no. 11, pp. 1-15, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Alejandro R. Marrero-Gonzalez et al., “Application of Artificial Intelligence in Laryngeal Lesions: A Systematic Review and Meta-Analysis,” *European Archives of Oto-Rhino-Laryngology*, vol. 282, no. 3, pp. 1543-1555, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Saska Tirronen, Sudarsana Reddy Kadiri, and Paavo Alku, “The Effect of the MFCC Frame Length in Automatic Voice Pathology Detection,” *Journal of Voice*, vol. 38, no. 5, 975-982, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Danilo Rangel Arruda Leite, Ronei Marcos de Moraes, and Leonardo Wanderley Lopes, “Different Performances of Machine Learning Models to Classify Dysphonic and Non-Dysphonic Voices,” *Journal of Voice*, vol. 39, no. 3, pp. 577-590, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[7] Rumana Islam, Esam Abdel-Raheem, and Mohammed Tarique, “Voice Pathology Detection Using Convolutional Neural Networks with Electroglottographic (EGG) and Speech Signals,” *Computer Methods and Programs in Biomedicine Update*, vol. 2, pp. 1-13, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] N.P. Narendra, and Paavo Alku, “Glottal Source Information for Pathological Voice Detection,” *IEEE Access*, vol. 8, pp. 67745-67755, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] Saska Tirronen, Sudarsana Reddy Kadiri, and Paavo Alku, “The Effect of the MFCC Frame Length in Automatic Voice Pathology Detection,” *Journal of Voice*, vol. 38, no. 5, pp. 975-982 ,2024 . [CrossRef] [Google Scholar] [Publisher Link]

[10] Vyom Verma et al., “A Novel Hybrid Model Integrating MFCC and Acoustic Parameters for Voice Disorder Detection,” *Scientific Reports*, vol. 13, no. 1, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Sneha Basak et al., “Challenges and Limitations in Speech Recognition Technology: A Critical Review of Speech Signal Processing Algorithms Tools and Systems,” *CMES-Computer Modeling in Engineering and Sciences*, vol. 135, no. 2, pp. 1053-1089, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Zhaopeng Qian, and Kejing Xiao, “A Survey of Automatic Speech Recognition for Dysarthric Speech,” *Electronics*, vol. 12, no. 20, pp. 1-23, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Hyun-Bum Kim et al., “Classification of Laryngeal Diseases Including Laryngeal Cancer, Benign Mucosal Disease, and Vocal Cord Paralysis by Artificial Intelligence using Voice Analysis,” *Scientific Report*, vol. 14, no. 1, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Yuyang Yan et al., “Optimizing MFCC Parameters for the Automatic Detection of Respiratory Diseases,” *Applied Acoustics*, vol. 228, pp. 1-9, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[15] Mohamed Cherif Amara Korba et al., “Improved Laryngeal Pathology Detection based on Bottleneck Convolutional Networks and MFCC,” *IEEE Access*, vol. 12, pp. 124801-124815, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Tuan D. Pham et al., “Diagnosis of Pathological Speech with Streamlined Features for Long Short-Term Memory Learning,” *Computers in Biology and Medicine*, vol. 170, pp. 1-14, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Tuan D. Pham, “Time-Frequency Time-Space LSTM for Robust Classification of Physiological Signals,” *Scientific Reports*, vol. 11, no. 1, pp. 1-11, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[18] Nuha Qais Abdulmajeed, Belal Al-Khateeb, and Mazin Abed Mohammed, “Voice Pathology Identification System using a Deep Learning Approach based on Unique Feature Selection Sets,” *Expert System*, vol. 42, no. 1, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Rehman, Mujeeb Ur et al., “Voice Disorder Detection using Machine Learning Algorithms: An Application in Speech and Language Pathology,” *Engineering Applications of Artificial Intelligence*, vol. 133, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Mohammed Zakariah, Muna Al-Razgan, and Taha Alfakih, “Pathological Voice Classification using MEEL Features and SVM-TabNet Model,” *Speech Communication*,” vol. 162, 2024. [CrossRef] [Google Scholar] [Publisher Link]