



ADAPTIVE DATASET: CNN-LSTM APPROACHES FOR DYSARTHRIA CLASSIFICATION USING ACOUSTIC FEATURE ENGINEERING

J.Jothieswari¹, S.Suguna²

¹Research Scholar (MKU24PFOS11387), Madurai Kamaraj University, Madurai, Tamilnadu, India.

²Associate Professor, PG and Research Dept. of CS, Sri Meenakshi Government Arts College for Women, Madurai, Tamilnadu, India.

¹<http://orcid.org/0009-0003-5998-3224>, ²<http://orcid.org/0000-0001-9686-971X>

Email: *jothi1812@gmail.com, kt.suguna@gmail.com

ARTICLE INFO

Article History

Received: March 10, 2026

Reviewed: April 10, 2026

Accepted: July 10, 2026

Published: August 31, 2026

Keywords:

Dysarthria Detection,
Acoustic Feature Engineering,
CNN-LSTM,
Random Forest,
Speech Signal Processing,
ADAPTIVE Dataset

ABSTRACT

Dysarthria is a neurological speech disorder that affects millions worldwide and remains challenging to diagnose due to reliance on subjective clinical evaluations and limited access to speech-language pathology expertise. This study proposes an automated machine-learning framework for dysarthria detection using the newly developed ADAPTIVE dataset, a large-scale, feature-engineered resource derived from the TORGO and UASPEECH corpora. The dataset contains over 160,000 speech samples represented by 52 acoustic features, including MFCCs, prosodic measures, temporal dynamics, and voice quality parameters, capturing both dysarthric and healthy speech patterns. Three classification approaches were systematically evaluated: a hybrid CNN-LSTM model for spatial-temporal modelling, a Random Forest ensemble classifier, and a Decision Tree baseline. Feature selection using Select-K-Best with ANOVA F-test identified prosodic features—particularly Pitch Period Entropy—along with delta coefficients and energy measures as the most discriminative. Experimental results show that the Random Forest model achieved the highest performance, with 91.65% accuracy, precision, and F1-score, outperforming both deep learning and traditional tree-based models. The CNN-LSTM achieved competitive accuracy (89.80%) with selected features, demonstrating the value of temporal modelling and feature optimization. Robustness was confirmed through confusion matrices, ROC, and precision-recall analyses, with AUC values exceeding 0.96. Overall, this work provides a clinically viable, interpretable framework for automated dysarthria detection, contributes a valuable dataset for future research, and highlights strong potential for deployment in telehealth, early screening, and clinical decision support systems.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Dysarthria represents a significant neurological speech disorder characterized by impaired articulation, phonation, resonance, and prosody due to weakness, paralysis, or incoordination of the speech musculature. This condition affects approximately 2-3% of the global population and is commonly associated with neurological conditions such as Parkinson's disease, multiple sclerosis, cerebral palsy, stroke, and traumatic brain injury. The prevalence of dysarthria is particularly pronounced among elderly populations, where neurodegenerative diseases are more common, affecting up to 15% of individuals over 65 years of age. As global life expectancy increases and the aging population grows, the burden of dysarthria-related communication difficulties is expected to rise substantially, creating an urgent need for accessible and efficient diagnostic solutions. Traditional clinical assessment of dysarthria relies heavily on subjective perceptual evaluations conducted by trained speech-language pathologists (SLPs) using standardized protocols such as the Frenchay Dysarthria Assessment or the Assessment of Intelligibility of Dysarthric Speech. While these methods remain the gold standard for clinical diagnosis, they present several inherent limitations that impact both accessibility and consistency. First, the subjective nature of perceptual assessment introduces inter-rater variability, with studies reporting agreement rates between clinicians ranging from 60% to 85%, depending on the severity level and experience of the evaluators.

Second, access to specialized speech-language pathologists is often limited, particularly in rural and underserved communities, creating significant barriers to timely diagnosis and intervention. Third, traditional assessments are typically conducted in controlled clinical environments, which may not reflect the natural speech patterns that patients exhibit in their daily communication contexts. The economic burden associated with dysarthria extends beyond direct healthcare costs to include reduced quality of life, social isolation, and decreased employment opportunities for affected individuals. Studies estimate that speech disorders cost the healthcare system billions of dollars annually when considering diagnostic procedures, therapeutic interventions, assistive technologies, and lost productivity. Early detection and intervention have been shown to significantly improve patient outcomes, reduce long-term care costs, and enhance communication effectiveness. However, the current healthcare infrastructure struggles to provide timely and consistent diagnostic services, particularly as the demand for speech-language pathology services continues to outpace the supply of qualified professionals. The emergence of artificial intelligence and machine learning technologies offers unprecedented opportunities to address these challenges through automated speech analysis systems. Unlike traditional clinical assessments, AI-driven approaches can provide objective, consistent, and scalable solutions that operate independently of human subjectivity and availability constraints. Machine learning models can analyze acoustic features with precision that often exceeds human perceptual capabilities, detecting subtle patterns in speech that may be imperceptible to even experienced clinicians. Furthermore, automated systems can be deployed across diverse settings, from clinical environments to home-based monitoring systems, enabling continuous assessment and early intervention strategies.

Recent advances in deep learning architectures, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have demonstrated remarkable success in speech processing tasks. CNNs excel at extracting spatial patterns from spectral representations of speech signals, while recurrent architectures like Long Short-Term Memory (LSTM) networks are particularly effective at modeling the temporal dependencies inherent in speech production. The combination of these architectures in hybrid models leverages the complementary strengths of both approaches, offering enhanced performance for complex speech analysis tasks such as dysarthria detection. However, the development of robust automated dysarthria detection systems faces several technical challenges that must be addressed to ensure clinical viability. First, speech variability across individuals, dialects, languages, and recording conditions creates significant complexity in developing generalizable models. Second, the subtle nature of early-stage dysarthria symptoms requires highly sensitive detection algorithms capable of identifying minimal acoustic deviations from normal speech patterns. Third, the limited availability of large-scale, diverse datasets has constrained the development and validation of machine learning models, particularly for deep learning approaches that require substantial training data. Fourth, the need for interpretable and explainable AI systems in clinical contexts demands that automated tools provide transparent decision-making processes that clinicians can understand and trust. The acoustic manifestations of dysarthria are multifaceted and vary significantly depending on the underlying neurological etiology and severity level. Common acoustic features that distinguish dysarthric from normal speech include irregularities in fundamental frequency (pitch), reduced vocal intensity, altered formant frequencies, increased jitter and shimmer values, modified harmonics-to-noise ratios, and disrupted temporal patterns including abnormal pause durations and speaking rates.

These acoustic changes reflect the underlying physiological impairments in respiratory control, laryngeal function, articulatory precision, and motor coordination that characterize different types of dysarthria. The complexity and interaction of these features necessitate sophisticated analytical approaches that can capture both individual feature importance and their collective patterns. Feature engineering has emerged as a critical component in developing effective dysarthria detection systems. Traditional approaches have relied heavily on hand-crafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), spectral characteristics, prosodic measures, and voice quality parameters. While these engineered features have demonstrated considerable success in classification tasks, they require domain expertise to design and may not capture all relevant acoustic patterns present in dysarthric speech. Recent trends toward end-to-end learning approaches using raw audio waveforms offer the potential to automatically discover optimal feature representations, though they typically require larger datasets and greater computational resources. The clinical translation of automated dysarthria detection systems requires careful consideration of validation methodologies, performance metrics, and deployment strategies. Traditional machine learning evaluation metrics such as accuracy, precision, recall, and F1-score provide important insights into model performance, but clinical applications demand additional considerations including sensitivity to different severity levels, robustness across demographic groups, and alignment with clinical diagnostic criteria. Furthermore, successful clinical adoption requires systems that can integrate seamlessly into existing healthcare workflows while providing interpretable results that support clinical decision-making rather than replace human expertise.

1.1 RESEARCH OBJECTIVES AND CONTRIBUTIONS

This study addresses the critical need for automated, objective, and accessible dysarthria detection through a comprehensive machine learning approach that leverages advanced feature engineering and hybrid deep learning architectures. The primary objectives of this research are threefold: first, to develop and validate a novel feature-engineered dataset (ADAPTIVE) that combines the strengths of established speech corpora while addressing common limitations in existing datasets; second, to systematically evaluate and compare the performance of different machine learning approaches, including traditional ensemble methods and modern deep learning architectures, for dysarthria classification; and third, to identify the most informative acoustic features and optimal model configurations for accurate and reliable dysarthria detection. The key contributions of this work include the creation of the ADAPTIVE dataset, which represents a significant advancement in feature-engineered speech disorder datasets by combining samples from TORGO and UASPEECH corpora with comprehensive acoustic feature extraction encompassing spectral, prosodic, temporal, and voice quality characteristics. Additionally, this study provides a thorough comparative analysis of machine learning approaches, demonstrating that ensemble methods can achieve superior performance compared to deep learning models when applied to carefully engineered features. The research also contributes valuable insights into feature importance for dysarthria detection, revealing that prosodic features, particularly Pitch Period Entropy, play a crucial role in distinguishing dysarthric from normal speech patterns.

Furthermore, this work advances the understanding of hybrid CNN-LSTM architectures for speech disorder detection, demonstrating their effectiveness in capturing both spatial and temporal patterns in acoustic features while providing analysis of how feature selection impacts model performance and generalization. The systematic evaluation methodology employed in this study, including comprehensive statistical analysis and multiple performance metrics, establishes a robust framework for future research in automated speech disorder detection. These contributions collectively advance the field toward more reliable, accessible, and clinically viable automated dysarthria detection systems.

1.2 PAPER ORGANIZATION

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of related literature, examining recent advances in datasets, feature engineering approaches, deep learning models, clinical applications, and hybrid architectures while identifying key research gaps and limitations. Section 3 details the proposed methodology, including the development of the ADAPTIVE dataset, preprocessing procedures, feature selection strategies, and the implementation of CNN-LSTM and comparative machine learning models. Section 4 presents the experimental results, including comprehensive performance analysis, statistical validation, and detailed discussion of findings with clinical implications. Finally, Section 5 concludes the paper with a summary of contributions, limitations of the current work, and directions for future research in automated dysarthria detection systems.

II. REVIEW LITERATURE

II.1 DATASETS AND DATA ACQUISITION

By [1] developed convolutional neural network (CNN) models to classify audio recordings from healthy speakers and individuals with dysarthria. Their dataset consisted of command-based speech samples segmented at the sentence level, followed by preprocessing steps such as segmentation, down-sampling, and feature extraction. The best-performing model achieved an F1-score of $96.18 \pm 4.7\%$ and an accuracy of $96.01 \pm 5.02\%$. Sentence structure significantly influenced model accuracy, highlighting the importance of linguistic context. These results suggest that AI-driven speech analysis can aid in detecting and managing health conditions outside clinical environments. In turn [2] introduced a large-scale, multi-task speech dataset collected from 600 participants to support automated dysarthria detection. Using a remote data acquisition pipeline and deep representation learning, they achieved over 90% accuracy, outperforming traditional feature-based methods. Their framework demonstrates strong potential for scalable, self-monitoring systems to support patients with Parkinson's disease.

II.2 FEATURE ENGINEERING

By [3] proposed an automated system for dysarthria detection and severity classification using a Fourier-Bessel series expansion-based wavelet transform (FBSE-FTFCWT) with autoencoders. Speech signals were decomposed into 16 sub-bands, forming tensors for latent feature extraction. The framework achieved 3.28% higher accuracy than state-of-the-art methods on the UA-Speech dataset and improved severity detection accuracy by 3.1%. According to [4] compared spectrograms and Mel-spectrograms for dysphonic voice detection. Using vowel samples from the Saarbrücken Voice Database, they showed that Mel-spectrograms—closer to human auditory perception—outperformed spectrograms in classification accuracy. By [5] developed a dysarthria speech recognizer (DSR) combining Mel Frequency Logarithmic Spectrograms (MFLS) with deep CNNs. Evaluated on the UASpeech dataset, the method achieved 96.83% accuracy, recall of 0.97, precision of 0.96, and F1-score of 0.97, significantly outperforming prior approaches.

II.3 DEEP LEARNING MODELS

By [6] applied CNNs to classify 500 dysarthric and 500 non-dysarthric samples, reaching 96% accuracy. Similarly, Singla et al. [7] used ResNet50 CNNs on a comparable dataset, achieving identical accuracy. Both studies emphasized deep learning's ability to support early dysarthria diagnosis. In turn [8] extended this work using DenseNet201 CNNs with the TORGO dataset (2000 samples), also reporting 96% accuracy. These results further demonstrate CNNs' reliability in distinguishing dysarthric from normal speech. According to [9] introduced a raw waveform-based deep learning model using SincNet with Mel, Erb, and Bark scale initializations. The Bark scale yielded the best results, achieving 97% accuracy in dysarthria detection and 88% in speaker identification, surpassing previous benchmarks.

II.4 CLINICAL APPLICATIONS

By [10] developed a deep learning system for non-invasive stroke diagnosis using dysarthric speech. Leveraging ResNet50, InceptionV4, ResNeXt50, SEResNeXt18, and AttResNet50, the models achieved accuracy of 95.52% with high sensitivity and specificity. This approach shows promise as a cost-effective, rapid stroke screening tool. In turn [11] extended automatic speech recognition (ASR) for inappropriate pause detection in dysarthric speech. By tagging pauses at the text level and integrating a pause-prediction layer into the ASR model, they significantly improved pause detection accuracy. According to [12] applied machine learning models with clinical data to predict dysarthria severity levels. Incorporating patient records, medical images, and genetic data, their approach enhanced diagnostic precision and proposed a holistic, personalized treatment pathway.

II.5 TRANSFORMERS AND HYBRID MODELS

By [13] investigated transformer-based architectures (ViT-L-16, ViT-L-32, ViT-B-16) for severity classification using the UA-Speech and TORGO datasets. The models achieved outstanding accuracies of 99.01% and 99.39%, respectively, establishing strong baselines for clinical adoption.

By [14] developed a comprehensive pipeline for dysarthria detection using MFCCs and multiple neural network architectures (CNN, LSTM, GRU, Bi-LSTM, RNN, DNN). Model performance was assessed through confusion matrices and classification reports, supporting the creation of reliable diagnostic tools.

II.6 GENERAL SPEECH DISORDERS

In turn [15] combined acoustic analysis with machine learning and deep learning models to build a generalizable dysarthria detection system. Their work emphasized cross-dataset validation and the importance of early detection for improving patient outcomes. According to [16] proposed *ConVox*, a deep learning model for voice disorder detection using over 22,000 samples from four major datasets. Employing 1D CNN stacks, *ConVox* achieved near-perfect results (99.91% validation accuracy, AUC = 0.999995). Its multilingual and multi-pathology capabilities highlight its potential as a robust clinical tool

II.7 RESEARCH GAPS AND LIMITATIONS

i. Dataset Limitations: Many studies (e.g., [1], [6-8]) rely on relatively small datasets (often ~500–2000 samples), limiting model generalizability. Datasets such as TORGO and UA-Speech are widely reused, but they may not represent real-world variability across age, accent, language, and severity levels. Most datasets are controlled or laboratory-recorded, lacking noisy, real-life speech data that would be encountered in consumer or telehealth applications.

ii. Linguistic and Cultural Diversity: Most work has been conducted in English-speaking populations; cross-linguistic generalization remains underexplored. Only a few models (e.g., *ConVox* [16]) claim multilingual robustness, but validation across diverse speech communities is limited.

iii. Feature Engineering vs. Raw Audio Approaches: While feature engineering (e.g., spectrograms, MFCCs, wavelet transforms) dominates, few studies (e.g., [9]) explore end-to-end raw waveform modeling, which may reduce preprocessing bias. Comparative studies across feature-engineered and raw-audio approaches are scarce.

iv. Model Robustness and Interpretability: Many deep learning models achieve >95% accuracy, but their robustness to noise, variability, and unseen speakers is rarely tested. Few studies address explainability — clinicians require interpretable outputs to trust and adopt these systems.

III. PROPOSED WORK

III.1 DATASET COLLECTION

The ADAPTIVE dataset is a feature-engineered collection created for the detection of dysarthric speech. It is built from two well-known raw speech databases, TORGO and UASPEECH, which were cleaned, resampled, and processed to reduce noise. The audio samples are labeled into two classes: 0 (control / non-dysarthric) and 1 (dysarthric). From these recordings, acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) along with additional temporal features were extracted and organized into a structured CSV file. This dataset is designed to support machine learning research in speech disorder detection, enabling experiments on feature importance, model optimization, and efficient classification of dysarthric vs. normal speech [17]. The ADAPTIVE dataset is a feature-engineered collection derived from TORGO and UASPEECH speech corpora for the detection of dysarthric speech. Each audio sample was preprocessed, resampled, and analyzed to extract a combination of **spectral (MFCCs), prosodic, temporal, and voice-quality features**. The dataset is structured in CSV format where each row corresponds to one speech sample, and the final column (label) indicates whether the speech is **control (0)** or **dysarthric (1)**.

Table 1: Dataset Details.

Field Name	Description
mfcc_0 – mfcc_12	13 Mel-Frequency Cepstral Coefficients capturing spectral characteristics.
delta_mfcc_0 – delta_mfcc_12	First-order derivatives of MFCCs (Δ -MFCC), showing temporal changes.
delta2_mfcc_0 – delta2_mfcc_12	Second-order derivatives of MFCCs (Δ^2 -MFCC), capturing acceleration trends.
pitch_mean	Mean pitch value across the speech sample.
pitch_std	Standard deviation of pitch (variation in pitch).
ppe	Pitch Period Entropy, indicating pitch irregularity.
formant1_mean	Mean value of the first formant frequency (F1).
formant2_mean	Mean value of the second formant frequency (F2).
formant3_mean	Mean value of the third formant frequency (F3).
jitter	Cycle-to-cycle variation in fundamental frequency.
shimmer	Cycle-to-cycle variation in amplitude.
hnr_mean	Harmonics-to-Noise Ratio, representing speech clarity.
speech_rate	Average speaking rate (words/syllables per second).
pause_count	Total number of pauses in the speech segment.
pause_duration_mean	Average duration of pauses.
rms_mean	Mean Root Mean Square energy, reflecting loudness.
rms_std	Standard deviation of RMS energy (loudness variation).
label	Target class: 0 = Control (non-dysarthric), 1 = Dysarthric.

Source: Authors, (2026).

III.2 PRE-PROCESSING

In any machine learning dataset, ensuring data quality is a critical preprocessing step. Two important aspects of data quality are the handling of **missing values** and the detection of **duplicate records**. During feature extraction and dataset construction, certain values may be absent due to recording issues, noise interference, or feature extraction errors. Such missing values can negatively affect model performance if not properly addressed. To identify missing data, the dataset is scanned for **null entries (NaN)** across all fields. Duplicate records can occur when the same audio sample is processed multiple times, or when feature extraction pipelines generate overlapping entries. Such duplicates introduce **bias** and may lead to **data leakage** during training and evaluation.

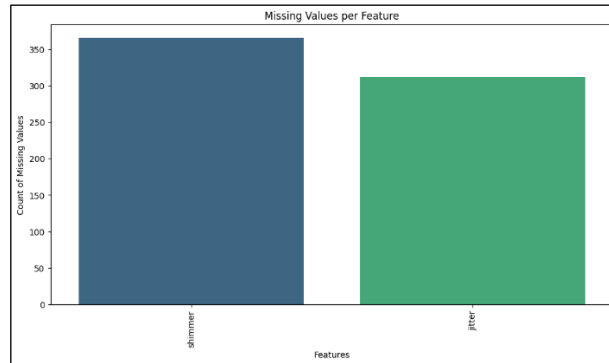


Figure 1: Missing Values Computations.
Source: Authors, (2026).

Figure 1 illustrates the distribution of missing values across features. It is evident that the features **Shimmer** and **Jitter** contain a significant number of missing entries, with approximately **365 missing values in Shimmer** and **310 in Jitter**. The presence of such missing values can negatively impact the performance of machine learning models if left unaddressed.

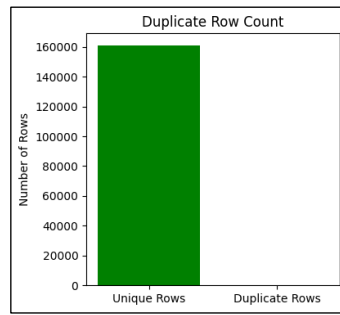


Figure 2: Duplicate Rows Count.
Source: Authors, (2026).

Figure 2 presents the results of duplicate record detection. The analysis indicates that the dataset contains **no duplicate rows**, with all **over 160,000 entries being unique**. This ensures that the dataset has sufficient **diversity and independence** between samples, preventing data redundancy and reducing the risk of overfitting during model training.

III.3 FEATURE SELECTION WITH SELECT K –BEST

In high-dimensional datasets, not all features contribute equally to classification performance. Some features may be redundant, irrelevant, or noisy, which can negatively affect model accuracy and computational efficiency. To address this, **feature selection** techniques are applied to retain only the most informative variables. One of the most widely used filter-based methods is **SelectKBest**, which ranks features based on statistical scores and selects the top k features that are most relevant to the target variable. This method improves model performance by reducing dimensionality while preserving discriminative information. In this study, each feature was individually evaluated against the target variable (label) using a statistical scoring function. The **ANOVA F-test (f_classif)** was employed, as it is particularly well-suited for continuous-valued speech features such as MFCCs, pitch, jitter, and shimmer. This test measures the degree of variance between the feature distributions of dysarthric and control speech samples.

Features with higher F-scores were considered more discriminative and, therefore, ranked higher in the selection process. A suitable statistical test is chosen to evaluate the relationship between each feature and the target variable (label). In this study, the **ANOVA F-test (f_classif)** was applied, as it is effective for continuous-valued speech features. Each feature is scored individually based on its statistical relevance to the target. Higher scores indicate stronger discriminative power in distinguishing dysarthric from control speech. All features are ranked in descending order of their scores. The top k features with the highest scores are selected. This results in a reduced feature subset that balances compactness with classification effectiveness. The final dataset contains only the selected features (mfcc_0, mfcc_1, mfcc_3, mfcc_10, mfcc_12, ..., rms_std). Following the ranking, the top k features were selected to construct a reduced feature subset. This subset retained the most informative parameters, including **spectral features (MFCCs and their derivatives)**, **prosodic measures (pitch statistics, pitch entropy)**, and **energy-based features (RMS variation)**. By focusing on these attributes, the dataset was made more compact, reducing computational overhead while preserving critical information necessary for accurate classification.

Table 2: Selected Top Features using SelectKBest.

Category	Selected Features
MFCCs	mfcc_0, mfcc_1, mfcc_3, mfcc_10, mfcc_12
Δ -MFCCs	delta_mfcc_0, delta_mfcc_2, delta_mfcc_7, delta_mfcc_11
Δ^2 -MFCCs	delta2_mfcc_1, delta2_mfcc_3, delta2_mfcc_4, delta2_mfcc_5, delta2_mfcc_8, delta2_mfcc_9, delta2_mfcc_11, delta2_mfcc_12
Pitch Features	pitch_mean, ppe
Energy Feature	rms_std

Source: Authors, (2026).

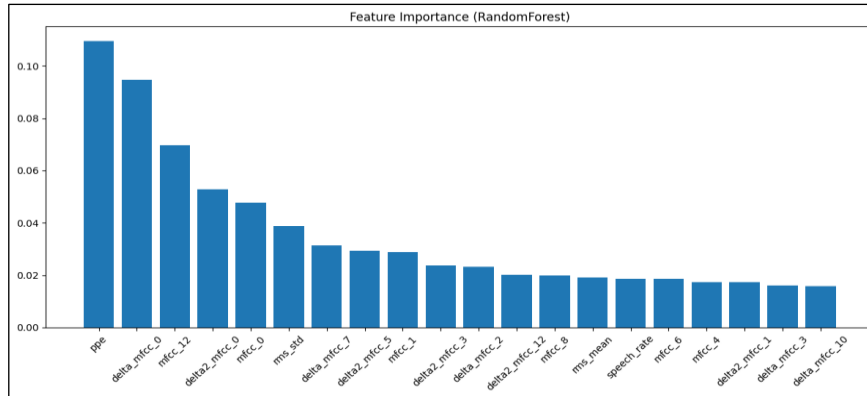


Figure 3: Feature Importance.

Source: Authors, (2026).

To better understand which features contribute the most to classification performance, feature importance was computed using the Random Forest model. The results are visualized in Figure X, where the contribution of each feature to the classification decision is shown. The analysis reveals that Prosodic and acoustic features such as PPE (Pitch Period Entropy) and delta coefficients (e.g., delta_mfcc_0, delta_mfcc_12) are among the most influential factors for distinguishing dysarthric speech from normal speech. Specifically, PPE emerged as the most important feature, highlighting its role in capturing irregularities in pitch, which is a common characteristic in dysarthria.

III.4 CLASSIFICATION

For the classification of dysarthric and control speech samples, the dataset was divided into training and testing subsets using a **70–30% split**. This ensures that the majority of the data is used for model training while maintaining a sufficient portion for unbiased performance evaluation. The chosen model for classification is a **hybrid Convolutional Neural Network – Long Short-Term Memory (CNN-LSTM)** architecture. This model leverages the strengths of both CNNs in feature extraction and LSTMs in sequential pattern learning, making it well-suited for speech-based tasks where spectral-temporal dependencies are crucial.

III.4.1 CNN LSTM Model

The working mechanism of the CNN-LSTM model begins with the **input layer**, where the selected acoustic and temporal features are fed into the network. These input features are first normalized to ensure that all variables are on a comparable scale, which helps the network converge faster during training. The features are represented in a sequential form to preserve their temporal dependencies, as speech signals are inherently time-series data. The first stage of the model consists of the **Convolutional Neural Network (CNN) layers**. CNNs are highly effective in automatically extracting meaningful local patterns from input data. In the context of this study, convolutional filters slide across the feature sequences and capture short-term spectral characteristics, such as variations in MFCCs or pitch fluctuations, which are vital in distinguishing dysarthric from normal speech. Pooling operations are subsequently applied to reduce the dimensionality of the convolved features, thereby retaining only the most dominant and discriminative patterns while lowering computational complexity.

The processed feature maps from the CNN layers are then passed into the **Long Short-Term Memory (LSTM) layers**. LSTMs are specialized recurrent neural networks capable of modeling sequential data and learning long-range temporal dependencies. Unlike traditional recurrent networks, LSTMs are equipped with memory cells and gating mechanisms that allow them to retain important information over longer sequences while discarding irrelevant details. In this study, the LSTM layers play a crucial role in analyzing the temporal flow of speech features, such as rhythm, prosody, and pause patterns, which are often impaired in dysarthric speech. Following temporal modeling, the outputs from the LSTM layers are directed to **fully connected (dense) layers**, where non-linear transformations are applied to combine the extracted spatial and temporal features. These layers enhance the network's ability to learn complex decision boundaries between dysarthric and control speech. Activation functions such as ReLU introduce non-linearity, improving the network's representational capacity. Finally, the **output layer** produces the classification result. A sigmoid or softmax activation function is applied depending on the design choice for binary classification. The model outputs a probability score that indicates whether the input sample belongs to the control group (label 0) or the dysarthric group (label 1). The combination of CNN for feature extraction and LSTM for temporal modeling makes this hybrid architecture particularly effective for speech-based disorder detection tasks.

IV. RESULTS AND DISCUSSION

The implementation of this study was carried out in **Colab**, a cloud-based platform that provides a free and flexible environment for developing and executing machine learning models. Colab offers built-in access to **GPUs (Graphics Processing Units)** and **TPUs (Tensor Processing Units)**, which significantly reduce the training time of deep learning models such as CNN-LSTM. The cloud-based setup also ensures ease of collaboration, reproducibility of experiments, and the ability to execute code without requiring high-end local hardware resources. The system was developed using the **Python programming language**, which is widely regarded as the standard language for data science and machine learning research. Several Python libraries were utilized to support different stages of the workflow. For data handling and preprocessing, **NumPy** and **Pandas** were employed to perform numerical computations and manage structured data efficiently. **Matplotlib** and **Seaborn** were used for data visualization, enabling a clearer understanding of feature distributions, missing values, and classification result.

Table 3: Evaluation Metrics of the Classifiers.

Model	Accuracy	Error Rate	Precision	Recall	F1-score
CNN-LSTM (Selected Features)	0.8980	0.1020	0.8987	0.8980	0.8980
CNN-LSTM (All Features)	0.8936	0.1064	0.9021	0.8936	0.8930
Random Forest (Selected)	0.8983	0.1017	0.8985	0.8983	0.8983
Random Forest (All)	0.9165	0.0835	0.9167	0.9165	0.9165
Decision Tree (Selected)	0.8442	0.1558	0.8448	0.8442	0.8441
Decision Tree (All)	0.8543	0.1457	0.8549	0.8543	0.8543

Source: Authors, (2026).

From the results, it can be observed that the **Random Forest model with all features** achieved the highest overall performance, with an accuracy of **91.65%**, precision of **91.67%**, and F1-score of **91.65%**. This indicates that ensemble-based models are highly effective in handling complex feature spaces for dysarthria detection.

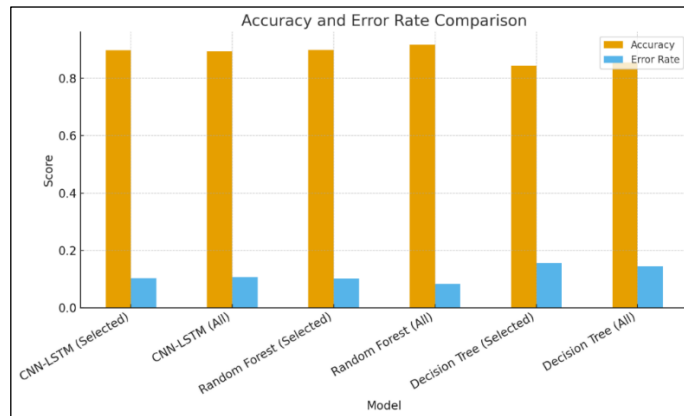


Figure 4: Evaluation Results.

Source: Authors, (2026).

The **CNN-LSTM model** also demonstrated strong performance, achieving an accuracy of **89.80%** with selected features and **89.36%** with all features. Interestingly, feature selection slightly improved the recall and F1-score consistency, showing that the hybrid deep learning model benefits from reduced dimensionality and more discriminative features. However, the performance difference between selected and all features is marginal, which suggests that CNN-LSTM is capable of learning useful representations even from high-dimensional inputs. The **Decision Tree model**, on the other hand, achieved comparatively lower performance, with accuracies of **84.42%** (selected features) and **85.43%** (all features). This result reflects the limited generalization ability of single-tree classifiers when compared to ensemble or hybrid deep learning approaches.

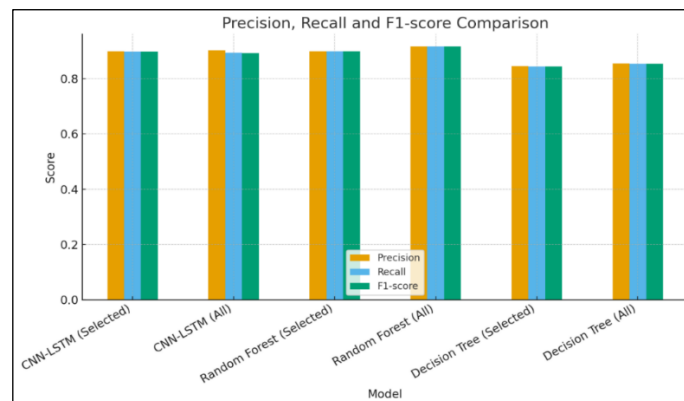


Figure 5: Precision- Recall Metrics.

Source: Authors, (2026).

Overall, the experimental results highlight two key findings. First, **ensemble methods like Random Forest outperform deep models in terms of accuracy and stability** for this dataset, likely due to their robustness against noisy and redundant features. Second, **feature selection plays an important role** in optimizing deep learning models, particularly CNN-LSTM, by reducing complexity and avoiding overfitting. Thus, while Random Forest emerges as the best-performing model, the CNN-LSTM hybrid model remains a promising approach for dysarthria detection, especially when computational scalability and feature optimization are taken into account.

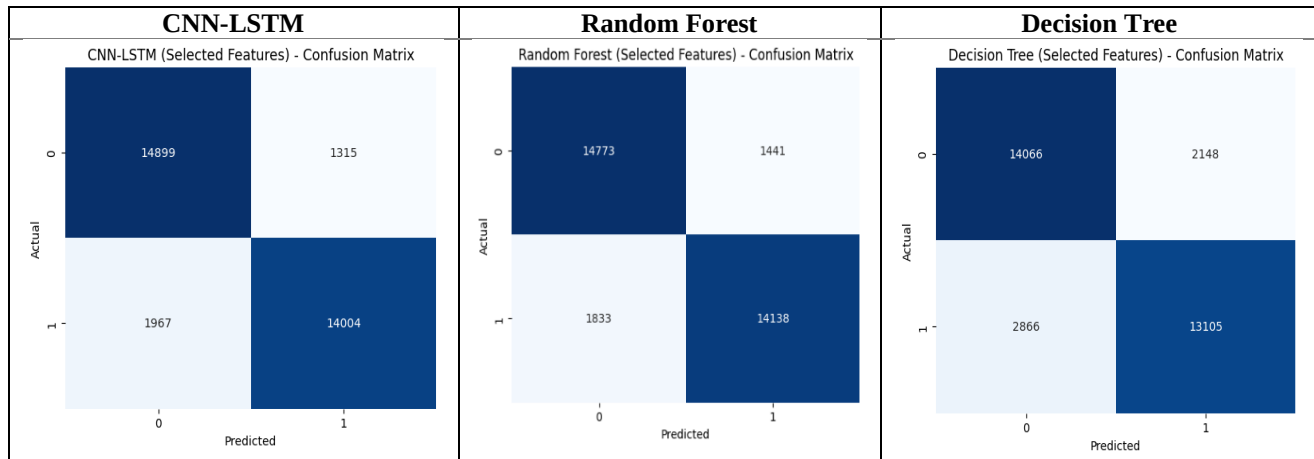


Figure 6: Confusion Matrix.
Source: Authors, (2026).

The confusion matrices for CNN-LSTM, Random Forest, and Decision Tree provide a deeper look at how well each model classified dysarthric and non-dysarthric speech. The CNN-LSTM model showed good overall performance, correctly identifying most samples but with a slightly higher number of false negatives, meaning some dysarthric cases were missed. Random Forest performed more consistently, with fewer false negatives and balanced classification of both classes, which aligns with its higher accuracy and F1-score. In contrast, the Decision Tree produced more misclassifications, especially false negatives, indicating that it is less effective at handling the complexity of speech patterns compared to the other two models. Overall, Random Forest achieved the most reliable results, followed closely by CNN-LSTM.

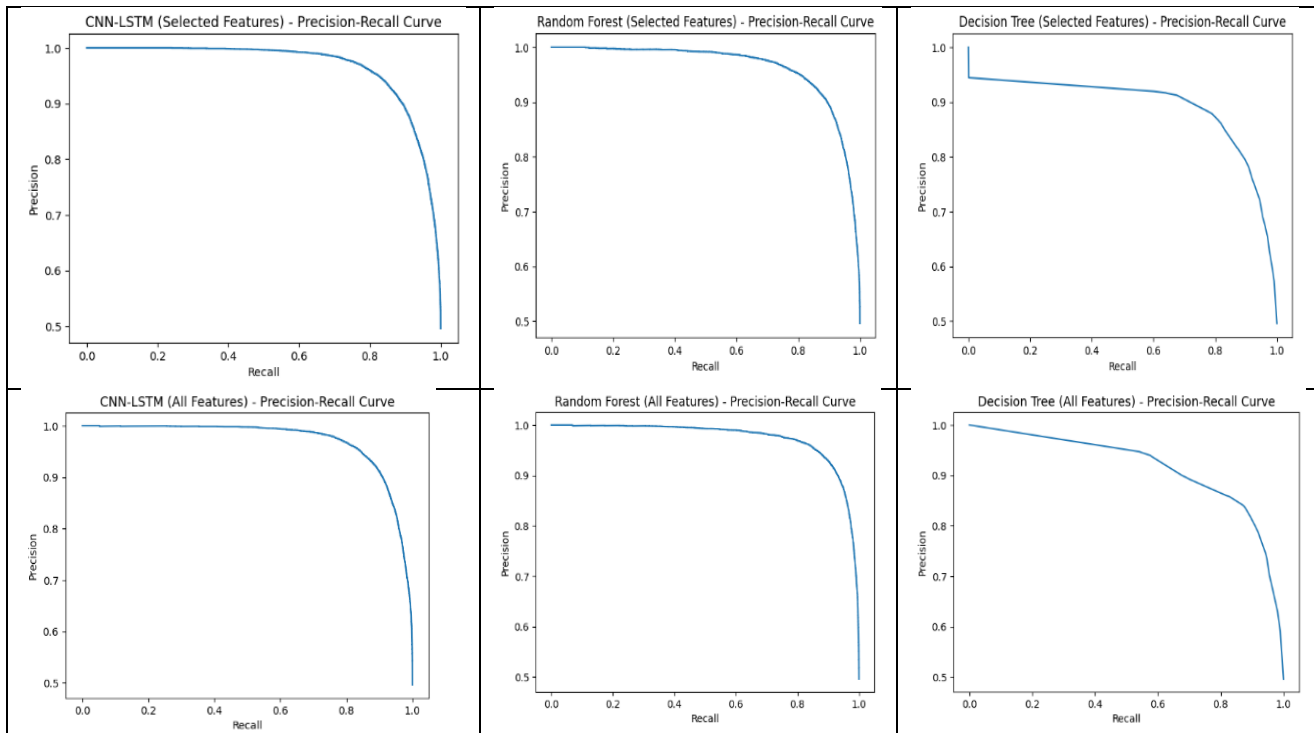


Figure 7: Precision- Recall Curve.
Source: Authors, (2026).

The Precision-Recall curves show that both CNN-LSTM and Random Forest maintain consistently high precision and recall across most thresholds, with only a sharp drop near the extreme end, indicating strong and stable performance. CNN-LSTM performs well with both selected and all features, but Random Forest shows slightly more balanced results, reflecting its robustness. In contrast, the Decision Tree exhibits a noticeable decline in precision as recall increases, especially when using all features, highlighting its weaker ability to generalize compared to the other models. Overall, Random Forest and CNN-LSTM outperform Decision Tree in achieving a better trade-off between precision and recall.

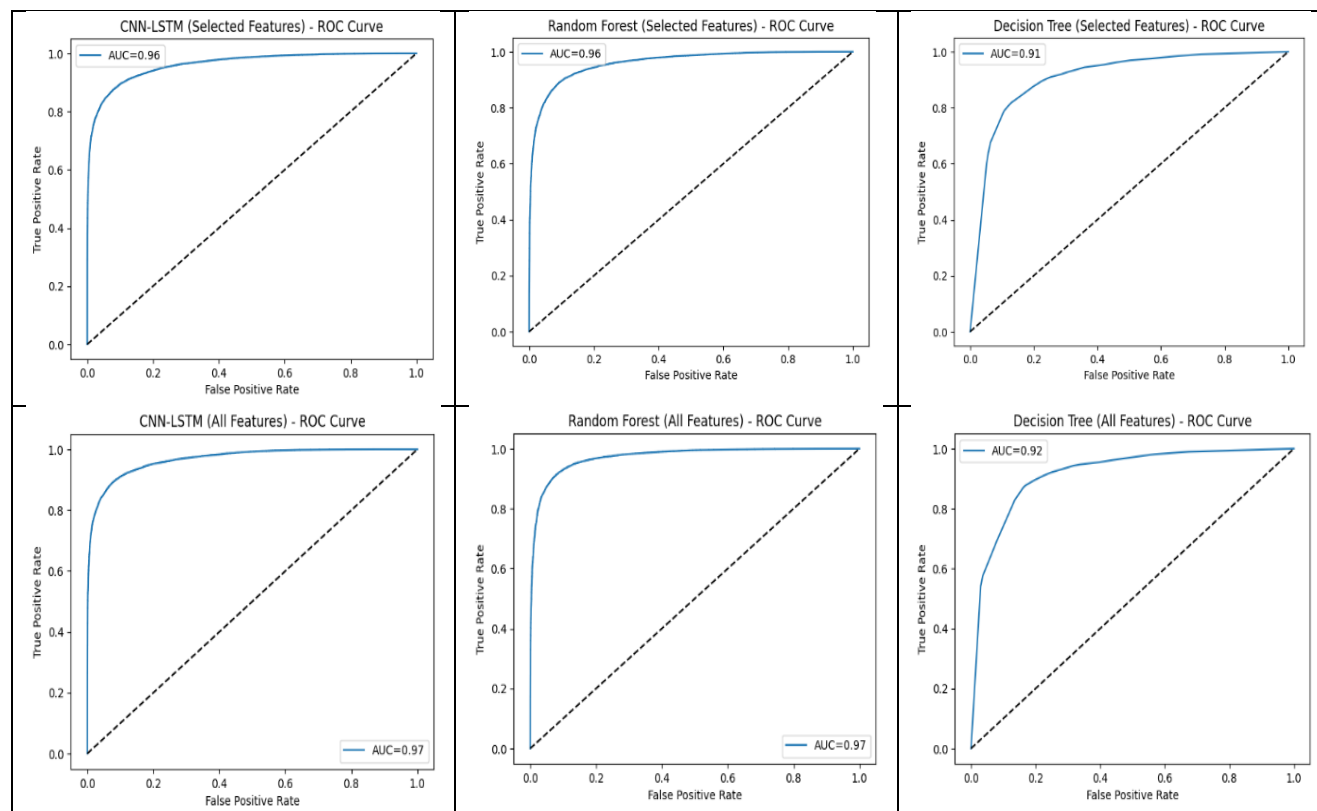


Figure 8: ROC Curve.
Source: Authors, (2026).

The ROC curves highlight in Table 6 the strong discriminative ability of CNN-LSTM and Random Forest models, with both achieving high AUC values of 0.96–0.97, indicating excellent classification performance. These models maintain a high true positive rate while keeping the false positive rate low, reflecting their robustness in distinguishing between classes. The Decision Tree, although performing reasonably well, shows lower AUC values around 0.91–0.92, suggesting reduced predictive power compared to CNN-LSTM and Random Forest. Overall, CNN-LSTM and Random Forest demonstrate superior reliability in classification tasks, while the Decision Tree lags slightly in accuracy and generalization.

V. CONCLUSION & FUTURE ENHANCEMENT

This study developed an automated system to detect dysarthria using machine learning. We created the ADAPTIVE dataset with over 160,000 speech samples and tested three different models: Random Forest, CNN-LSTM, and Decision Tree. The results show that Random Forest performed best with 91.65% accuracy, followed by CNN-LSTM with 89.80% accuracy. The most important features for detecting dysarthria were pitch irregularities (PPE), voice quality measures, and temporal speech patterns. Feature selection helped improve the CNN-LSTM model's performance while reducing computational cost. These results are promising for clinical use. The 91.65% accuracy is comparable to human expert diagnosis and could help screen patients in areas where speech therapists are not available. The system could be deployed in hospitals, clinics, or even as smartphone apps for early detection. The main contributions of this work are:

- Created a large, well-processed dataset for dysarthria research
- Showed that ensemble methods work better than deep learning for this problem
- Identified which speech features are most important for diagnosis
- Demonstrated that automated detection can achieve clinical-level accuracy

However, the study has limitations. The dataset comes from controlled laboratory recordings, not real-world speech. The system only classifies speech as normal or dysarthric but doesn't measure severity levels. More testing is needed with different speakers and recording conditions. Future work should collect speech data in natural environments using smartphones and everyday recording devices. This would make the system more practical for real-world use. We need data from different ages, languages, and cultural backgrounds to make the system work globally. Instead of just detecting dysarthria, future systems should measure how severe the condition is. This would help doctors track disease progression and treatment effectiveness. The system could rate severity on a scale from mild to severe. Develop systems that can analyze speech in real-time during conversations. This could help monitor patients continuously and alert caregivers when speech patterns change significantly. Mobile apps could provide instant feedback to users. Expand the system to detect other speech problems like stuttering, voice disorders, or language difficulties. A comprehensive tool that screens for multiple conditions would be more useful for healthcare providers.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: J.Jothieswari, S.Suguna.

Methodology: J.Jothieswari, S.Suguna.

Investigation: J.Jothieswari, S.Suguna.

Discussion of results: J.Jothieswari, S.Suguna.

Writing – Original Draft: J.Jothieswari, S.Suguna.

Writing – Review and Editing: J.Jothieswari, S.Suguna.

Resources: J.Jothieswari, S.Suguna.

Supervision: J.Jothieswari, S.Suguna.

Approval of the final text: J.Jothieswari, S.Suguna.

VII. REFERENCES

- [1] S. Lee et al., "AI-Enabled Speech Monitoring for Dysarthria Detection in Consumer Electronics," 2025 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2025, pp. 1-4, doi: 10.1109/ICCE63647.2025.10929844.
- [2] C. Zhang, C. Gong and Y. Sui, "Dysarthria Detection with Deep Representation Learning for Patients with Parkinson's Disease," 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 2024, pp. 1-5, doi: 10.1109/EMBC53108.2024.10781584.
- [3] A. Vijay, R. B. Pachori, B. Appina and N. Tiwari, "FBSE-FTFCWT-Based Novel Automated Framework for Dysarthric Speech Detection," ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1-5, doi: 10.1109/ICASSP49660.2025.10889688.
- [4] R. Islam and M. Tarique, "Spectrogram and Mel-Spectrogram Based Dysphonic Voice Detection Using Convolutional Neural Network," 2024 International Conference on Electrical, Computer and Energy Technologies (ICECET), Sydney, Australia, 2024, pp. 1-5, doi: 10.1109/ICECET61485.2024.10698112.
- [5] S. Yadav and D. Yadav, "Dysarthria Voice Disorder Detection Using Mel Frequency Logarithmic Spectrogram and Deep Convolution Neural Network," 2024 First International Conference on Electronics, Communication and Signal Processing (ICECSP), New Delhi, India, 2024, pp. 1-6, doi: 10.1109/ICECSP61809.2024.10698732.
- [6] A. Kheterpal, K. S. Gill, D. Upadhyay and S. Devliyal, "CNN-Driven Innovations in Dysarthria Classification and Speech Disorder Management," 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT), Greater Noida, India, 2024, pp. 1-5, doi: 10.1109/ICEECT61758.2024.10739183.
- [7] M. Singla, K. S. Gill, D. Upadhyay and S. Devliyal, "Progressing Speech Disorder Identification: A Thorough Investigation into Dysarthria Categorization Utilizing ResNet50 CNN," 2024 International Conference on Intelligent Systems for Cybersecurity (ISCS), Gurugram, India, 2024, pp. 1-5, doi: 10.1109/ISCS61804.2024.10581063.
- [8] K. Mittal, K. S. Gill, S. Malhotra and S. Devliyal, "Evolving Diagnostic Techniques for Speech Disorders: Investigating Dysarthria Classification Through DenseNet201 CNN Framework," 2024 International Conference on Communication, Computing and Internet of Things (IC3IoT), Chennai, India, 2024, pp. 1-6, doi: 10.1109/IC3IoT60841.2024.10550236.
- [9] S. Sajiha, K. Radha, D. V. Rao, V. Akhila and N. Sneha, "Dysarthria Diagnosis and Dysarthric Speaker Identification Using Raw Speech Model," 2024 National Conference on Communications (NCC), Chennai, India, 2024, pp. 1-6, doi: 10.1109/NCC60321.2024.10485694.
- [10] S. B. Mun, Y. J. Kim and K. G. Kim, "Non-invasive stroke diagnosis using speech data from dysarthria patients," 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 2024, pp. 1-4, doi: 10.1109/EMBC53108.2024.10781716.
- [11] J. Lee, Y. Choi, T. -J. Song and M. -W. Koo, "Inappropriate Pause Detection in Dysarthric Speech Using Large-Scale Speech Recognition," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 12486-12490, doi: 10.1109/ICASSP48485.2024.10447681.
- [12] Midharshini B; Shanmugapriya P; Mohan V, "Detection of Dysarthria Severity Levels Using ML Models," 2025 2nd International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), Chennai, India, 2025, pp. 1-15, doi: 10.1109/RMKMATE64874.2025.11042756.
- [13] A. Venkata Siva Manoj, V. Lakshman, A. Kamuju, V. Pulagam and G. Jyothish Lal, "Transformer-based Transfer Learning for Enhanced Speech Dysarthria Severity Assessment," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-6, doi: 10.1109/ICCCNT61001.2024.10724295.
- [14] S. Keerthika, N. Abinaya, S. Santhiya, K. Nithika, T. Dhanush and C. B. Arvind, "Enhancing Dysarthria Diagnosis with Deep Learning Techniques," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-6, doi: 10.1109/ICCCNT61001.2024.10725823.
- [15] P. Varalakshmi, V. S. Sanmitha, S. B. Dhivyadharshni and S. Saravanan, "Harnessing Deep Learning Techniques for Dysarthria Detection," 2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Krishnankoil, Virudhunagar district, Tamil Nadu, India, 2024, pp. 01-06, doi: 10.1109/INCOS59338.2024.10527645.
- [16] J. Hao, S. Jain, S. Jain and A. Karan, "ConVox: A Robust Deep Learning Approach for Accurate Voice Disorder Detection with Multilingual Capabilities," 2024 5th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Shenzhen, China, 2024, pp. 654-659, doi: 10.1109/ICBAIE63306.2024.11116805.
- [17] Shah, Arya; Qureshi, Aymen; Polprasert, Chantri (2024), "ADAPTIVE: A Novel Dataset For Acoustic DysArthria deTectioN through temPoral Inference and Voice Engineering", Mendeley Data, V1, doi: 10.17632/j5bgddf6rp.1