Grenze International Journal of Engineering and Technology, June Issue

Adaptive Feature based Explainable Ensemble Model for Dysarthria Speech Detection using OpenSMILE Acoustic Features and Multi-Level Feature Selection

Pramod G G[1], Nischaykumar Hegde[2], Sanjana V[3], J Sheethal[4] and Bhavitha[5]
[1-5]Canara Engineering College, Benjanapadavu, Karnataka, India
Email: Pramodggkalasa8338@gmail.com, sanjanavinu87@gmail.com, sheethalshivani26@gmail.com, bhavithagolthila14@gmail.com, nischayhegde@canaraengineering.in

***Abstract*— Speech-based effective communication is essential to human contact and has a significant impact on an individual's quality of life. Dysarthria is a motor speech disorder that impedes natural expression and intelligibility by altering voice acoustics. However the conventional system often face the challenges due to noisy and redundant features. To address this challenge, the study proposed the Adaptive feature based Explainable ensemble mod-el for the automated detection of dysarthria. Subsequently many acoustic features of TORGO and UA-Speech data are extracted with the core feature extraction of module of OpenSMILE utilized in this study. The extracted features are selected with the Adaptive feature selection process to chosen the most informative features. The selected features are classified with the ensemble learning stacking model of Random forest, XG Boost and SVM. Finally the core model interprets with the global and local levels with hybrid developed model of SHAP and LIME.**

***Index Terms*— Dysarthria, speech detection, SHAP, LIME, RF, SVM.**

I. INTRODUCTION

Effective speech-based communication is crucial to human interaction and greatly affects a person quality of life. However, some people have neurological issues that cause them to have trouble speaking, which can manifest as motor speech issues like dysarthria [1-4]. Different subjects experience this injury in different ways, which can result in a variety of dysarthria. The production, rhythm, pitch, rate, volume, quality, and duration of speech can all be impacted by speech disorders. People with dysarthria include those who have a primary speech issue or those who suffer from a speech disorder due to a condition like Parkinson disease (PD) or amyotrophic lateral sclerosis (ALS). Dysarthria reduces speech intelligibility, thus affecting social interaction and quality of life [5-7]. Due to advancements in signal processing, machine learning (ML), and deep learning (DL), automatic detection and severity level classification of dysarthria from speech have drawn a lot of attention in recent years [8,9]. The two primary stages of the so-called pipeline system feature extraction and classification correspond to the techniques used in dysarthria identification [10, 11]. These systems are trained under supervision with gathered speech data and labels (dysarthria vs. healthy). Speech-language pathologists; examinations provide the (binary) labels. Numerous studies have looked into different feature extraction techniques that highlight the key elements of producing dysarthria speech [12, 13].

Grenze ID: 01.GIJET.12.2.77_23

Even though, there are numerous model developed in the past studies. Still there is the need to development of adaptive model due to presence of noisy and redundant features. To overcome these limitation, we proposed Adaptive feature based Explainable ensemble model to detect the dysarthria disease. The main contributions are,

- To reduce the noise and normalize the raw data, the technique of wiener filter and normalization was processed in this study. The obtained pre-processed output are further fed into the feature extractor of OpenSMILE signal processing toolkit.
- From the extracted features, with the Tri-fused feature selection technique of Mutual Information, Recursive Feature Elimination, and Lasso Regression to identify the most discriminative and non-redundant feature.
- To classify the selected features and interpret the global and local feature contribution, the proposed framework develop the explainable ensemble model by the integration of ensemble stacking classifier with the hybrid SHAP-LIME XAI.

II. Literature Review

Dysarthria, as a motor-speech disorder induced by a neuromuscular impairment has been extensively research in acoustic and machine learning to create objective and automated diagnostic tools. The recent research has investigated numerous signal- processing and learning methods to describe articulatory and phonatory deviations. The important early signal-based techniques were based on hand-crafted acoustic parameters. The literature review shows the current developments on the field of detecting dysarthria and its severity based on artificial intelligence and self-supervised learning methods. In [1], the self-supervised models used include wav2vec2-BASE and distilALHuBERT on TORGO and UA-Speech, which produce 99% detection and 95% severity classification results, which are better than baseline MFCC-based models. The study of [2] examines the application of acoustic biomarkers such as harmonics-to-noise ratio, jitter, shimmer and F0 among ALS patients, and it reaches a high accuracy rate with decision tree classifiers as 86.6, which implies that these tools are reliable in the early diagnosis. In [3] and [4], Farhad Javanmardi and Amina Hamza use RNN, LSTM and SVM models with 37 acoustic features using the Nemours corpus, RNN had 99.69 percent detection and SVM 97 percent severity classification, which prove that the combination of MFCC, prosodic, and spectral features will be important. The approach suggested in [5] by Yan Xiong et al. presents a Task-specific Gradient Projection Multi-Task Learning (TGP-MTL), which yields an increase in the accuracy of dysarthria recognition by 3.18 per cent and t-SNE visualization, which promotes a better class separability. As Saraswati Patil et al. [6] use CNN on TORGO with MFCC features, with an accuracy of 96 percent, the ability of CNN to be used in clinical diagnosis in real-time is confirmed. In the review article [7] by Shaik Sajiha et al. AI methods such as CNN, RNN, and SVM are reviewed with more than 95 percent accuracy and because of the value of multi-feature fusion to strong detection. The method of [8] by M. Mahendran et al. uses CNNs with MFCCs, spectral centroid, roll-off and zero-crossing rate to achieve 93.87% accuracy demonstrating the robustness of the method in noisy conditions. Lastly, a Conformer- based ASR model with CNNs and Transformers, suggested by Mark Johnson et al. in [9], after domain adaptation, outperforms RNN-Transducer models in recognition accuracy and word error rates. Together, these articles indicate that advanced deep learning and self-supervised techniques are useful in improving the detection and classification of dysarthria speech by a significant margin, which opens the door to efficient and automated clinical assessment systems. Although early studies were based on MFCCs, jitter, shimmer, and prosodic cues to characterize them, modern models are more likely to incorporate feature selection, representation learning, and explainable attention mechanisms. Overall, the analyzed studies show that to achieve successful dysarthria detection, there should be a trade-off between the acoustic interpretability, model flexibility, and computational efficiency. Although this has been achieved, there are still a few issues to consider: small and skewed data sets like TORGO or UA Speech restrict generalization; inter-speaker discrepancy creates bias; and deep networks do not always provide transparency to clinical use. These loopholes are the causes that encourage the current research to design a machine-learning system that integrates Advance feature extraction technique and the conventional feature selection (Mutual Information, RFE, Lasso regression).The integrated hybrid designs can provide both data-efficiency and clinical interpretability that bridges the classical and advance approaches for dysarthria detection.

III. Methodology

The proposed methodology integrates the denoising, acoustic feature extraction, selection and classification with the utilization various baseline models. The raw data pre-processed to get the denoised and normalized signal. From this, with the utilization OpenSMILE tool kit core acoustic features are extracted. Among the extracted

feature, informative features are selected and classified with the ensemble classifier fused model. Finally to ensure the reliability interpreted with the hybrid XAI model illustrated in Figure 1.

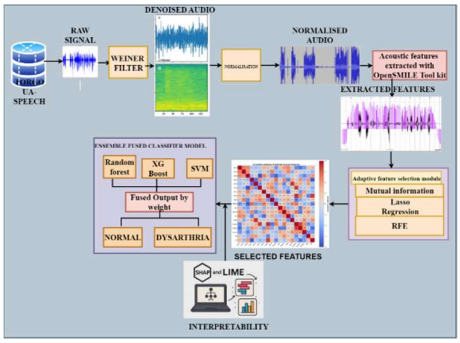

Figure 1. Architecture of the proposed model

A. Dataset

In TORGO dataset, eight patients (three females and five men) with cerebral palsy or amyotrophic lateral sclerosis (ALS) were included in this database, along with seven healthy control speakers (three females and four males). The age range of the patients is sixteen to fifty years old. Three kinds of speech signals—non-words, words, and sentences—make up the database. The non-words category includes five to ten repeats of vowels spoken at high and low pitches for five seconds. A head-mounted microphone, an array microphone, and a 16 kHz sample were used to capture speech simultaneously. The array microphone recordings were used for the current investigation. The other dataset of UA-speech dataset is one of the widely used dysarthria speech corpora created by the University of Illinois. It consists of the 19 speakers recording diagnosed with dysarthria caused by cerebral palsy. Every speaker tends to produce 765 words with the inclusion of digit, vocabulary, commands and alphanumeric items. The dataset also contains 13 healthy control speakers that allows the comparison between the impaired and typical speech.

B. Preprocessing

The input raw data was often accompanied with the unwanted noise due to microphone variability and speaker movement which reduces the quality of the features and accuracy during the detection. Therefore the denoising technique of wiener filter was employed in this study to remove the redundant noise. Followed by the process of denoising, z-score normalization also performed in order to standardize the amplitude. This process ensures the uniform scaling to mitigate the variation of interspeaker and record variation. The z-score normalization can be represented as Subsequent paragraphs, however, are indented.

$$zi = \frac{z-\mu}{\sigma} \tag{1}$$

C. Feature Extraction

After pre-processing, the data fed into the process of feature extraction where the most relevant acoustic features of each samples are extracted. The extracted features are mainly includes of spectral, prosodic and phonatory characteristics of pre-processed subset. Open SMILE is a comprehensive toolkit were used to extract the various acoustic features by integrating the several types of classical and advanced signal processing techniques. It represents the framework that comprises of multiple well established feature extraction methods into a one unified system. The important features extracted from OpenSMILE includes of prosodic, spectral, voice quality, temporal as well as their statistical functional. Prosodic features of fundamental frequency, energy, duration are extracted using techniques of autocorrelation, cepstral analysis and energy analysis. By the utilization of FFT based spectral analysis, the Spectral based features of MFCC, spectral centroid, spectral flux and band width are obtained. Finally the other set of features of Jitter, shimmer and HNR are also computed as cycle to cycle variations to capture the subtle voice irregularity on the frequency and amplitude. Addition to that for capturing information about rhythm and extraction rate, the temporal features Like ZCR and pause duration also measured. After extracting all the features, statistical function were applied to convert the variable length signals to fixed-length suitable vectors.

D. Feature Selection

After the extraction of features, the phase of feature selection was processed to extract the refined and optimized feature subset for the detection of dysarthria. The main purpose of projecting this phase is used to identify and retain the most relevant features. That informative features significantly contributes to distinguish the severity level from the normal level of dysarthria speech. Through this process irrelevant features are deteriorated, which reduces the noise as well as the computational load of the process. In this study, the phase of feature selection plays the pivotal role in improving model accuracy and enhances the interpretability. In this study three feature selection techniques of Mutual information, Recursive Feature Elimination (RFE) and Lasso Regression was employed. Subsequent paragraphs, however, are indented.

Mutual information is statistical based technique that quantifies the dependency between the extracted features and target label with the selected features. This techniques measures the amount of informative feature and the target label. It quantifies with the higher values of score obtained as relevant

$$I(xi, y) = \sum xi \in Xi \ \sum y \in Y \ P(xi, y) \log \frac{P(xi,y)}{P(xi),P(y)} \quad (2)$$

Recursive filter elimination, wrapper based feature selection method that recursively removes the redundant features on the basis of performance weight on the model. Train with the all features, from which importance weights computed and ranked the features based on the weights. The whole process will repeated until the optimal subset reached.

$$Xi = |w_i|^2 \quad (3)$$

As Lasso Regression method is highly efficient in high dimensional feature spaces. It performs the embedded feature selection process by integrating the penalty term of L1 which forces the co-efficient of the most redundant features to zero. The objective function is

$$min\alpha(\frac{1}{2N}\sum_{i=1}^{N}(xi - \alpha_0 - \sum_{k=1}^{q}\alpha_k yik))^2 + \lambda \sum_{k=1}^{q}|\alpha k|) \quad (4)$$

All the selected features through the concatenation combines the various multiple feature vectors into a single comprehensive representation for the better classification that captures the complementary speech features.

E. Adaptive ensemble stacking classifier

The selected features are fed forward into the classifier of dysarthria detection framework. In this stage, various supervised learning model are adopted to classify the selected features as dysarthic or normal. To enhance the accuracy and robustness of the detection system, an ensemble based classifier was developed in this study. The learned representations of SVM, XG boost and Random forest classifier within the fusion framework. The important ideology of this study is to integrate the multiple classifiers that observes the various aspect of data and combine their outputs to achieve a most reliable prediction

In this study, the SVM classifier constructs an optimal separating hyper plane which maximizes the spaces between the binary classes of dysarthria and normal speech.

$$fSVM\ (x) = sign\ (w^s x + A) \quad (5)$$

The model employs the gradient boosting mechanism which lowers the classification error by training multiple learners iteratively that combines them to construct a strong predictive model. Its function represented as,

$$fXGB\ (x) = \sum_{m=1}^{m} \gamma mhm(x) \quad (6)$$

Random forest classifier constructs the multiple independent decision trees using random subset of features and average its prediction to deteriorate overfitting. It can be represented as,

$$fRF\ (x) = \frac{1}{t}\sum_{m=1}^{m} hm(x) \quad (7)$$

The probability score produced by each classifier indicate the likelihood of selected features belongs to the class of dysarthria. To obtain the final decision made by the classifier, a fusion strategy was applied that combine the output of all classifier. In this study, a weighted score-level averaging was adopted to balance the individual performance to preserve interpretability. The final ensemble score computed as

$$S_f = \frac{\sum_{k=1}^{3} w_k S_k}{\sum_{k=1}^{3} w_k} \quad (8)$$

Based on the threshold based rule, final decision is determined. For balanced dataset, threshold set as 0.5

F. Interpretability

As an obtained output from the ensemble model, the hybrid SHAP-LIME models interprets with the decision making process by contributing the feature at global and local levels. SHAP is the game theoretic approach which offers the each features contribution on the model. This module provides the global interpretability by assigning each feature with the value of SHAP. The Lime approach focuses on the local interpretability as it explains the model behaviour for the specific instances thus by fitting over the model at a particular point. For the prediction based, LIME learns its approximation from the obtained loss value

The integration of global and local interpretability proceed under the hybrid SHAP-LIME integration. This hybrid approach leverages the global interpretability of each features contribution to the output and local interpretability of lime explains the prediction locally by simple surrogate function. A hybrid scores are defined as

$$I = \propto \emptyset_k + (1-\propto) w_k \qquad (9)$$

IV. Results and Discussion

The experimental analysis was conducted to detect the effectiveness of the proposed system. All the experiments are performed with TORGO and UA-speech dataset which includes of dysarthria and normal speech samples. Various performance metrics such as Accuracy, precision, recall and F1-scores are calculated to assess the reliability of the model. The following section present the derived results and comparative analysis of the study.

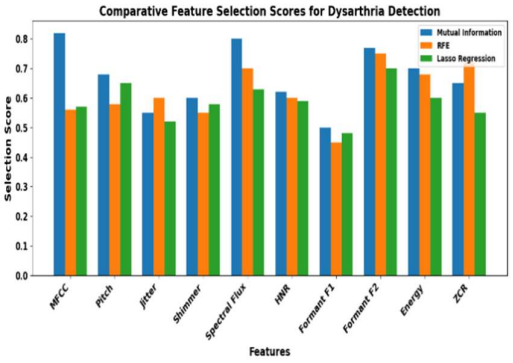

Figure 2. Feature importance scores

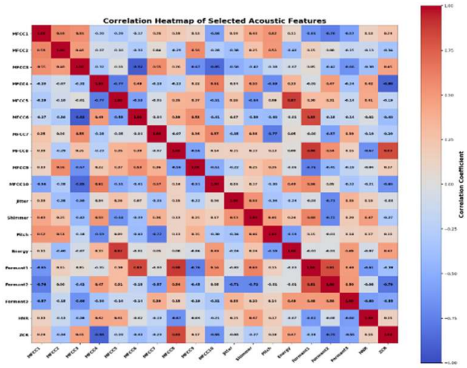

Figure 3. Heat map of selected feature

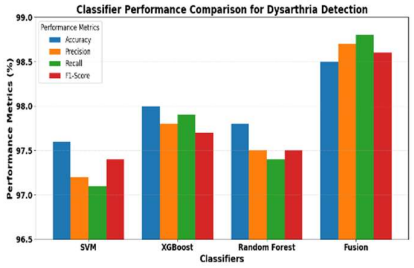

Figure 4. Comparative Analysis

The comparative analysis of feature selection process of mutual information, RFE and Lasso regression was depicted in the figure 2 and 3. While the MI achieves the highest feature relevance scores indicates its efficiency to capture the non-linear dependencies. Another technique of RFE performs well by eliminating the less important features by assuming weights. While the Lasso Regression technique exhibited the better scores due to the sparsity constraint. As core all the features fusion gives good linear representation vector with more discriminant features. The proposed ensemble classifier was evaluated using the metrics of accuracy, precision, recall and F1-scores. As shown in figure 4, comparative study was conducted with the individual classifiers of SVM, Random forest, XG Boost and proposed model. Compared with the other baseline models the proposed model shows the superior performance with superior accuracy of 98.5%, precision 97.9%, recall and precision of

98.0%. This achievement in performance metrics highlights synergetic features achieved with traditional combined ML features. This ensemble method mitigates the overfitting and capture the informative features across classifiers leads to the improvement in detection and interpretability of this study.

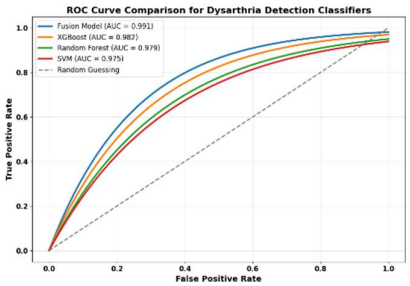

Figure. 5. ROC Analysis

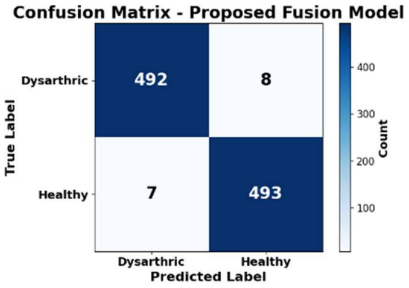

Figure 6. Confusion matrix

To evaluate the detecting capability of proposed model, the confusion matrix is depicted in figure 6. It attains the true positive and true negative rate of 492 and 493 while false negative and false positive as 8 and 7 respectively. The ROC Curve was employed to find the discriminative ability present in the proposed classifier in the detection of dysarthria. As illustrated in Figure, all the baseline models demonstrated the values of Area under curve exceed than 97. The SVM achieved AUC of 0.975while the other models of XG Boost and Random forest achieved the AUC values of 98.2% and 97.9% respectively. As the proposed models achieved the best AUC score of 99.1% among them which indicated the best sensitivity and specificity rate compared with the baseline classifier. Thus the best achieved scores intimates that the robustness, reliability as well as clinical applicability of the model for automated dysarthria detection.

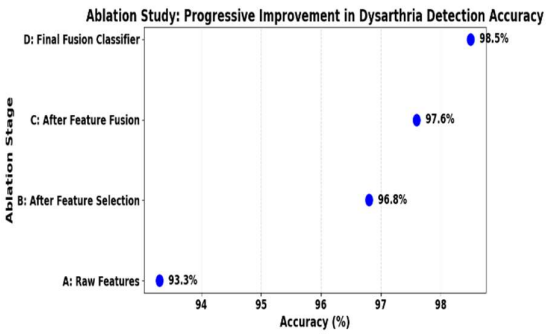

Figure 7. Ablation study

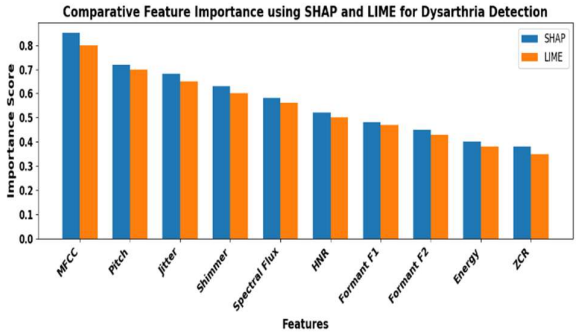

Figure 8. Interpretability of data

To systemically evaluate the contribution of each models integrated, the ablation study was performed across the four stages depicted in Figure 7. At the initial stage, the raw input data with the baseline models attained the accuracy of 93.2% that establishes the challenges in the detection mechanism. At the second stage after application of feature extraction, i.e. with extracted features it achieves the accuracy of 96.8%. While after completion of feature selection, the model attained the accuracy level of 97.6%. Finally the core proposed model achieves the accuracy of 98.5%. The attained accuracy clearly visualise the contribution of feature optimization and classifier fusion of the develop detection system. Figure 8 visualizes the interpretability of explainable model. By integrating both the methods the system achieves the balanced data of acoustic features which influence the feature globally as well as locally. The most influential acoustic parameters contributes for the dysarthria detection. Thereby the proposed model ensures the role of explainable AI in the clinical speech assessment domain. Subsequent paragraphs, however, are indented. The proposed explainable ensemble model with adaptive feature-based is shown to be performing better in terms of dysarthria detection, with an accuracy of 98.5 and an ROC-AUC of 99.1 and outperforms single SVM, Random Forest, and XGBoost classifier. The study of ablation establishes that optimization of feature extraction, tri-fused feature selection and stacking of ensembles all help in performance enhancement. Moreover, SHAP and LIME integration increases the transparency of the models because it gives significant global and local readings of acoustic characteristics, which promotes clinical applicability.

V. CONCLUSION

This study developed the efficient framework in the detection of dysarthria with the optimised feature extraction, hybrid feature selection and ensemble classification model. The past existing study shows that the baseline models of SVM and Random Forest typically achieve accuracies around 95%. While the advanced XG Boost model lift up he accuracy to 97% by capturing the complex speech pattern. It highlights that the proposed approaches outperforms the every classifier by achieving 98.5% in the performance metrics. The proposed model attained performance of 98.5% accuracy shows that the robustness and reliability of the framework in the dysarthria detection. As well as the Explainability also enhanced through the hybrid based SHAP-LIME interpretation model. In future this framework may be extended to access various multilingual dataset with deep learning framework and real-time monitoring detection environment. Additionally, real-time clinical deployment and robustness evaluation under noisy and spontaneous speech conditions will be explored.

REFERENCES

[1] B Sanjay, Priyadharshini Mk, P Vijayalakshmi, and T Nagarajan. Severity classification and dysarthric speech detection using self-supervised representations. In Proceedings of the 21st International Conference on Natural Language Processing (ICON), pages 621–628, 2024.

[2] Raffaele Dubbioso, Myriam Spisto, Laura Verde, Valentina Virginia Iuzzolino, Gianmaria Senerchia, Giuseppe De Pietro, Ivanoe De Falco, and Giovanna Sannino. Precision medicine in als: Identification of new acoustic markers for dysarthria severity assessment. Biomedical Signal Processing and Control, 89:105706, 2024.

[3] Farhad Javanmardi, Sudarsana Reddy Kadiri, and Paavo Alku. Pre-trained models for detection and severity level classification of dysarthria from speech. Speech Communication, 158:103047, 2024.

[4] Amina Hamza, Djamel Addou, and Hamza Kheddar. Machine learning approaches for automated detection and classification of dysarthria severity. In 2023 2nd International Conference on Electronics, Energy and Measurement (IC2EM), volume 1, pages 1–6. IEEE, 2023.

[5] Yan Xiong, Visar Berisha, Julie Liss, and Chaitali Chakrabarti. Improving speech-based dysarthria detection using multi-task learning with gradient projection. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pages 902–906, 2024.

[6] Saraswati Patil, Shruti Borude, Chanchal Budhwani, Ajinkya Bhandare, and Prathamesh Bhujbal. Deep learning based dysarthria detection: A comprehensive approach. In 2024 Fourth International Conference on Multimedia Processing, Communication & Information Technology (MPCIT), pages 262–266. IEEE, 2024.

[7] Afnan Al-Ali, Somaya Al-Maadeed, Moutaz Saleh, Rani Chinnappa Naidu, Zachariah C Alex, Prakash Ramachandran, Rajeev Khoodeeram, and Rajesh Kumar. The detection of dysarthria severity levels using ai models: A review. IEEE Access, 2024.

[8] M Mahendran, R Visalakshi, and S Balaji. Dysarthria detection using convolution neural network. Measurement: Sensors, 30:100913, 2023.

[9] Qianli Wang, Zihan Zhong, Satwinder Singh, Clarion Mendes, Mark Hasegawa-Johnson, Waleed Abdulla, and Seyed Reza Shahamiri. Dysarthric speech conformer: Adaptation for sequence-to-sequence dysarthric speech recognition. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2025.

[10] Hassan, E., Saber, A., Abd El-Hafeez, T., Medhat, T., & Shams, M. Y. (2025). Enhanced dysarthria detection in cerebral palsy and ALS patients using WaveNet and CNN-BiLSTM models: A comparative study with model interpretability. Biomedical Signal Processing and Control, 110, 108128.

[11] Jothieswari, J., & Suguna, S. (2026). Dysarthria Speech Disorder Detection: A Recent Review. In International Conference on Hybrid Intelligence: Theories and Applications (pp. 173-187). Springer, Singapore.

[12] Zhang, Z., Wang, T., Hu, Z., Yang, L. Z., & Li, H. (2025). Multivariate Time Series Approach Integrating Cross-Temporal and Cross-Channel Attention for Dysarthria Detection From Speech. Neurocomputing, 130708.

[13] Lee, S., Lee, S., Park, G., Won, D. O., Kim, C. H., Lee, J. J., & cheol Jeong, I. (2025, January). AI-Enabled Speech Monitoring for Dysarthria Detection in Consumer Electronics. In 2025 IEEE International Conference on Consumer Electronics (ICCE) (pp. 1-4). IEEE.