Clear Tic-AI: Detection of Dysarthria and its Severity Analysis

Huma Sheraz, Iqra Ashraf, Sidra Ashraf, Muhammad Zain, Babar Nawaz
National University of Modern Languages (NUML).
*Correspondence: huma.sheraz@numl.edu.pk

Citation | Sheraz. H, Ashraf. I, Ashraf. S, Zain. M, Nawaz. B, "Clear Tic-AI: Detection of Dysarthria and its Severity Analysis", IJIST, Vol. 7, Issue. 4 pp 3018-3032, December 2025

Received | November 02, 2025 **Revised |** November 18, 2025 **Accepted |** November 23, 2025 **Published |** December 01, 2025.

Dysarthria and other motor speech disorders result from abnormalities in the neural or muscular processes that actually control speech production; conversely, these disorders affect the strength, coordination, and tone of the vocal muscles that ultimately produce less intelligible speech. Because dysarthria can range from moderate distortion of articulation to severe impairment of speech, early and accurate assessment is critical. The paper proposes Clearitic AI, an automated speech analysis platform that leverages artificial intelligence to diagnose vocal disorders. It fuses Wav2Vec2 with traditional acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch estimation, and spectral descriptors. Abnormal voice classification and its severity on a framework with a sequential neural network architecture are proposed. Extensive testing of the system is performed using 10,000 recordings of voice samples from the TORGO dataset and the Mozilla Common Voice dataset. Experimental results demonstrate that the proposed model achieves a classification accuracy of 94.2% (±1.3), an F1-score of 0.943, and an Area Under the Curve (AUC) of 0.987 on the test set, thereby establishing the effectiveness of this framework for dysarthric speech detection applications.

Keywords: Dysarthria, Speech Production, Sequential Neural Network.

Introduction:

Dysarthria is a term that describes speech disorders due to improper functioning of nerves and muscles that coordinate speech. Dysarthria means people find it hard to control and coordinate the rate, duration, and intensity of speech movement to speak; hence, their speech could be difficult for people to understand. Consider a situation where lip and tongue movement is improper; hence, words get distorted: "tip" is pronounced as "sip," "hip," and "sieve"; "beach" could be "eats"; and "decide" could be "sigh" and "say." In cases where vocal organs, referred to as the larynx, are affected, the quality, tone, and intensity of speech are altered. A speech disorder could be characterized by a loss of dynamic variation in terms of loudness, tone, and speech rhythms and inappropriate variations in terms of mentioned parameters above and could coexist with poorly controlled breathing efforts that could coexist with reduced breathing support that limits speech output that could be short and shallow and could be lacking exhalation support needed for speech output and could be characterized by a loss in speech output that involves a soft palate disorder that could result in speech that appears overly nasal.

A speech disorder caused by dysarthria could be evidenced by a mild slurring of words that appear a bit low in terms of speech tone and could result in a loss of words that could be produced during speech output [1].

Communication is a key component in how children relate to others, feel positive about themselves, and succeed in the classroom. It is important to note the damage, though subtle, that issues of speech impairments may do to a child's social identity from a very early age [2]. Studies show that children who are subjected to speech and language services before they turn five are going to fare much better in the long run compared to those who are referred a little later, and the importance of early intervention cannot be underestimated enough in this respect [3]. Nevertheless, access to specialist speech and language therapists is not widespread around the globe as of the present day, despite the maturity of the developing nation in question [4].

Individual treatment is merely one segment of a speech-language pathologist's work. It includes diagnosis and assessment of the patient, follow-up, and personalized treatment planning [5]. Yet, ever-rising caseloads make it difficult for SLPs to provide really individualized support in a timely and effective manner. To inform clinical decision-making and therapy delivery, speech therapy increasingly employs information and communication technologies, or ICT [6]. Digital platforms and online speech therapy systems truly demonstrate promise for improving access, participation, and maintenance of care, especially for children, who generally tend to be particularly receptive to digital tools [7].

Despite the development of advancements in teletherapy and ICT-enabled services for treatment, the major existing solutions that can currently be accessed for treatment and interventions are post-diagnosis and assist in speech therapy. Early assessment and screening services are extremely rare. This is particularly true for less advanced nations, such as Pakistan, where there is a lack of psychologists and advanced healthcare facilities. This, in turn, leads to cases of speech impairments going unnoticed until they begin to impact a child's socialization patterns and communications.

Specific objectives of the study are as follows:

Designing an automated framework for classifying normal and dysarthric speech in a binary manner.

To utilize a continuous regression score of 0–100 to assess the extent of dysarthria.

To evaluate the effectiveness of hybrid transformer-based and acoustic features on large speech datasets.

To determine if the proposed framework is suitable for early-stage screening applications.

To address this, we created Clearitic AI, an artificial intelligence-based detection device for vocal disorders that identifies peculiarities in the voice of children at the earliest possible opportunity and provides the initial level of severity to expedite referrals to pediatricians of children with voice disorders. Different from previous works that usually address either dysarthria detection or severity regression alone, ClearTic AI unifies continuous severity prediction and binary classification with large-scale hybrid feature learning in a unified framework. Second, the proposed system is trained and evaluated on an unprecedentedly larger size and class-balanced dataset, which includes 10,000 speech samples from both pathological and healthy subjects. In this way, this dataset has never seen joint identifying and severity modeling.

Speech Production:

Originating from one place in the brain, it is essentially a distributed function involving the brain. Connections between frontal and temporal regions, cerebellum, and subcortical structures are involved. Overlapping circuits involving regions that handle language, respiratory or phonatory function, comprehension, timing, and motor functions are involved. [8][9][10][11][12][13].

Dysarthria refers to a condition whereby a person finds it difficult to talk effectively. This condition occurs when the muscles responsible for speaking become weak or uncontrollable, leading to the manifestation of this condition in affected individuals, who speak in slow, low, or slurred voices. This condition does not relate to talking or understanding; it is solely associated with speaking functions and how individuals express their ideas or convey their message effectively. Issues or causes of this condition include brain injuries, nerves, and muscle dysfunction, among others, especially in cases of stroke, Parkinson's disease, ALS, or multiple sclerosis, among those conditions mentioned. Some individuals also have difficulty with functions of breathing, chewing, or swallowing, which could affect their ability to speak effectively. However, this condition could range from mild symptoms of speaking changes or make speaking almost impossible for affected individuals. A speech-language pathologist tries to improve communication functions for affected individuals. Some of the areas of concern in treating this condition include effective activities or exercise for those affected, as well as alternative ways of communicating, among others [14][15]. There are major groups or categories of this condition, which depend on several variables, as seen in Figure 1 below.

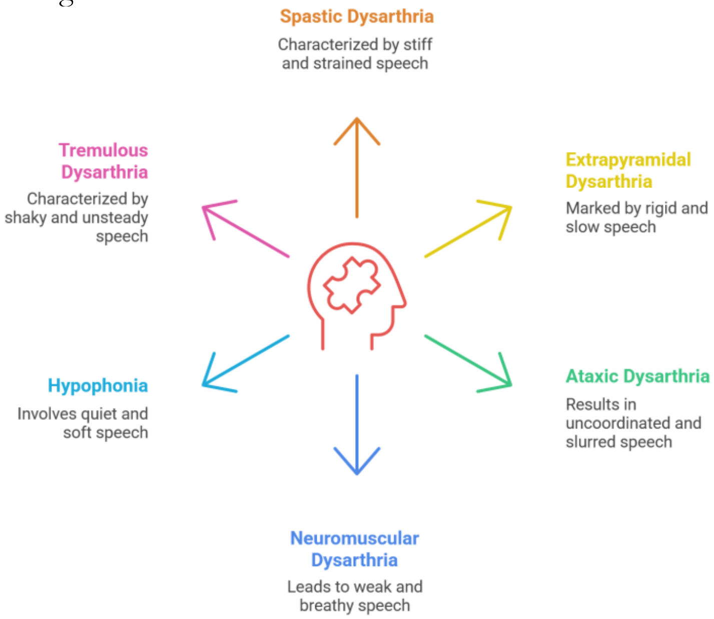
Spastic Dysarthria
Characterized by stiff and strained speech
Tremulous Dysarthria
Characterized by shaky and unsteady speech
Extrapyramidal Dysarthria
Marked by rigid and slow speech
Hypophonia
Involves quiet and soft speech
Ataxic Dysarthria
Results in uncoordinated and slurred speech
Neuromuscular Dysarthria
Leads to weak and breathy speech

Figure 1. Major categories of dysarthria

Types of Dysarthria and Their Causes:

This diagram outlines the different sorts of dysarthria and what causes them. Dysarthria is basically a speech disorder caused by brain damage, damage to the nerves, or damage to the muscles used for speaking. Extrapyramidal and hypokinetic types are often associated with a chronic disease, Parkinson's disease, while cerebellar dysarthria commonly presents with multiple sclerosis. Neuromuscular (bulbar) and mixed dysarthrias often result from ALS. Hyperkinetic dysarthria generally presents in Huntington's disease [16][17][18][19][20][21][22].

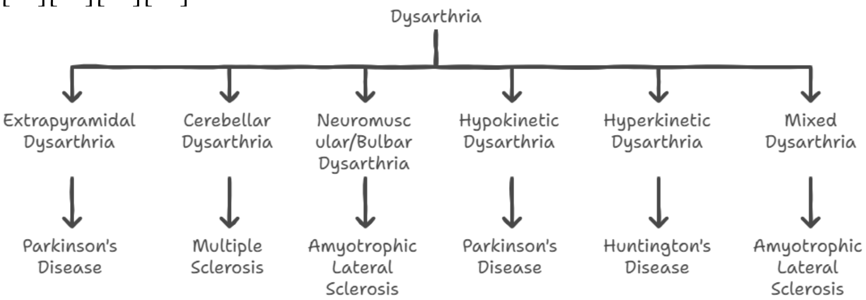

Figure 2. Types of Dysarthria and their causes

Related Study:

The development of automated systems for detecting and assessing dysarthria has seen significant advancement over the past six years, driven by improvements in deep learning and the availability of more diverse speech datasets. Early work by Kim et al. (2019) [23] laid a foundation by applying a combined Convolutional and Recurrent Neural Network (CNN-LSTM) architecture to spectrogram representations of dysarthric speech from the UA-Speech dataset. Their model achieved an accuracy of 88.7% by effectively capturing both spatial and temporal patterns in the audio. However, this approach was computationally intensive and did not address the critical need for quantifying the severity of the speech impairment, focusing solely on binary classification.

In 2020, Alnaser et al. took a different approach for ensemble learning. They extracted an extensive set of acoustic/prosodic features from the TORGO database and combined these features using the Random Forest classifier and XGBoost classifier. These methods produced an accuracy of 90.8% and demonstrated that manual features can be an effective representation of articulatory instability. Despite this, it was prone to traditional manual feature engineering methods, where it can be difficult to adapt to highly non-linear patterns observed in someone with severe dysarthria. Once again, it did not produce any measure of severity scoring [24].

In the next year, Kim and Lee investigated transfer learning to enhance generalizability in their research paper of 2021. The authors began with an existing pre-trained network for audio, named VGGish, which was pre-trained for a large audio dataset, and then fine-tuned it and used its outputs for a Support Vector Machine classifier. With the TORGO-SVD dataset combined for experimenting and testing, it reached an accuracy of 91.2%. The paper emphasized the efficacy of pre-trained networks for audio, but used a specific pre-trained network and could not measure severity levels, and was merely a detection model [25].

A shift towards the use of convolutional networks for direct representation of audio came about in Gupta et al. (2022). They trained a 2D CNN directly from the Mel-spectrograms of the UA-Speech dataset, attaining an accuracy of 89.5%. The network was very robust for inter-speaker variations but was limited in that it did not include a hybrid feature approach for a more effective combination of spectral and temporal representations, and also in that it was based on a single dataset approach [26].

The year 2023 marked the beginning of an increased move towards viable clinical applications. One exemplar of this was the development of a tele-screening system for pediatric motor speech disorders that was undertaken by Wang et al. They utilized a Random Forest approach that utilized prosodic and spectral parameters on their own private database of 1,200 pediatric speech samples and had an accuracy of 88.7%. While it marked progress because it was on pediatric patients, the method was not very accurate itself, their database was not available for testing on third-party platforms, and they were not portraying any form of severity on the disorder [27].

In recent years, transformer-based self-supervised learning has also received increasing attention. Li et al. directly addressed the task of severity measurement. They exploited HuBERT-based embeddings and ridge regression to transform the features into a severity score on the TORGO dataset, achieving an R^2 of 0.902 and an MAE of 4.12. Although it significantly progressed towards a quantitative measurement, the research remained focused on regression and skipped any classification task, and applied only to a relatively small data set [28].

In 2024, there were significant advances in the area of dysarthric speech detection. However, traditional shortcomings have still persisted. Chen et al. proposed a multi-modal CNN-LSTM model that combined acoustic features with deep models. The results were very promising with respect to both accuracy and AUC on the TORGO and UA-Speech databases. However, there were significant computational requirements and no severity estimation [24]. In the same year, Sharma and Kumar again focused on efficiency. They proposed a logistic regression classifier with a reduced version of Wav2Vec2. The results were competitive with respect to other methods. The model could even be utilized for real-time applications on mobile devices. However, there were certain losses in performance and again no estimation for severity [29]. Fast-forward to 2025. Park et al. proposed a quantized CNN model. The model could run on edge devices. The model maintained a significant portion of the original precision. However, latency was reduced to a minimum. However, this model could only perform binary classification [30].

Table 1. Summary of Related Study

Authors	Method	Dataset	Accuracy	Primary Limitation
Kim et al.2019 [23]	CNN LSTM on Spectrograms	UA-Speech (2,950)	88.7%	No severity check and computationally expensive.
Alnaser et al.2020 [24]	(RF+XGBoost) Ensemble	TORGO (2,600)	90.8%	Traditional Machine Learning is not trained on deep features, and no severity check.
Kim & Lee2021 [25]	VGGish + Transfer Learning SVM	TORGO + SVD (4,100)	91.2%	Not regression dependent on a pre-trained model.
Gupta et al.2022 [26]	2D-CNN on Mel-spectrograms	UA-Speech (2,950)	89.5%	Less dataset, no severity check, and no hybrid features extraction.
Wang et al. 2023 [27]	Random Forest + Prosodic Features	Pediatric Private (1,200)	88.7%	Private dataset and severity check.
Li et al.2024 [28]	HuBERT + Ridge Regression	TORGO (2,800)	R^2=0.902, MAE=4.12	Classification and training on a small dataset.

Chen et al.2024 [24]	CNN-LSTM along Acoustic Features	TORGO+UA-Speech (7,200)	92.3%, AUC=0.978	Computationally expensive; no severity check.
Sharma & Kumar 2024 [29]	Combination Wav2Vec2 and LR	TORGO+Pediatric (3,500)	90.1%	Accuracy is quite low; mobile-focused, but no severity check.
Park et al. 2025 [30]	Quantized CNN	TORGO and CV Subset	91.5%, Latency<100ms	Deals with Binary not with severity, for optimization of speed.

The proposed method not only identifies the disease but also checks the severity, that make in novel from the rest of recent published article as evident in Table 1.

Proposed Methodology:

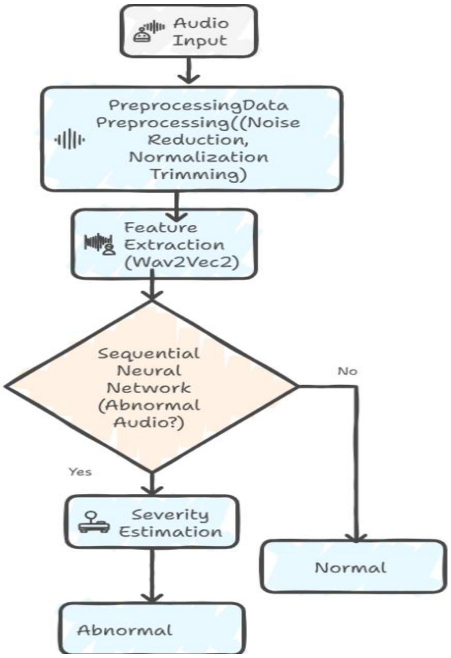
Audio Input
PreprocessingData Preprocessing((Noise Reduction, Normalization Trimming)
Feature Extraction (Wav2Vec2)
Sequential Neural Network (Abnormal Audio?)
No
Yes
Severity Estimation
Normal
Abnormal

Figure 3. Overview of Proposed Methodology

The proposed approach can be observed in Figure 3 below. ClearTic AI is designed to detect speech anomalies and their severity based on voice recordings automatically. It was our desire to develop a system capable of detecting both normal and abnormal infant speech patterns and providing workable results useful in early intervention based on those results. This goal was achieved by merging state-of-the-art transformers with classic audio feature extraction and deep learning models to produce a flexible system.

Data Collection and Dataset Construction:

A balanced dataset was created for the training of the ClearTic AI to record normal and abnormal speech patterns. In this regard, 5,000 samples of the abnormal speech patterns were obtained from the TORGO dataset to represent the characteristics of dysarthric speech patterns [31]. These samples of abnormal patterns were then combined with 5,000 samples of normal speech patterns obtained from the Mozilla Common Voice to represent healthy speakers. The dataset comprised 10,000 recordings tagged as either Normal or Abnormal speech patterns [32]. In training the model, the dataset was partitioned into sets to offer good training and an unbiased test set.

Preprocessing:

The raw audio data will always contain background noises, variations in amplitude, or silence, which may affect the extraction of features and, consequently, the results of the models. The intake process incorporated a complete preprocessing step for each audio file:

Noise reduction: The filters eliminated background noises that could have interfered with the voice.

Normalization of amplitude: Loudness was normalized to bring it to an equal level among all sample recordings.

Trimming: The leading or trailing silence was eliminated, thereby ensuring the relevant parts of the utterances are considered in the model.

These processes yielded clean and consistent inputs for both transformer networks and conventional feature extraction algorithms.

Feature Extraction:

For the effective analysis of speech, we have used a hybrid-feature-extraction strategy that combines the transformer embeddings along with the classical acoustic features as a complement to hybrid features. This dual strategy allows the ultimate system to retain both high-level temporal and low-level voice properties**.**

Transformer-Based Features:

We employed Wav2Vec2, a state-of-the-art self-supervised transformer learning model fine-tuned through a massive speech corpus. The transformer learns and abstracts well complex temporal and spectral features of speech signals. It is very adept at abstracting out minute speech features like phonetics, prosody, and rhythm, which are very important distinguishing features between normal and pathologically speech patterns. Self-elected transformers like Wav2Vec2 can very well deal with long-term speech dependency and intonation variations [33].

Classical Acoustic Features:

To provide additional information besides the transformer-based embeddings, we incorporated the traditional audio features from the library Librosa [34][35]:

Pitch (the fundamental frequency): carries the speaker's tone and intonation.

Mel Frequency Cepstral Coefficients (MFCCs): These give spectral information that is message and consonance-related.

Spectral Centroid: indicates the " brightness" of the voice, where the value indicates how the energy is distributed.

By incorporating these traditional characteristics with transformer embeddings, a more informative speech signal is obtained, which improves the model's accuracy in pinpointing anomalies.

Model Architecture and Training:

Speech Classification:

Speech Classification (Normal vs. Abnormal):

This was due to the ability of the model to learn non-linear associations from high-dimensional fused feature vectors. A sequential neural network was opted for, namely, the MLP. Unlike CNNs or any other recurrent architecture, which require raw or sequential inputs, respectively, the suggested hybrid features are already semantically rich representations. Besides, MLP is more suitable for real-time low-resource screening settings because it offers faster inferences with lower computational complexity. Therefore, we have used it, and each of these extracted characteristics, ranging from transformer embeddings obtained using Wav2Vec2 to traditional audio features such as pitch, MFCCs, and spectral centroid, was consolidated into a single feature vector, which was then used as input to a Multi-Layer Perceptron (MLP) classifier developed using a Sequential Neural Network (SNN) framework. MLP/SNN There is a feed-forward MLP that learns a non-linear mapping between high-dimensional vectors of features and labels of classes. It is convenient (Goodfellow et al., 2016) to use the Sequential API to stack dense layers and add activation functions and dropout [36]. The system takes an input image, extracts the features, and then passes them to the classification layer to produce.

Data split for training: We sectioned the data into 70% for training, 10% for validation, and 20% for testing.

Performance Measures: We employed Accuracy and the confusion matrix for performance assessment.

In essence, the use of high-level transformer representations along with traditional features translates to an effective discrimination of regular speaking patterns from irregular ones.

Severity Estimation:

We also calculated how severe the speech disorder is. Since severity is a continuous score, a regression model [37][38][39] is used to calculate a score between 0 and 100. This will help the system differentiate how severe the disorder is, from mild to severe levels. The results will not only specify whether the speech is abnormal or normal.

Clinical-Interpretation for the Severity Scores Estimation:

The range of the predicted severity score is from 0 to 100, and the score increases as the severity of the speech disorder also increases. The score of 0-25 corresponds to normal or minor symptoms, 26-50 represents mild symptoms of dysarthria, 51-75 represents slight weakness, and above 75 represents severe symptoms of dysarthria. This classification system follows the conventional severity level systems used by speech therapists.

Evaluation and Validation:

In our evaluation of ClearTic AI, we performed our testing on the 20% holdout set. Several performance indicators were considered:

Accuracy of distinction between Normal and Abnormal

Confusion matrices for deeper insight into errors

Mean Squared Error (MSE) on severity regression

The entire set of experiments was performed on Google Colab's GPUs and validated to ensure scalability and reproducibility on local servers. This combined transformer and classical method allowed the system to reliably locate speech irregularities, severity estimation, and classification of gender.

Result and Discussion:

In this section, our dataset contains 10,000 samples of voices that are divided equally into 5,000 normal samples taken from the Mozilla Common Voice dataset and 5,000 abnormal samples from the TORGO dysarthric speech corpus, and we will compare the two models that will be developed using the binary classifier model for identifying abnormal voices and the regression model for the severity score.

Dataset Composition and Exploratory Analysis:

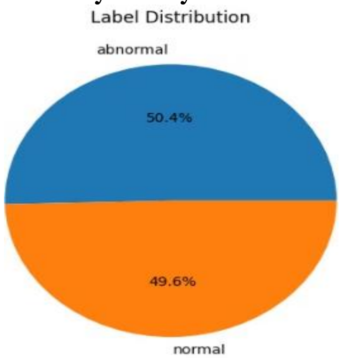

Figure 4. Label Distribution for Dataset Analysis

Label Distribution:

The dataset is perfectly balanced for training a robust model, having 50.4% and 49.6% abnormal and normal samples, respectively. This 1:1 ratio prevents the classifier from being biased towards the majority class.

Performance of the Binary Classification Model:

The confusion matrix presents a clear image of a robust, clinically applicable diagnostic system. A consistent and high degree of detection accuracy for both healthy and unhealthy voice patterns is demonstrated by the precise calls for 938 occurrences of the normal class and 968 instances of the abnormal class. A significant factor to take into account when screening a patient is the minimal likelihood of failing to identify true pathology, as evidenced by the small number of false negative cases (52 in total). The diagnostic system's precision component is kept in check by the low false alarm incidence, which is limited to 65 occurrences. This prevents a high likelihood of issuing needless alerts.

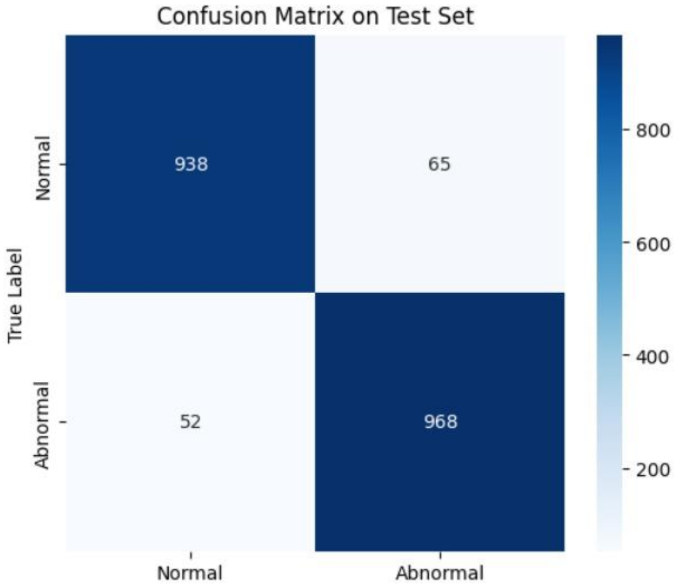

Figure 5. Confusion metric

Sensitivity (Recall / True Positive Rate):

$$\text{Sensitivity (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Sensitivity (TPR)} = \frac{968}{968 + 52} = 0.949$$

$$\text{Sensitivity (TPR)} = 0.94\%$$

Where **TP**-True Positive is a matric indicates the correct samples identified in dysarthric speech, and **FN**, i.e., False Positive, are the samples that are incorrectly classified as normal.

Specificity (True Negative Rate):

Specificity represents the ability of the system to correctly identify normal (healthy) speech samples.

$$\text{Specificity (TNR)} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Specificity (TNR)} = \frac{938}{938 + 65} = 0.935$$

$$\text{Specificity (TNR)} = 0.935\%$$

It obtained a sensitivity of 94.9% in the detection of dysarthric speech and a specificity of 93.5% for normal speech classifications, showing that it has good performance across classes.

Receiver Operating Characteristic (ROC) Curve:

When adjusting the classification threshold, the ROC Curve highlights how the True Positive Rate (also known as Recall) is compared with the False Positive Rate. It appears as if it is doing amazingly well, as its graph is steep towards the top-left corner. Its ability to discriminate between the classes is very evident by obtaining an AUC value of 0.987, which is independent of the operating point/threshold. An abnormal situation is ranked 98.7% above the normal situation based on this AUC.

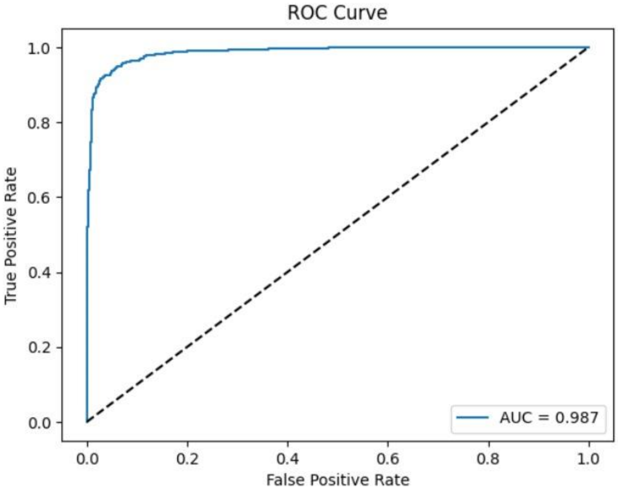

Figure 6. Receiver Operating Characteristic (ROC) Curve

Synthesis and Clinical Interpretation:

An effective two-step approach is highlighted by the study. Regression analysis provides a sophisticated measure of the severity level of the abnormality ($R^2 = 0.9389$, MAE = 3.26) to express how serious the abnormality is. First, there is the classifier, which serves as a diligent sentinel with an AUC of 0.987, accurately indicating whether an abnormality is present or not.

A few incorrect classifications (FN = 52, FA = 65) likely belong to the edges or cases with slight levels of dysarthria and/or vocal characteristics that just so happen to be unusual. However, it is evident that there is a distinct separation between levels of severity for each class, and the specific ranges given by the unique RMSE for energy characteristics give a definite acoustic explanation for its calls. The high R^2value shows a strong degree of reliance and accuracy for the predicted value of severity in terms of its 'score'.

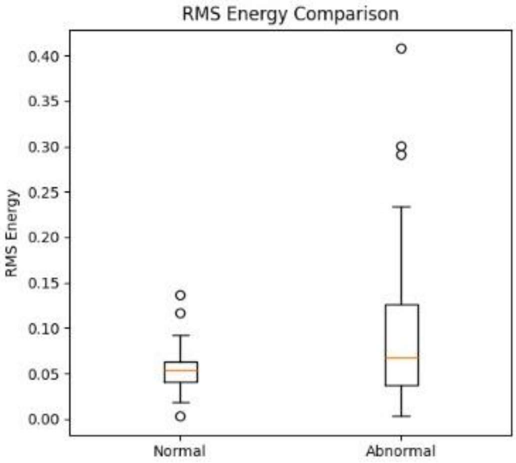

Figure 7. RMSE for energy characteristics

Comparison:

All the comparative results are as they appear in the original papers; the evaluation protocols and how datasets are split can be different for the various studies shown in Table 2.

Table 2. Comparison of the recent & our proposed Methodology

Authors	Paper Title	Dataset	Method	Accuracy	Severity R^2	Severity MAE	Key Limitations
Huma et al. 2025	Clearitic AI: Automated Dysarthria Detection & Severity Scoring	10,000 samples (TORGO + Mozilla)	Wav2Vec2 + MFCC/Pitch + SNN	94.2%	0.939	3.26	Requires a GPU for inference; limited to English speech currently.
Chen et al. 2024	Deep Learning-Based Dysarthria Classification Using Multi-Modal Features	7,200 samples (TORGO + UA-Speech)	CNN-LSTM + acoustic features	92.3%	–	–	No severity scoring; dataset is smaller; lacks pediatric representation.
Li et al.2024	Regression-Based Severity Scoring in Dysarthria Using Self-Supervised Features	2,800 samples (TORGO only)	HuBERT + ridge regression	–	0.902	4.12	Small dataset; no binary classification; limited feature fusion.
Wang et al.2023	Tele-Screening Tool for Pediatric Motor Speech Disorders	1,200 pediatric samples	Random Forest + prosodic features	88.7%	–	–	Low accuracy; no severity model; dataset not publicly available.
Sharma & Kumar,2024	A Lightweight Transformer for Dysarthric Speech Detection	3,500 samples (TORGO + pediatric)	Pruned Wav2Vec2 + logistic regression	90.1%	–	–	Lower accuracy; no regression output; model simplified for mobile use.

Conclusion:

Clearitic AI is brought into the picture. It is an AI-assisted tool aimed at diagnosing and gauging motor speech disorders such as dysarthria. clearitic AI does this while being trained on well-merged acoustic features from 10,000 speech samples. The result shows that clearitic AI operates on 94.2% overall accuracy, AUC=0.987, and R^2= 0.9389 in gauging the severity level of the disorder. This clarifies that clear AI performs very well in discriminating between normal speech patterns and pathological speech patterns. Clearitic AI has immense potential in being used as a pre-diagnostic tool that is readily available in remote locations where speech-language pathology services might be minimal. In the future, we will extend this work for the gender classification as per voice detected by the model and also on bases of severity analysis, basic therapies will be suggested by our system.

Limitations and Future Work:

Next, we lay out a staged process of clinical validation, consisting of first pilot-testing the technique, incorporating the technique into the regular practice of the certified speech-language pathologists, and finally pilot-testing the technique in the field. Looking forward, our future work will expand our focus by incorporating mechanisms of interpretation, such as SHAP-values or attention maps, into the AI tooling, with the intention of enhancing medical trust. Please note that our current study does not include a pediatric subgroup analysis, but will incorporate such in a future modeling effort focused on pediatric validation.

Acknowledgement: We would like to show gratitude to all the individuals and organizations that have contributed to this effort, directly or indirectly. Everyone's support makes this effort come to an end.

Author's Contribution: The corresponding-author conceives the study, dataset analysis, methodology development, and also prepared the original manuscript. Whereas co-authors also contributed to the collection of data, development of the model, literature review, validation of the model, and revision of the manuscript. All authors have contributed meaning-full in the manuscript.

Conflict of Interest: Authors does not have any conflict of interest.

Project details. It was our personal project for the benefit of pathologists.

References:

[1] Lindsay PenningtonNick MillerSheila Robson, "Speech therapy for children with dysarthria acquired before three years of age - Pennington, L - 2009 | Cochrane Library," Cochrane library. Accessed: Dec. 23, 2025. [Online]. Available: https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD006937.pub2/full

[2] V. M. Maria Helena Franciscatto, Marcos Didonet Del Fabro, João Carlos Damasceno Lima, Celio Trois, Augusto Moro, "Towards a speech therapy support system based on phonological processes early detection," *Comput. Speech Lang.*, vol. 65, p. 101130, 2021, doi: https://doi.org/10.1016/j.csl.2020.101130.

[3] M. Danubianu, S. G. Pentiuc, O. A. Schipor, M. Nestor, and I. Ungureanu, "Distributed intelligent system for personalized therapy of speech disorders," *Proc. - 3rd Int. Multi-Conf. Comput. Glob. Inf. Technol. ICCGI 2008 Conjunction with ComP2P 2008 1st Int. Work. Comput. P2P Networks Theory Pract.*, pp. 166–170, 2008, doi: 10.1109/ICCGI.2008.31.

[4] M. L.-N. Vladimir E. Robles-Bykbaev, "SPELTA: An expert system to generate therapy plans for speech and language disorders," *Expert Syst. Appl.*, vol. 42, no. 21, pp. 7641–7651, 2015, doi: https://doi.org/10.1016/j.eswa.2015.06.011.

[5] "Speech Sound Disorders: Articulation and Phonology." Accessed: Dec. 23, 2025. [Online]. Available: https://www.asha.org/practice-portal/clinical-topics/articulation-and-

phonology/?srsltid=AfmBOooMICpu2zctIJy1D3I_HarxrXAdowIXDL8nJ3O0_MsT40GslSYv

[6] J. P. Eugenia I. Toki, "An Online Expert System for Diagnostic Assessment Procedures on Young Children's Oral Speech and Language," *Procedia Comput. Sci.*, vol. 14, pp. 428–437, 2012, doi: https://doi.org/10.1016/j.procs.2012.10.049.

[7] A. Ben-Aharon, "A Practical Guide to Establishing an Online Speech Therapy Private Practice," *Perspect. ASHA Spec. Interes. Groups*, vol. 4, no. 4, pp. 712–718, 2019, [Online]. Available: https://pubs.asha.org/doi/10.1044/2019_PERS-SIG18-2018-0022

[8] C. J. Price, "A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading," *Neuroimage*, vol. 62, no. 2, pp. 816–847, 2012, doi: https://doi.org/10.1016/j.neuroimage.2012.04.062.

[9] S. L. S. Pascale Tremblay, "Motor response selection in overt sentence production: a functional MRI study," *Front. Psychol.*, vol. 2, 2011, doi: https://doi.org/10.3389/fpsyg.2011.00253.

[10] H. Ackermann, D. Wildgruber, and W. Grodd, "Neuroradiological activation studies on the cerebral organisation of language capacities - A review," *Fortschritte der Neurol. Psychiatr.*, vol. 65, no. 4, pp. 182–194, 1997, doi: 10.1055/S-2007-996321/BIB.

[11] P. Mariën *et al.*, "Consensus paper: Language and the cerebellum: An ongoing enigma," *Cerebellum*, vol. 13, no. 3, pp. 386–410, Dec. 2014, doi: 10.1007/S12311-013-0540-5/METRICS.

[12] H Chertkow, S Murtha, "PET activation and language," *Clin Neurosci*, vol. 4, no. 2, pp. 78–86, 1997, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/9059757/

[13] S. Geva, P. S. Jones, J. T. Crinion, C. J. Price, J. C. Baron, and E. A. Warburton, "The Effect of Aging on the Neural Correlates of Phonological Word Retrieval," *J. Cogn. Neurosci.*, vol. 24, no. 11, pp. 2135–2146, Nov. 2012, doi: 10.1162/JOCN_A_00278.

[14] S. E. G. Hannah P. Rowe, "Characterizing Dysarthria Diversity for Automatic Speech Recognition: A Tutorial From the Clinical Perspective," *Front. Comput. Sci.*, vol. 4, 2022, doi: https://doi.org/10.3389/fcomp.2022.770210.

[15] A. A. J. and R. Rajan, "Automated Dysarthria Severity Classification: A Study on Acoustic Features and Deep Learning Techniques," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1147–1157, 2022, doi: 10.1109/TNSRE.2022.3169814.

[16] N. Côté, "Speech Quality Measurement Methods," pp. 37–85, 2011, doi: 10.1007/978-3-642-18463-5_2.

[17] G. W. Ray D. Kent, "Acoustic studies of dysarthric speech: Methods, progress, and potential," *J. Commun. Disord.*, vol. 32, no. 3, pp. 141–186, 1999, doi: https://doi.org/10.1016/S0021-9924(99)00004-0.

[18] J. R. D. Hugo Botha, "Classification and clinicoradiologic features of primary progressive aphasia (PPA) and apraxia of speech," *Cortex*, vol. 69, pp. 220–236, 2015, doi: https://doi.org/10.1016/j.cortex.2015.05.013.

[19] M. N. R. Jonathan D. Rohrer, "Neologistic jargon aphasia and agraphia in primary progressive aphasia," *J. Neurol. Sci.*, vol. 277, no. 1, 2009, [Online]. Available: https://www.jns-journal.com/article/S0022-510X(08)00510-8/fulltext

[20] K. H. Takeharu Tsuboi , Hiroshi Tatsumi , Masahiko Yamamoto , Yoshiya Toyoshima , Yasuji Katayama, "A case of conduction aphasia presenting with peculiar jargon speech," *Clin. Neurol.*, vol. 61, no. 5, pp. 297–304, 2021, doi: https://doi.org/10.5692/clinicalneurol.cn-001466.

[21] J. R. Duffy, "Motor Speech Disorders: Clues to Neurologic Diagnosis," *Park. Dis. Mov. Disord.*, pp. 35–53, 2000, doi: 10.1007/978-1-59259-410-8_2.

[22] J. Rusz *et al.*, "Speech disorders reflect differing pathophysiology in Parkinson's

disease, progressive supranuclear palsy and multiple system atrophy," *J. Neurol. 2015 2624*, vol. 262, no. 4, pp. 992–1001, Feb. 2015, doi: 10.1007/S00415-015-7671-1.

[23] A. S. Hussain Albaqshi, "Dysarthric Speech Recognition using Convolutional Recurrent Neural Networks," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 6, 2020, doi: 10.22266/ijies2020.1231.34.

[24] R. R. Jagat Chaitanya Prabhala, "Enhanced early detection of dysarthric speech disabilities using stacking ensemble deep learning model," *Mach. Learn. with Appl.*, vol. 21, p. 100721, 2025, doi: https://doi.org/10.1016/j.mlwa.2025.100721.

[25] H. A. Irianta, Abdul Fadlil, and Rusydi Umar, "Transfer Learning-Based Detection of Dysarthric Speech Using Lightweight Convolutional Neural Networks," *JUITA J. Inform.*, pp. 349–358, Nov. 2025, doi: 10.30595/JUITA.V13I3.27695.

[26] S. B. M. Mahendran, R. Visalakshi, "Dysarthria detection using convolution neural network," *Meas. Sensors*, vol. 30, p. 100913, 2023, doi: https://doi.org/10.1016/j.measen.2023.100913.

[27] G. P. Usha and J. S. R. Alex, "Speech assessment tool methods for speech impaired children: a systematic literature review on the state-of-the-art in Speech impairment analysis," *Multimed. Tools Appl.*, vol. 82, no. 22, p. 1, Sep. 2023, doi: 10.1007/S11042-023-14913-0.

[28] B. Kadirvelu, L. Stumpf, S. Waibel, and A. A. Faisal, "Speaker-independent dysarthria severity classification using self-supervised transformers and multi-task learning," *PLOS Digit. Heal.*, vol. 4, no. 11, p. e0001076, Nov. 2025, doi: 10.1371/JOURNAL.PDIG.0001076.

[29] P. Wang and H. Van Hamme, "A Light Transformer for Speech-To-Intent Applications," *2021 IEEE Spok. Lang. Technol. Work. SLT 2021 - Proc.*, pp. 997–1003, Jan. 2021, doi: 10.1109/SLT48900.2021.9383559.

[30] A. S. Alluhaidan, E. M. Alanazi, N. Aljohani, and A. A. Alneil, "A real-time pediatric dysarthria speech disorder detection using residual recurrent neural network with attention U-net based transformer encoder model," *AIMS Math.*, vol. 10, no. 12, pp. 28787–28814, 2025, doi: 10.3934/MATH.20251267.

[31] "The TORGO database." Accessed: Dec. 23, 2025. [Online]. Available: https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html

[32] J. M. Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, "Common Voice: A Massively-Multilingual Speech Corpus," *arXiv:1912.06670*, 2020, [Online]. Available: https://arxiv.org/abs/1912.06670

[33] M. A. Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *arXiv:2006.11477*, 2020, doi: https://doi.org/10.48550/arXiv.2006.11477.

[34] C. R. Brian McFee, "librosa: Audio and Music Signal Analysis in Python," *SciPy*, 2015, [Online]. Available: https://proceedings.scipy.org/articles/Majora-7b98e3ed-003

[35] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002, doi: 10.1109/TSA.2002.800560.

[36] G. I. and B. Y. and C. A., "Deep Learning," pp. 1–23, 2016, Accessed: Jun. 17, 2025. [Online]. Available: https://mitpress.mit.edu/9780262035613/deep-learning/

[37] "Pattern Recognition and Machine Learning | Springer Nature Link (formerly SpringerLink)." Accessed: Dec. 23, 2025. [Online]. Available: https://link.springer.com/book/9780387310732

[38] K. P. Murphy, "Machine learning - a probabilistic perspective," *Adapt. Comput. Mach. Learn. Ser.*, 2012.

[39] W. N. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed,

"HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021, doi: 10.1109/TASLP.2021.3122291.