Mathematics

https://www.aimspress.com/journal/Math

AIMS Mathematics, 10(12): 28787–28814.
DOI: 10.3934/math.20251267
Received: 12 August 2025
Revised: 03 October 2025
Accepted: 22 October 2025
Published: 08 December 2025

Research article

A real-time pediatric dysarthria speech disorder detection using residual recurrent neural network with attention U-net based transformer encoder model

Ala Saleh Alluhaidan[1,*], Eman M Alanazi[2], Nasser Aljohani[3] and Amani A Alneil[4,5]

[1] Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Saudi Arabia
[2] Department of Health Informatics, College of Health Sciences, Saudi Electronic University, Saudi Arabia
[3] Department of Information Systems, Faculty of Computer and Information Systems, Islamic University of Madinah, Medina 42351, Saudi Arabia
[4] Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam bin Abdulaziz University, AlKharj, Saudi Arabia
[5] King Salman Centre for Disability Research, Riyadh 11614, Saudi Arabia

* **Correspondence:** Email: asalluhaidan@pnu.edu.sa.

Abstract: Speech disorders have a significant impact on quality of life, as they decrease the ability to define one's character, exercise autonomy, and frequently affect relationships and self-esteem, particularly in young children. Dysarthria is a neurological illness that affects motor speech pronunciation. Young children who experience this disorder have no issue with their understanding, but they have a problem expressing their words. They might struggle to communicate precisely and smoothly with their friends and family members due to this illness. A dysarthric child has significant trouble with communication, as this disorder causes poorly pronounced phonemes and poor speech articulation. To address this condition, numerous speech assistive technologies have been developed for consumers with dysarthria, tailored to the level of severity. Currently, deep learning (DL) systems offer potential for objective evaluation, thereby improving diagnostic accuracy. Its goal is to systematically analyze present approaches for detecting dysarthria based on severity levels. In this manuscript, a novel pediatric dysarthria disorder detection framework using residual recurrent neural network and transformer (PD3F-RRNNT) technique is proposed. The PD3F-RRNNT technique aims to develop a real-time recognition method for accurately detecting dysarthria speech disorders in children, supporting early diagnosis and intervention. Initially, the audio processing phase involved various steps, including voice activity detection (VAD), noise removal, pre-emphasis, framing, windowing, and normalization, to transform and extract significant data from audio signals. Furthermore, the PD3F-RRNNT method utilizes the transformer-attention-based U-Net (TransAttUnet) technique for feature extraction. Finally, the residual bidirectional gated recurrent unit (RBG) method is employed to detect and classify speech disorders accurately. The experimental validation of the PD3F-RRNNT model is performed under the dysarthria and non-dysarthria speech dataset. The comparison analysis of the PD3F-RRNNT model revealed a superior accuracy value of 99.50% compared to existing techniques.

Keywords: dysarthria; speech disorder; residual bidirectional gated recurrent unit; transformer-attention-based U-net; deep learning
Mathematics Subject Classification: 37M10

1. Introduction

Speech disorders significantly impact the quality of life; they limit the ability to express one's personality, assert independence, and often influence relationships and confidence, particularly in young children [1]. These children struggle with communication issues, and even though they use hand gestures, they still fail to deliver messages precisely and clearly to others. One of the major disorders impacting them is dysarthria. Dysarthria is a nerve-related condition that harms motor speech control. It is a motor speech disorder affecting countless children by altering language due to brain injury [2]. A child with dysarthria finds communication very challenging, as the condition causes unclear or absent phonemes and weak speech control. Simply put, it is a disorder where muscle issues hinder speech production, often making word pronunciation hard. There are various types of dysarthria based on the part of the nervous system affected, yet all types impact consonant and vowel clarity, resulting in slurred speech [3]. These disorders can only be identified before the age of 3 in children born with them. Dysarthria is a speech impairment that results from weakness in the muscles involved in speech production [4]. It is linked to conditions that result in brain damage, such as amyotrophic lateral sclerosis, cerebral palsy, and brain injury. Speech disorder in children influence their speech flow and clarity. Dysarthria affects the prosodic elements of speech, such as speed, volume, and rhythm. Children with dysarthria often speak either very slowly or very quickly, with inconsistent volume and rhythm, resulting in irregular prosody and articulation errors [5]. Signs of dysarthria may vary significantly depending on the underlying cause and severity, resulting in speech that ranges from slightly slurred to extremely unclear as the condition progresses. Delayed identification and care raise the chance of social challenges and educational difficulties [6]. Due to the significant shortage of speech-language pathologists (SLPs), there is growing interest in computer-aided speech therapy (CAST). Several CAST applications are specifically designed to target children with speech impairments. Yet, the majority of these applications focus on the automatic presentation of speech therapy prompts, allowing the session to be directed solely by the SLP or the child [7]. Only Vocaliza, STAR, and Tabby Talks apply speech evaluation to give automatic responses and assist in guiding the therapy tasks [8]. Still, the precision of all of these automatic impaired speech evaluation systems falls short for medical use [9]. Additionally, child speech exhibits greater inter- as well as intra-speaker differences than adult speech, which makes it more challenging to process [10]. Highly trained SLPs typically employ the conventional approach; however, the growing need for effective and comprehensive assessment techniques has driven a significant shift in practice. Signaled by the rise of deep learning (DL), we can now combine various AI-based methods with the conventional SLA framework, offering a substantial advancement in the domain of speech and language analysis.

1.1. Motivation for real-time detection of pediatric dysarthria

It becomes crucial for every child to maintain effective social communication, and the child's ability to express thoughts and emotions clearly is affected by speech impairments, specifically those affecting motor control. Therefore, it is crucial to detect speech impairments to provide timely interventions that enhance communication skills and overall quality of life. Automated systems utilizing advanced neural networks offer promising potential for efficiently and objectively detecting speech disorders. Developing robust real-time detection models can help clinicians and caregivers monitor and manage these conditions more effectively.

1.2. Main contributions of the study

This manuscript presents a novel framework for detecting pediatric dysarthria disorder, utilizing a residual recurrent neural network and transformer (PD3F-RRNNT) model. The main contributions of this paper are as follows:

- The multilevel audio pre-processing, including voice activity detection (VAD), noise removal, pre-emphasis, framing, windowing, and normalization, is used to improve the transformation of audio signals. These steps enhance the clarity and quality of the input data, enabling more accurate feature extraction. This pre-processing process assists robust speech analysis, which is crucial for effective disorder detection and classification.
- The TransAttUnet technique is utilized for performing robust feature extraction from pre-processed audio, improving the capture of both global and local contextual data. The capability of the approach is also enhanced by precisely representing intrinsic speech patterns. The method also improves feature richness for subsequent classification tasks by integrating multi-scale skip connections and self-aware attention mechanisms (AMs).
- The residual bidirectional gated recurrent unit (RBG) technique is used to improve the accuracy of speech disorder detection and classification by effectively capturing temporal dependencies in audio data. The residual connections (RC) improved gradient flow, enabling deeper network training and better feature learning. This architecture facilitated precise detection of speech disorder patterns, improving overall diagnostic performance.

The integration of TransAttUnet and RBG for extraction and classification presents a novel framework for effectively handling intrinsic audio signals. This integration utilizes advanced AMs and temporal modelling to capture complex speech patterns. Its unique architecture enhances detection accuracy and robustness. Overall, it presents an innovative approach to diagnosing speech disorders

2. Literature review on speech disorder methods

Sravya et al. [11] focused on creating a hybrid machine learning (ML) and DL approach for improved autism spectrum disorder (ASD) recognition through multimodal data sources. This approach merges behavioural questionnaire data classified and processed through an XGBoost paradigm and MobileNet framework enhanced by a recurrent neural network (RNN). Pre-processing stages included feature selection (FS) through convolutional feature extraction, TF-IDF encoding, and calculation of mutual information scores. Bindu and Devi [12] aimed to accurately execute ASD classification and gain a deeper understanding of the classification step to identify key features suitable for predicting the disorder. It utilizes the NASNet-Mobile framework for ASD recognition that is combined with an XAI model. Gao and Song [13] presented a new methodology, called HE-MF, that contains a hierarchical feature extractor mechanism (HFEM) and a multimodal deep feature integration mechanism (MDFIM) technique. HFEM aims to develop a multilevel, fine-grained feature extractor that enhances the methodology's discriminatory capability by gradually capturing discriminatory functional connections at both the overall subject and intra-group levels. Hu et al. [14] presented an inclusive technique for detecting unique speech patterns by examining examiner-patient dialogues. The authors used ML for regression and classification tasks to examine speech features. This classification method concentrates on distinguishing between non-ASD and ASD cases. Dia et al. [15] proposed a supervised learning model by using video frames from YouTube. Mengash et al. [16] introduced an automatic ASD identification by employing an owl search algorithm (OSA) alongside an ML (ASDC-OSAML) technique. This technique primarily aims to recognize and classify ASD. Furthermore, the beetle swarm antenna search (BSAS) method is applied to detect and classify ASD. Chola Raja and Kannimuthu [17] suggested a DL alongside a meta-heuristic algorithm to carry out the FS and feature extraction process. The FS step comprises two sub-phases: MH-driven FS and DL-driven feature extraction. The efficient CNN framework is applied for extracting key features. The hybrid heuristic algorithm known as the seagull-elephant herding optimizer algorithm (SEHOA) method is utilized from the extractor features of CNN. Almadhor et al. [18] proposed a spatio-temporal dysarthric ASR (DASR) scheme that uses a multi-head attention transformer (MHAT) and a spatial CNN (SCNN) for speech feature extraction. This approach enables DASR to acquire the phoneme shapes pronounced by individuals with dysarthria. The UA-Speech dataset is employed with various speech fluency levels. However, due to the proportion of practical speech data to the number of distinct classes, the recommended DASR scheme utilizes transfer learning (TL) to generate synthetic influence and images.

Klempir et al. [19] compared Wav2Vec 1.0 and Wav2Vec 2.0 for the detection of Parkinson's disease (PD). Utilizing techniques such as transformer-based self-supervised learning (SSL), transfer learning, and multi-criteria decision analysis (TOPSIS), the efficiency is evaluated. The study compared Hidden-unit bidirectional encoder representations from transformers (HuBERT), whisper, waveform language model (WavLM), data to vector (data2vec), and seamless multilingual multimodal machine translation (SEAMLESSM4T) models. Attaluri, CHVS, and Chittepu [20] introduced a technique that utilizes the OpenAI Whisper model, integrated with fine-tuned large language models (LLMs), such as GPT-4. o, LLaMA 3.1 70B, and Mistral 8x7B. Mahum et al. [21] proposed a model to improve the Swin transformer (ST)'s ability to learn local features. Four key modules, namely network for local feature capturing (NLF), convolutional patch concatenation, multi-path (MP) integration, and multi-view block ST (MVB-ST), are utilized to achieve accurate disorder classification. Ng et al. [22] developed a fully automated microscopic and macroscopic approach for detecting speech sound disorders (SSDs) by utilizing a deep neural network (DNN) technique for phonological error detection and speaker-level embeddings for holistic speech analysis. Shahamiri, Mandal, and Sarkar [23] introduced a method by employing depthwise separable convolution (DSC) neurons and RC. A deep acoustic model is also used to enhance recognition accuracy and mitigate speech variability. Sung et al. [24] developed an automated system for detecting SSD in children by utilizing deep learning and neural network methods. Manoswini, Sahoo, and Swetapadma [25] introduced a methodology to detect speech language impairment (SLI) in children. It evaluates a range of methods, including relative spectral transform (RASTA), wavelet packet transform (WPT), linear predictive coding (LPC), perceptual linear prediction (PLP), mel-frequency cepstral coefficients (MFCC), complex quantisation cepstral coefficients (CQCC), and perceptual noise cepstral coefficients (PNCC). These features are integrated with DL techniques, such as transformer, temporal convolutional networks (TCN), and TabNet to improve detection accuracy and support clinical diagnosis. Kim et al. [26] improved SSD detection by employing a multi-head model with shared feature extraction and age-dependent classifiers to reduce age bias. Mun, Kim, and Chung [27] developed a cascaded multimodal framework for assessing ASD severity by utilizing ASR for transcription. The linguistic and acoustic features are extracted using speech-language foundation models. Moreover, dual speech foundation models are also used for capturing atypical segmental and suprasegmental speech characteristics in ASD. Ziani et al. [28] proposed an adaptive educational system for children with ASD by utilizing speech recognition technology and generative adversarial networks (GAN), specifically MelGAN. Comparison analysis of existing speech disorder recognition systems in Table 1.

Table 1. A comparative study of different advanced techniques.

Authors	Years	Objectives	Methods	Datasets	Performance analysis
Sravya et al. [11]	2025	To advance a hybrid ML and DL methodology for improved identification of ASD to employ data resources. The presented model combines approaches, such as behavioral and questionnaire data, to employ XGBoost	XGBoost, MobileNet, and RNN	ASD dataset	Accuracy of 94.1% and 89.64%
Bindu and Devi [12]	2025	The primary goal is to accurately classify ASD and interpret the classification procedure in a preferred manner to identify key aspects suitable for disease prediction	NASNet-Mobile	-	Accuracy of 0.9607

Continued on next page

Authors	Years	Objectives	Methods	Datasets	Performance analysis
Hu et al. [14]	2024	Projects a comprehensive method to recognize distinct speech patterns by the identification of examiner-patient dialogues. To employ ML for either classification or regression tasks to examine these speech features. The classification technique focused on discriminating between ASD and non-ASD cases	ML	-	Accuracy of 87.75%
Dia et al. [15]	2024	A supervised learning model for classifying ASD and assessing the level of impact among autistic children. The utilization of a model for medical patient observation and management	Supervised learning	SSBD, AffectNet, IEMOCAP, and CK+	Accuracy of 63% and 0.998
Mengash et al. [16]	2023	Presents a novel method for the identification of ASD. The process of the BSAS model helps regulate the parameter values of the ID3 classifier	ASDC-OSAML, BSAS	-	Accuracy of 99.66%
Chola Raja and Kannimuthu [17]	2023	The phase of FS has dual sub-stages, such as DL-based feature extraction and MH-based FS	CNN and SEHOA	ABIDE dataset	Accuracy of 98.6%

Continued on next page

Authors	Years	Objectives	Methods	Datasets	Performance analysis
Almadhor et al. [18]	2023	Presents a DASR for visually eliminating the speech features, and the DASR learn the shape of phonemes pronounced by dysarthric people	DASR, MHAT, SCNN	UA-Speech dataset	Accuracy of 20.72%
Klempir et al. [19]	2025	To compare Wav2Vec 1.0 and 2.0 for PD speech detection	Wav2Vec 1.0 and Wav2Vec 2.0, SSL, TOPSIS	Three multilingual PD speech datasets	Non-Spontaneous Speech: w2v1-FEA Accuracy ≈ 0.75
Attaluri, CHVS, and Chittepu [20]	2024	To develop an advanced framework using LLMs and multimodal techniques	Speech-to-text conversion, fine-tuned and benchmark LLMs	TORGO and Google speech data (emotion-labelled)	High accuracy in speech reconstruction and emotion detection
Mahum et al. [21]	2025	To improve local feature extraction for accurate dysarthria detection using an ST-based model	NLF, convolutional patch concatenation, MP integration, MVB-ST	Mel-spectrograms from dysarthric speech samples	Accuracy: 98.66%, reduced false positives
Ng et al. [22]	2024	To develop fully automatic microscopic and macroscopic systems in young children	DNN, phonological error detection, speaker-level embeddings, similarity Score Aggregation	Impaired Child speech recordings	Unweighted average recall (UAR) 84.0%–91.9%
Shahamiri, Mandal, and Sarkar [23]	2025	To develop a deep dysarthric ASR model for improved word recognition	DSC, RC	UA-Speech (15 dysarthric subjects)	WRR improvement: 22.58%, average gain: 10.81%, moderate cases: +14.26%

Continued on next page

Authors	Years	Objectives	Methods	Datasets	Performance analysis
Sung et al. [24]	2024	To develop an automated SSD detection system in children using DL models	Audio classification, neural network models, DL	573 children's speech recordings (92 SSD cases)	Unweighted average recall (UAR): 73.9%
Manoswini, Sahoo, and Swetapadma [25]	2025	To detect the optimal feature extraction technique in children using DL models	RASTA, WPT, LPC, PLP, MFCC, CQCC, PNCC, transformer, TCN, TabNet	Children's recorded speech samples	Accuracy: 100.00%, best combo: PNCC + TabNet
Kim et al. [26]	2024	To improve the automatic model by applying debiasing techniques to address age and speaker biases in audio data	Multi-head model, age-dependent classifier, data augmentation	Korean SSD sataset	Significant performance improvement
Mun, Kim, and Chung [27]	2025	To develop a fully automated multimodal framework using speech and language features	Raw audio processing, linguistic and acoustic feature extraction, dual apeech foundation models, coattention mechanism integration	ASD speech recordings with human severity ratings	Spearman's correlation: 0.5629
Ziani et al. [28]	2024	To develop an adaptive educational system using speech recognition and generative adversarial networks	Speech recognition technology, MelGAN	Algerian children with ASD (Algerian dialect)	Promising accuracy, similarity scores

Although existing studies have demonstrated improvements in ASD and speech disorder detection using ML, DL, and transformer-based models, several limitations remain. Few studies create a research gap in data availability and diversity due to the scarcity of labelled datasets. Various techniques also face difficulty with variability in speech patterns due to age, severity, and environmental noise, highlighting a requirement for more robust and adaptive frameworks. Furthermore, the accuracy of multimodal and hybrid systems is improved, but they tend to increase computational complexity and compromise real-time applicability. Diverse models focus on specific disorders or modalities, lacking a comprehensive model for various speech impairments. Moreover, limited integration of emotion recognition with speech correction leaves a gap in holistic communication assistance. Addressing these research gaps can significantly enhance the generalizability of the technique, improve efficiency, and increase clinical usability.

3. System design and workflow

In this manuscript, a novel PD3F-RRNNT approach is proposed. The PD3F-RRNNT technique aims to develop a real-time recognition method for accurately detecting dysarthria speech disorders in children, supporting early diagnosis and intervention. To achieve this, the PD3F-RRNNT technique involves audio data preparation, deep feature representation using TransAttUnet, and DL-based speech classification. Figure 1 illustrates the workflow of the PD3F-RRNNT technique.

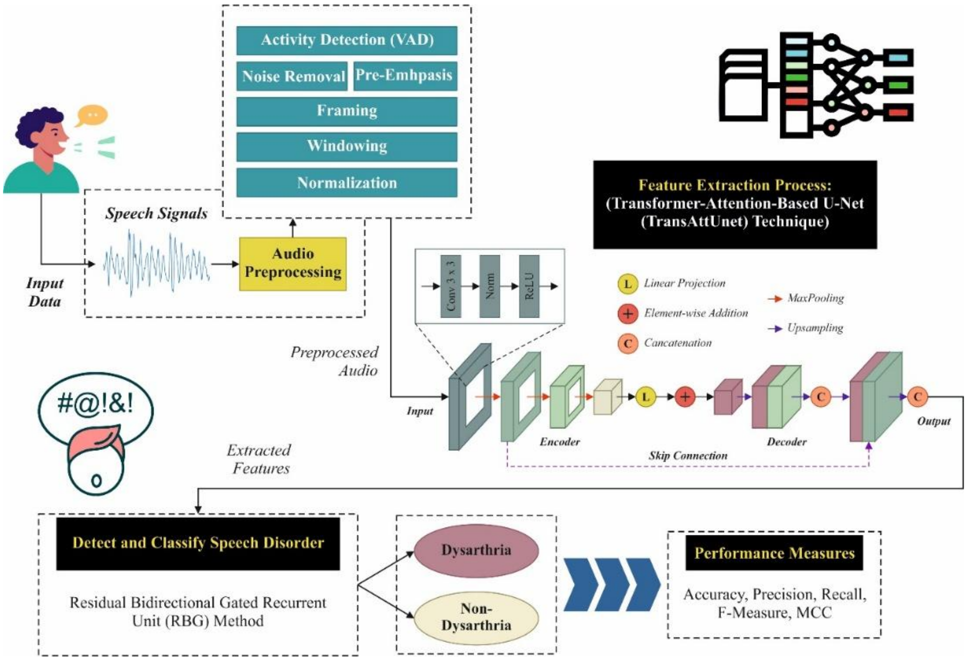


Figure 1. Workflow of the PD3F-RRNNT approach.

3.1. Pre-processing of speech inputs

Generally, speech signal pre-processing is implemented to produce the speech signal, removing features [29]. The speech signal is pre-processed in these ordered stages. Initially, the VAD is performed to keep only the voice part of the speech signal. Subsequently, noise is reduced to balance the signal's frequency spectrum. Moreover, a normalization and windowing function of the vocal tract is accomplished to enhance the SNR and signal spectrum.

VAD: VAD is a substantial model in speech signal pre-processing, which recognizes the voiced part of speech. During speech activity recognition, ASR is employed to regulate the beginning and end of speech words, thus enabling it to meet the inadequate hardware capabilities and enhance CPU efficiency. The VAD model depends upon the choice of features, which depict the discriminative property of noise and speech. Dual classes of features are differentiated; the primary class comprises time-domain features, such as zero-crossing rate (ZCR), short-time average magnitude (STAM), and short-term energy (STE), which are widely employed features. The second class is the frequency-domain features, namely the spectral power (SP) and spectral flatness (SF).

Noise removal: Environmental noise significantly impacts speech pre-processing, and denoising a speech signal enhances speech quality and improves its robustness. Noise extraction is the process of separating noise from various sounds, including speech and other environmental sounds.

Pre-emphasis: A high-pass filter is applied to the speech signal to enhance the high-frequency components, thereby achieving a balanced spectrum. Because of the glottal source, voiced sounds typically exhibit a -12 dB/octave slope, which is offset by a + 6 dB/octave boost stimulated by acoustic energy radiating from the lips. The pre-emphasis filters extract some glottal effects in the vocal tract parameter. Typically, a pre-emphasis filter is used, with a transfer function given by Eq (1).

$$H(z) = 1 - bz^{-1}. \tag{1}$$

The frequency response of this filter increases with frequency, amplifying high-frequency components. Thus, the speech spectrum is balanced by compensating for the natural roll-off introduced by the vocal tract. Here, b regulates the slope of the filter between 0.4 and 1.0 to improve the high-frequency components while maintaining signal stability. This range ensures a balanced emphasis to counteract the natural spectral tilt of voiced sounds without overly distorting the original speech signal.

Framing and windowing: The next pre-processing phase involves segmenting the instances into short frames of 30ms duration to produce quasi-stationary signals. Every frame is overlapped with the neighboring frame by 60 per cent of the frame size and windowed to preclude discontinuity among succeeding frames. The Hamming window function, a tapered and flattening mathematical function, is applied to the edges of the window by multiplying every frame of the signal by the window function to decrease the effect of spectral leakage and artefacts that occur from framing. Windowing is primarily beneficial when processing signals to utilize Fourier-based algorithms, such as the discrete Fourier transform (DFT). Samples of other window functions include the Hanning, rectangular, Blackman, and Hamming windows, each with distinct features. Mathematically, the Hamming window function is given in Eq (2).

$$w(n) = 0.54 - 0.46\,cos\left(\frac{2\pi n}{N-1}\right), 0 \leq n \leq N-1. \tag{2}$$

Now, N = number of samples in every frame

$$y(n) = x(n) \cdot w(n). \tag{3}$$

Here, $y(n)$ depicts an outcome of windowing, and (n) indicates the discrete signal. The DFT of every windowed frame is executed to employ Eq (4), attaining the magnitude spectrum of the signal.

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-j\frac{2\pi nk}{N}};\ 0 \leq k \leq N-1. \tag{4}$$

Here, N indicates the number of points employed to calculate the DFT.

Normalization: Speech signal normalization is the last phase that involves balancing the signal spectrum and transforming the signal information into a standardized form. It is primarily used to reduce the effects of noise, channel distortion, and speech signal distortion. In the normalization phase, the SNR is enhanced, and the signal's spectral variation is eliminated or normalized.

3.2. Deep feature representation with transAttUnet

Next, the PD3F-RRNNT model utilizes the TransAttUnet technique for the feature extraction process [30]. This model demonstrates efficiency in capturing both local and global contextual data through the integration of transformer-based self-attention modules (AMs) and the U-Net's encoder-decoder architecture. This technique efficiently models the long-range dependencies and spatial relationships, and also improves feature representation from intrinsic audio signals. The multi-scale skip connections of the model facilitate the better integration of semantic features at diverse resolutions, thereby enhancing segmentation and discrimination tasks. Furthermore, the dual self-aware attention modules provide richer contextual awareness, making transAttUnet more robust to noise and variability in speech data compared to conventional methods. This results in enhanced accuracy and generalization in speech disorder detection. The RBG utilizes hyperparameters, such as the number of hidden units and dropout rate, to prevent overfitting. The RCs help reduce gradient vanishing by allowing gradients to bypass recurrent layers, thereby enabling better gradient flow during backpropagation. For TransAttUnet, audio signals are transformed into image-like inputs by transforming them into spectrograms or mel-spectrograms with dimensions $C \times H \times W$, where c represents channels (e.g., 1 for grayscale), and H, W represent time-frequency resolution.

An input image $X \in \mathbb{R}^{C\times H\times W}$, where $H \times W$ depicts the spatial resolution of the image sample and C indicates channel counts. The objective of automated image segmentation is to forecast the equivalent pixel-wise semantical mapping label size of HxW. TransAttUnet utilizes multilevel, complementary self-aware attention units, along with multi-scale skipping connections, to enhance the semantic segmentation image quality. The spectrograms are generated using a window size of 25 ms, a hop length of 10 ms, and 128 Mel frequency bins, which define the dimensions H and W.

Self-aware AM

Primarily, the transformer-attention-based U-Net expands the standard U-Net by incorporating a strong and effective self-aware attention module situated at the bottom of a U-shaped framework, serving as the connection between the decoder and encoder sub-networks. It comprises two individual self-attention modules, namely the global spatial attention module (GSAM) and transformer self-attention (TSAT), which assist in acquiring a richer and broader contextual representation.

I. TSAT: The multi-head SA function from the Transformer that enables the method for assisting semantic data from global representation sub-spaces. Acquire the contextual data of relative and absolute position. The element of TSAT presents the learnt position encoded as input to the encoder feature, which is shared through every attention layer for a specified sequence of query/key-value. Particularly, the features of encoder $F \in \mathbb{R}^{c\times h\times w}$ are embedded into three inputs, comprising the matrix of queries $Q \in \mathbb{R}^{c\times(h\times w)}$, the matrix of keys $K \in \mathbb{R}^{c\times(h\times w)}$, and $V \in \mathbb{R}^{c\times(h\times w)}$.

$$Q = F \cdot W_q, K = F \cdot W_k, V = F \cdot W_v. \tag{5}$$

Now, W_q, W_k, and W_v refer to embedding matrices of diverse linear projections. Subsequently, a scaled dot-product operation among Q and the transposed version of K produces the matrix of attention mapping $A \in \mathbb{R}^{c\times c}$, which depicts the similarity of the specified component from Q with respect to global units of K. The aggregation of value, weighted by attention weight, is achieved through the contextual attention mapping A, which multiplies V.

$$TSA(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{6}$$

Now, $\sqrt{d_k}$ denotes the dimension of the query/key-value sequence. Finally, refine the optimized mapping feature to achieve the final output of TSAT.

II. GSAM: SAA utilizes the GSAM element's global context to learn features and encode wider contextual data positions into local aspects, thereby enhancing intra-class compactness and improving feature representation.

Primarily, dual diverse kinds of convolution operations are employed to encode features. F_{en} creates the mapping features $F_p^{c'} \in \mathbb{R}^{c' \times h \times w}$ and $F_p^c \in \mathbb{R}^{c \times h \times w}$, $c' = c/8$. Then, $F_p^{c'}$ is reshaped and transposed to map features $M \in \mathbb{R}^{(h \times w) \times c'}$ and $N \in \mathbb{R}^{c' \times (h \times w)}$, whereas F_p^c is transposed to $W \in \mathbb{R}^{c \times (h \times w)}$, correspondingly. Previously, a matrix multiplication operation with softmax normalization was carried out $on\, M$ and N, resulting in a position attention mapping $B \in \mathbb{R}^{(h \times w) \times (h \times w)}$.

$$B_{i,j} = \frac{exp(M_i \cdot N_j)}{\sum_{i=1}^{n} exp(M_k \cdot N_j)}. \tag{7}$$

Now, $B_{i,j}$ determines the effect of $the\ i^{th}$ location on the j^{th} location, and $n = h \times w$ indicates the pixel counts. Subsequently, W is multiplied by B, and the resultant feature is calculated at every position.

$$GSA(M, N, W)_p = \sum_{q=1}^{h \times w} \left(W_q \cdot B_{p,q}\right). \tag{8}$$

Now, W and B represent the weight and attention matrix, where $B_{p,q}$ indicates the attention weights.

III. Attention embedding fusion: A weighted integration method, utilizing attained contextual data and spatial relations, is employed to combine the original and attention feature embeddings at the end of SSA.

$$F_{SAA} = \lambda_1 F_{tsa} + \lambda_2 F_{gsa} + F_{en}. \tag{9}$$

Now, λ_1 and λ_2 indicate the scaling parameter, which regulates the significance of self-attention and spatial attention mappings, respectively.

Multi-scale skip connection

Various sophisticated works illustrate the efficacy of multi-scale feature fusion in encoding both local and global contexts. Notably, the multi-scale skipping connection method aims to collect features from different semantic scales through a sequence of transition operations, comprising concatenation, convolution, and up-sampling.

I. Cascade connection: The mapping features of different semantic scales from every block are up-sampled to a common resolution across bilinear interpolation, and directly concatenated to a unified feature representation.

$$F = f_n(v_1(F_1) \oplus v_2(F_2) \oplus \ldots \oplus F_n). \tag{10}$$

Now, $v_n(\cdot)$ and $f_n(\cdot)$ refer to up-sample and mixed convolution operations in the n^{th} phase and $\oplus$ indicate concatenation operations, correspondingly.

II. RC: In each decoder block, the input mapping features are up-sampled to the output resolution by bilinear interpolation, and are formerly concatenated with the output mapping feature as the input of the subsequent block.

$$F_n = f_n\big((F_n) \oplus v_{n-1}(F_{n-1})\big). \tag{11}$$

In Eq (11), the F_n in the parentheses indicates the input feature, while the initial F_n denotes the output after processing with f_n.

III. Dense connection (DC): The up-sampling features of preceding blocks in the encoder are combined with the input of the existing block, and the mapping features of the output are employed as inputs for every succeeding block.

$$F_n = f_n\big(v_1(F_1) \oplus v_2(F_2) \oplus \ldots \oplus v_{n-1}(F_{n-1})\big). \tag{12}$$

Specifically, the transformer-attention-based U-Net utilizes a dual, diverse, and multi-scale skipping connection method, comprising residual and DC connections, to regulate the up-sampling process in the decoder sub-network.

3.3. DL-based speech classification

Finally, the RBG method is employed to accurately detect and classify speech disorders [31]. This method effectively captures temporal dependencies in sequential speech data from both past and future contexts, which is considered significant for precisely detecting speech disorders. This model also helps the RCs mitigate the vanishing gradient problem. The technique also enables deeper network training and better data flow across layers, compared to standard RNNs or LSTMs. The convergence speed and stability of the model are also improved by the residual design of the technique, making it appropriate for handling the intrinsic and noisy characteristics of speech input. The BiGRU methodology presents a more effective architecture with fewer parameters while maintaining robust performance. Figure 2 illustrates the structure of BiGRU.

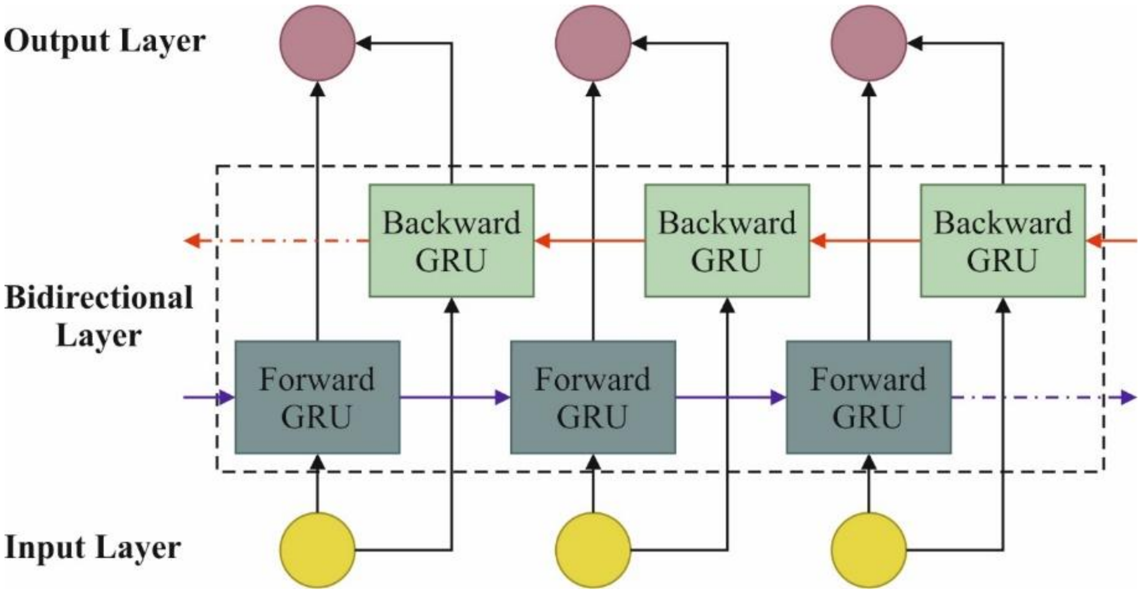


Figure 2. Structure of the BiGRU model.

The conventional GRU network depends upon the unidirectional propagation of data while learning temporal features. It fails to fully employ data from either previous or upcoming directions, resulting in the inadequate removal of temporal features or affecting the efficacy of image sequence identification. Thus, it utilizes a bi-directional propagation BiGRU network, where every time step comprises data from either the preceding or upcoming time steps. This allows the system to utilize the time-based correlation between frames and learn a more effective feature. The BiGRU network architecture comprises two unidirectional GRUs operating in opposite directions. One GRU handles the sequence from beginning to end (forward direction), while the other handles it from end to beginning (backward direction). The mathematical model for their output is specified in Eqs (13)–(15).

$$h_{t(fwd)} = G_{fwd}\left(x_t, h_{t-1(fwd)}\right). \tag{13}$$

$$h_{t(rwd)} = G_{rwd}\left(x_t, h_{t+1(rwd)}\right). \tag{14}$$

$$h_{t(bi)} = h_{t(fwd)} \oplus h_{t(rwd)}. \tag{15}$$

Now, $h_{t(rwd)}$ indicates the state of reverse GRU, $h_{t(fwd)}$ depicts forward GRU, and $\oplus$ refers to the concatenation of dual states $h_{t(fwd)}$ and $h_{t(rwd)}$, which leads to the output $h_{t(bi)}$ of BiGRU. Furthermore, G_{fwd} and G_{rwd} depict the forward and reverse GRU functions, respectively.

The BiGRU network is designed to address the problem of gradient vanishing that arises with an increase in the GRU network unit counts. To enhance the capability for learning temporal features, we address the issue of gradient vanishing, which occurs from excessive GRU network components. Each residual block comprises two BiGRU components, resulting in an overall total of eight GRU components. The network of RGB incorporates features through adjacent frames, accumulates temporal data across the entire sequence, and finally utilizes a softmax classification layer to regulate the classes.

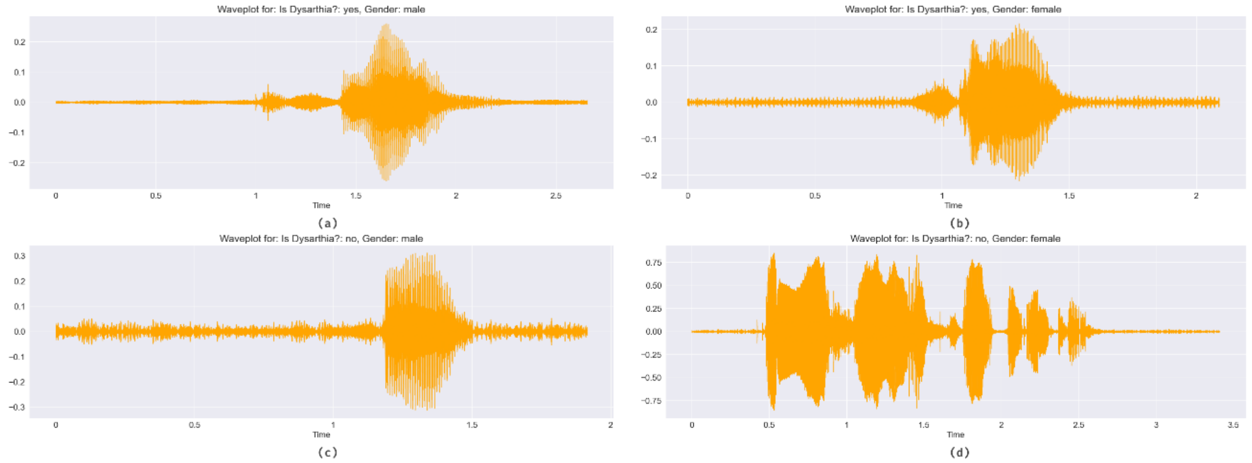
4. Empirical results and discussion

The performance analysis of the PD3F-RRNNT model is examined under the Dysarthria and Non-Dysarthria Speech dataset [32]. This dataset contains a total of 200 counts under dual class labels. The complete details of this dataset are presented in Table 2 below. The technique is simulated using Python 3.6.5 on a PC with an i5-8600k, 250GB SSD, GeForce 1050Ti 4GB, 16GB RAM, and 1TB HDD. Parameters include a learning rate of 0.01, ReLU activation, 50 epochs, a dropout rate of 0.5, and a batch size of 5.

Table 2. Details of the dataset.

Class labels	No. of count
Dysarthria	100
Non-dysarthria	100
Total count	200

Figure 3 illustrates the waveplots of the input speech signals. It can be seen that during tasks requiring minimal time, the presence of slurred speech is noticeable. This is also illustrated by the irregularities and reduced clarity in the waveform, which indicate the typical characteristics of dysarthric speech.



(a)

(b)

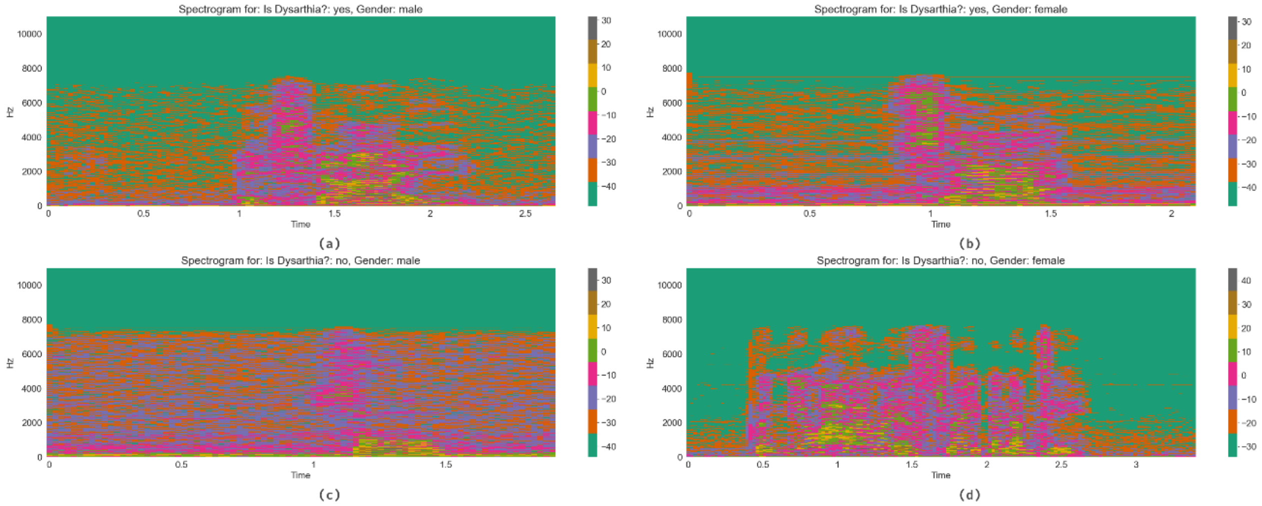
(c)

(d)

Figure 3. Waveplots graph of the PD3F-RRNNT model.

Figure 4 represents the Spectrogram graph of the PD3F-RRNNT approach. The following displays the spectrograms of the instances mentioned above.

- For dysarthric .wav files, the energy magnitudes of frequencies are more spread out across time, indicating slow and slurred speech, or because the words are coming out more rapidly and are overlapping with each other. Similar patterns may also be possible for an individual who is dysarthric and speaks monotonously,
- For .wav files that are not dysarthric, it is observed that the energy magnitudes are more concentrated in specific parts, which are regularly paced.



(a)

(b)

(c)

(d)

Figure 4. Spectrogram graph of PD3F-RRNNT model.

Figures 5–7 illustrate the confusion matrices and corresponding precision-recall (PR) and ROC curves across diverse training epochs (500–3000) in a model's ability to classify between dysarthria and non-dysarthria speech. Under epoch 500, the figure exhibits misclassification, with 13 wrongly predicted dysarthric samples. However, as training continues, specifically from epochs 2000 to 3000, the number of misclassifications decreases, reaching low prediction values at epoch 3000. The PR curves depict near-perfect PR values, approaching the ideal value of 1.0. This trend reflects the improved discrimination ability of the model, as it learns to more accurately separate the two classes, with both confusion matrices and PR curves validating the performance gain.

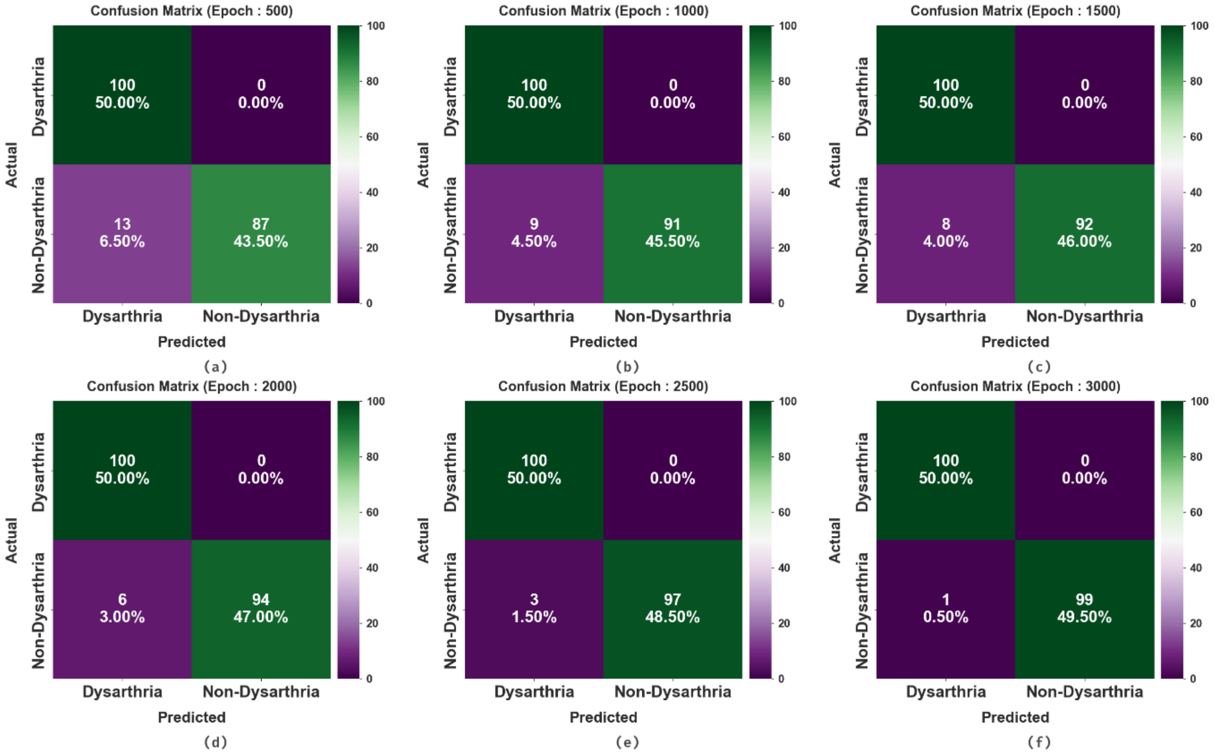


Figure 5. Confusion matrices of the PD3F-RRNNT technique illustrating classification performance at diverse training epochs.

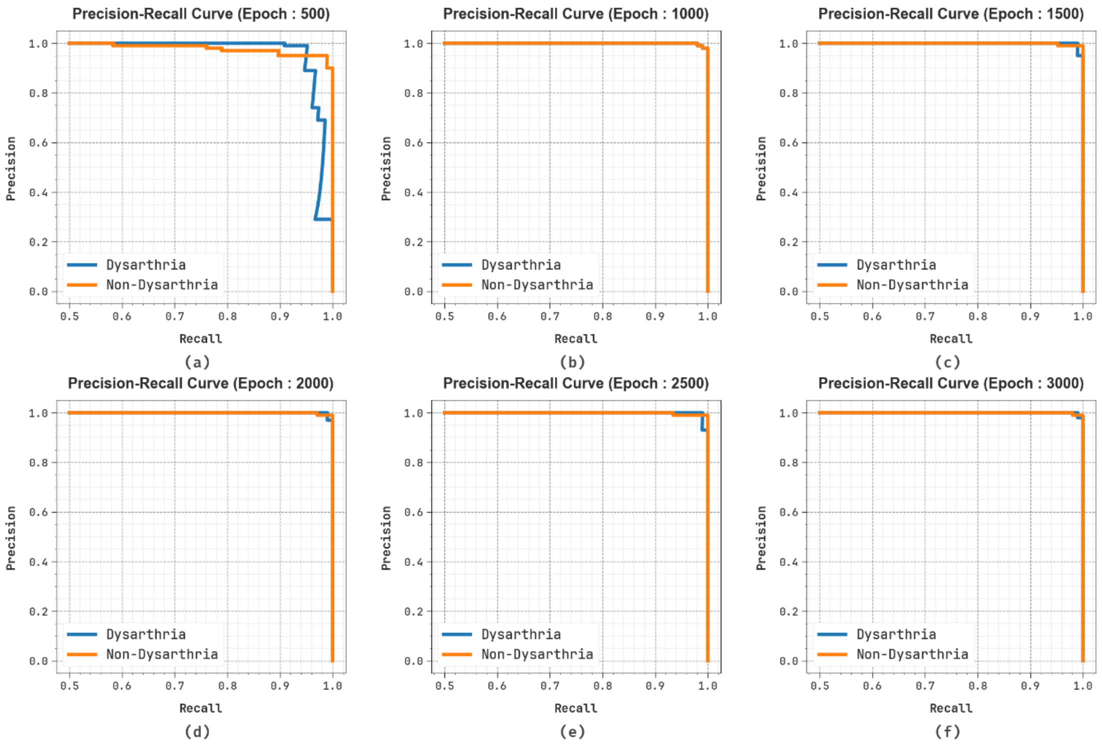


Figure 6. PR curves of the PD3F-RRNNT technique across epochs 500, 1000, 1500, 2000, 2500, and 3000.

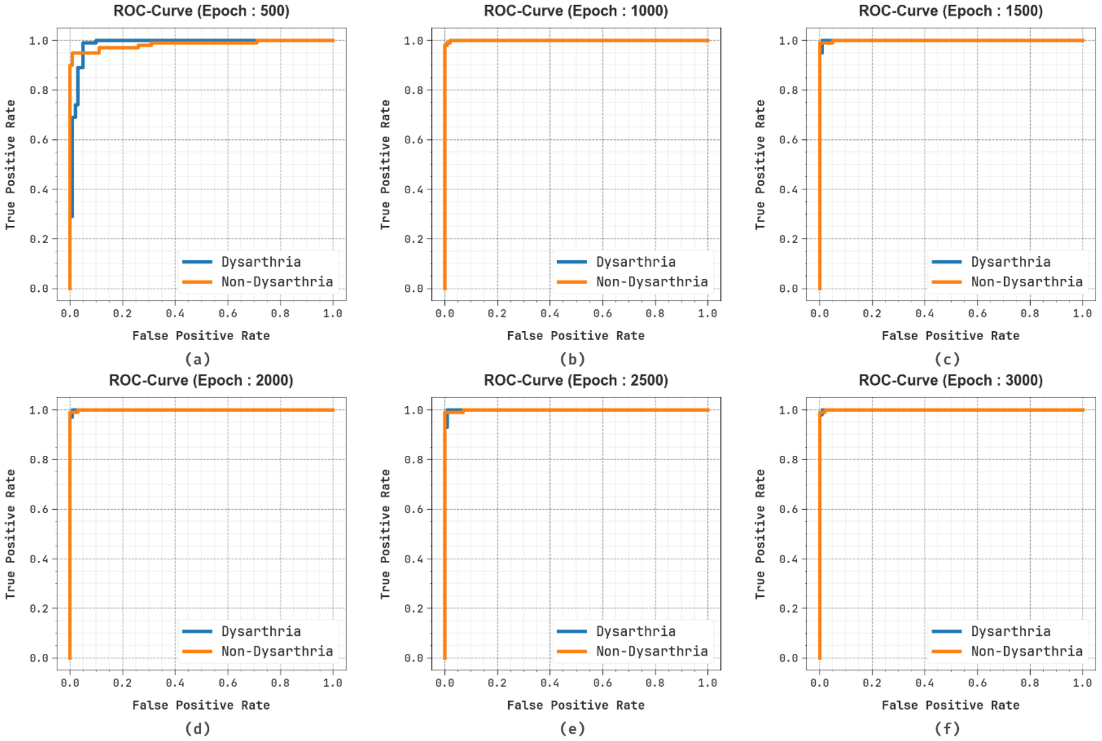


Figure 7. ROC curves of the PD3F-RRNNT technique across epochs 500, 1000, 1500, 2000, 2500, and 3000.

The speech disorder detection outcome of the PD3F-RRNNT model is illustrated under diverse epochs in Table 3 and Figure 8. Under 500 epochs, the PD3F-RRNNT model presents an average $accu_y$ of 93.50%, $prec_n$ of 94.25%, $reca_l$ of 93.50%, $F_{Measure}$ of 93.47%, and MCC of 87.74%. Similarly, at 1500 epochs, the PD3F-RRNNT technique presents an average $accu_y$ of 96.00%, $prec_n$ of 96.30%, $reca_l$ of 96.00%, $F_{Measure}$ of 95.99%, and MCC of 92.30%. Also, at 2500 epochs, the PD3F-RRNNT technique presents an average $accu_y$ of 98.50%, $prec_n$ of 98.54%, $reca_l$ of 98.50%, $F_{Measure}$ of 98.50%, and MCC of 97.04%. Finally, after 3000 epochs, the PD3F-RRNNT technique presents an average $accu_y$ of 99.50%, $prec_n$ of 99.50%, $reca_l$ of 99.50%, $F_{Measure}$ of 99.50%, and MCC of 99.00%.

Table 3. Speech disorder detection of the PD3F-RRNNT technique under distinct epochs.

Class labels	$Accu_y$	$Prec_n$	$Reca_l$	$F_{Measure}$	MCC
Epoch - 500					
Dysarthria	100.00	88.50	100.00	93.90	87.74
Non-dysarthria	87.00	100.00	87.00	93.05	87.74
Average	93.50	94.25	93.50	93.47	87.74
Epoch - 1000					
Dysarthria	100.00	91.74	100.00	95.69	91.37
Non-dysarthria	91.00	100.00	91.00	95.29	91.37
Average	95.50	95.87	95.50	95.49	91.37
Epoch - 1500					
Dysarthria	100.00	92.59	100.00	96.15	92.30
Non-dysarthria	92.00	100.00	92.00	95.83	92.30
Average	96.00	96.30	96.00	95.99	92.30
Epoch - 2000					
Dysarthria	100.00	94.34	100.00	97.09	94.17
Non-dysarthria	94.00	100.00	94.00	96.91	94.17
Average	97.00	97.17	97.00	97.00	94.17
Epoch - 2500					
Dysarthria	100.00	97.09	100.00	98.52	97.04
Non-dysarthria	97.00	100.00	97.00	98.48	97.04
Average	98.50	98.54	98.50	98.50	97.04
Epoch - 3000					
Dysarthria	100.00	99.01	100.00	99.50	99.00
Non-dysarthria	99.00	100.00	99.00	99.50	99.00
Average	99.50	99.50	99.50	99.50	99.00

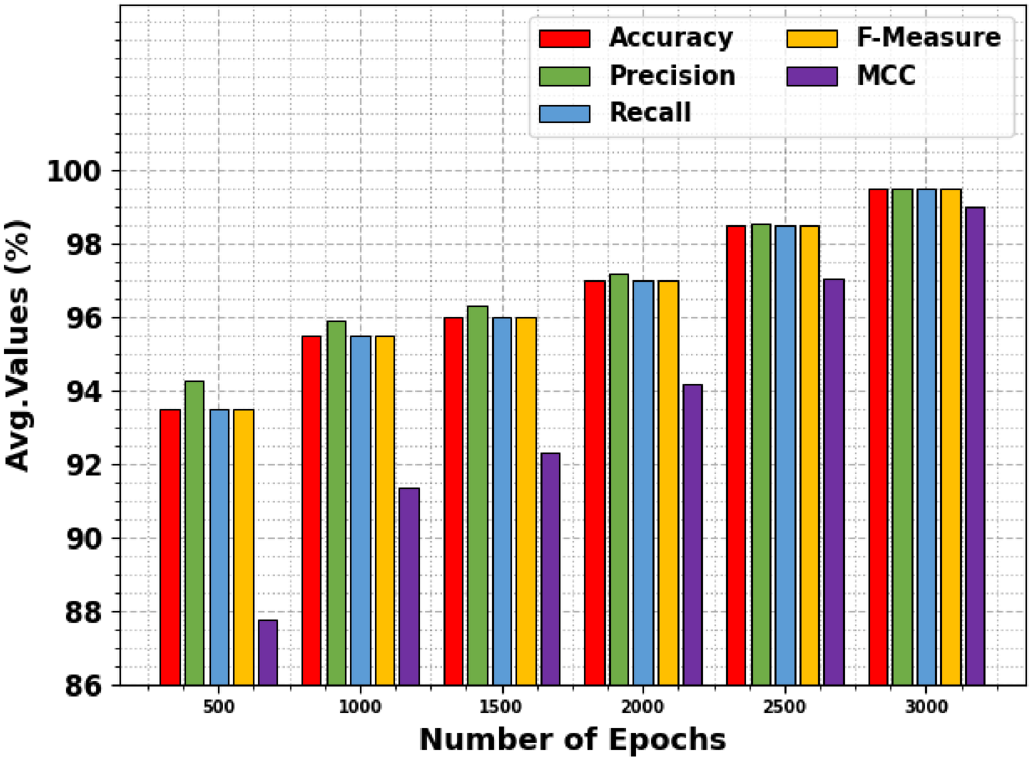


Figure 8. Average values of PD3F-RRNNT technique (a–f), Epochs 500–3000.

Figure 9 exemplifies the training (TRAIN) $accu_y$ and validation (VALID) $accu_y$ of a PD3F-RRNNT method at various epochs. At first, both TRAIN and VALID $accu_y$ rise quickly, denoting effective pattern learning from the data. Around the epoch, the VALID $accu_y$ minimally exceeds the training accuracy, indicating good generalization without overfitting. As training advances, it reflects maximum performance and a minimum performance gap between TRAIN and VALID. The close alignment of both curves in training suggests that the method is well-regularized and generalized. This determines the model's robust capability in learning and retaining valuable features across both seen and unseen data.

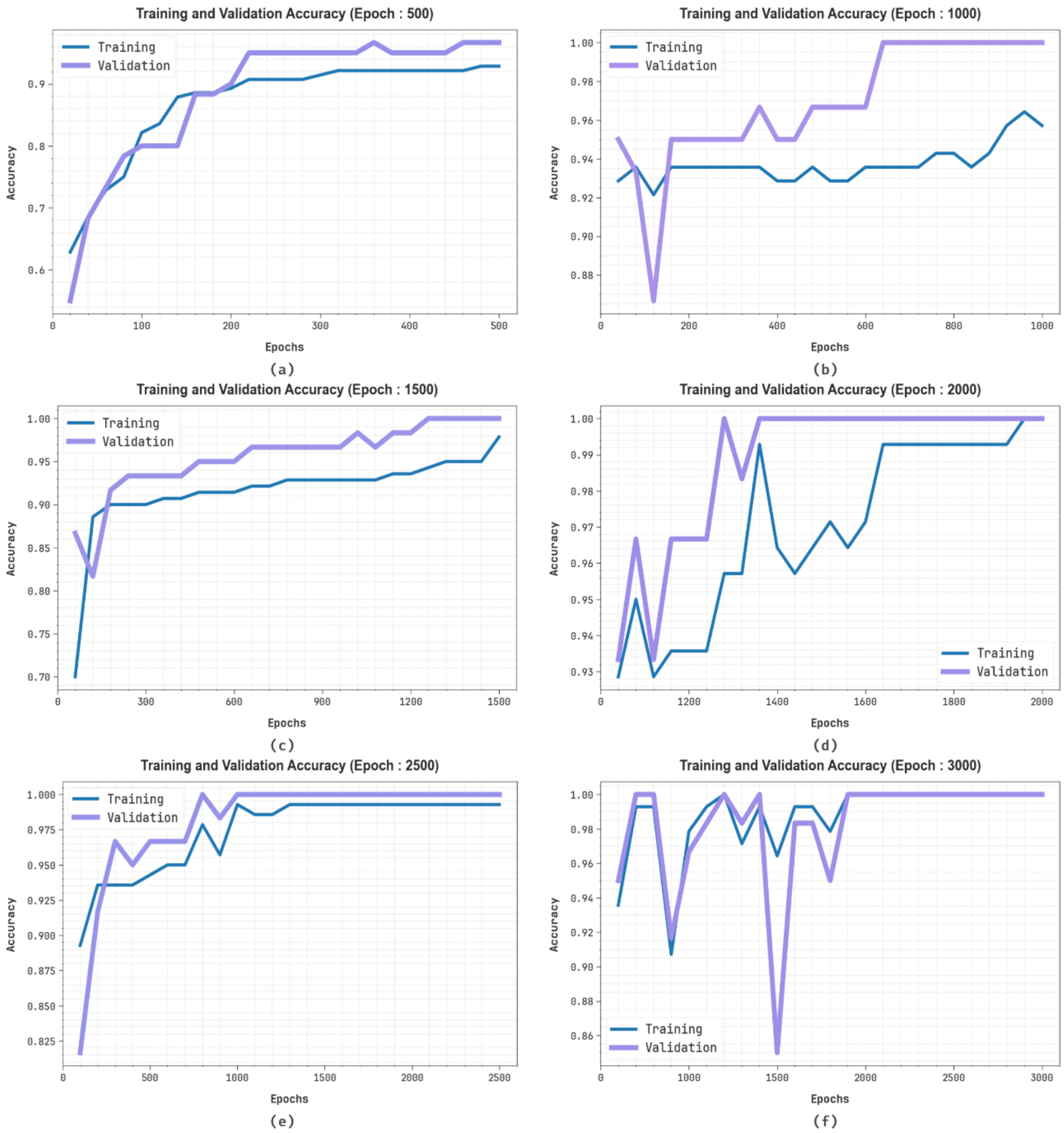


Figure 9. $Accu_y$ curve of PD3F-RRNNT method (a–f), Epochs 500–3000.

Figure 10 exemplifies the TRAIN and VALID losses of the PD3F-RRNNT model at various epochs. Initially, both TRAIN and VALID losses are higher, denoting that the model begins with a partial understanding of the data. As training evolves, both losses persistently reduce, displaying that the model is efficiently learning and updating its parameters. The close alignment between the TRAIN and VALID loss curves during training indicates that the model hasn't over-fitted and retains good generalization to unseen data. This persistent and steady reduction in loss reveals a stable, well-trained, and reliable DL model.

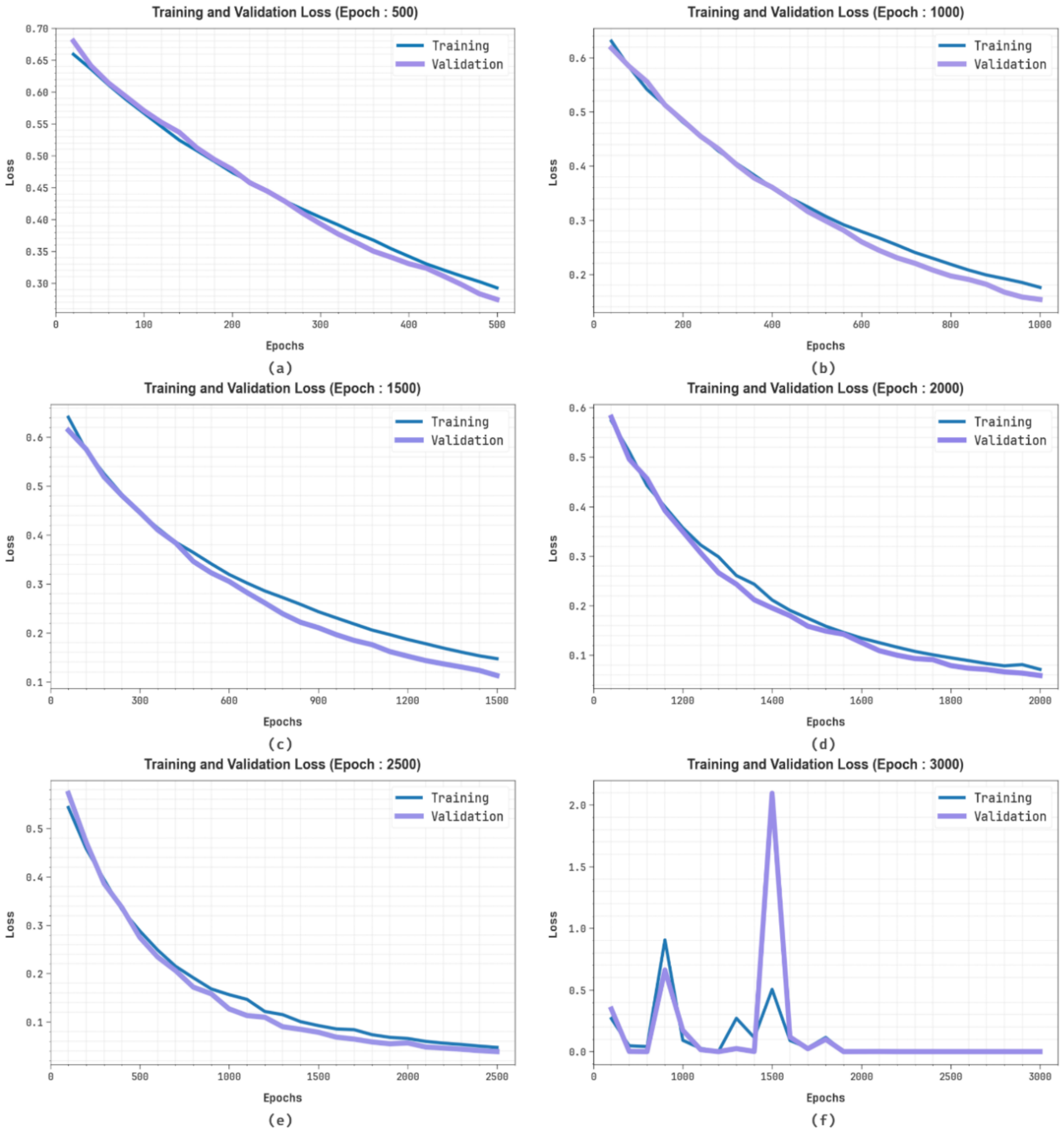


Figure 10. Loss curve of PD3F-RRNNT method (a–f), Epochs 500–3000.

Table 4 and Figure 11 present a comparative analysis of the PD3F-RRNNT methodology with existing techniques under various metrics [19, 33–35]. The outcomes underscored that the PD3F-RRNNT model attained the highest $accu_y$, $prec_n$, $reca_l$, and $F_{Measure}$ of 99.50%, 99.50%, 99.50%, and 99.50%, respectively. While the present methodologies, such as Wav2Vec 2.0, HuBERT, SEAMLESSM4T, MFCCs-CNN, XGBoost, KNN, MLP, deep autoencoder (DAE), Google ASR, and Transformer-CTC, have illustrated worse performance.

Table 4. Comparative study of PD3F-RRNNT methodology with existing techniques.

Methodology	$Accu_y$	$Prec_n$	$Reca_l$	$F_{Measure}$
Wav2Vec 2.0	93.99	92.24	95.41	89.02
HuBERT	89.97	89.98	91.30	92.15
SEAMLESSM4T	90.85	93.73	96.71	93.16
MFCCs-CNN	93.24	91.64	94.88	88.39
XGBoost	89.47	89.34	90.67	91.60
KNN	90.14	93.01	96.04	92.65
MLP	91.60	89.71	95.20	97.24
DAE	95.90	97.73	94.20	93.53
Google ASR	98.97	99.13	90.07	95.07
Transformer-CTC	98.46	95.27	88.37	90.66
PD3F-RRNNT	99.50	99.50	99.50	99.50

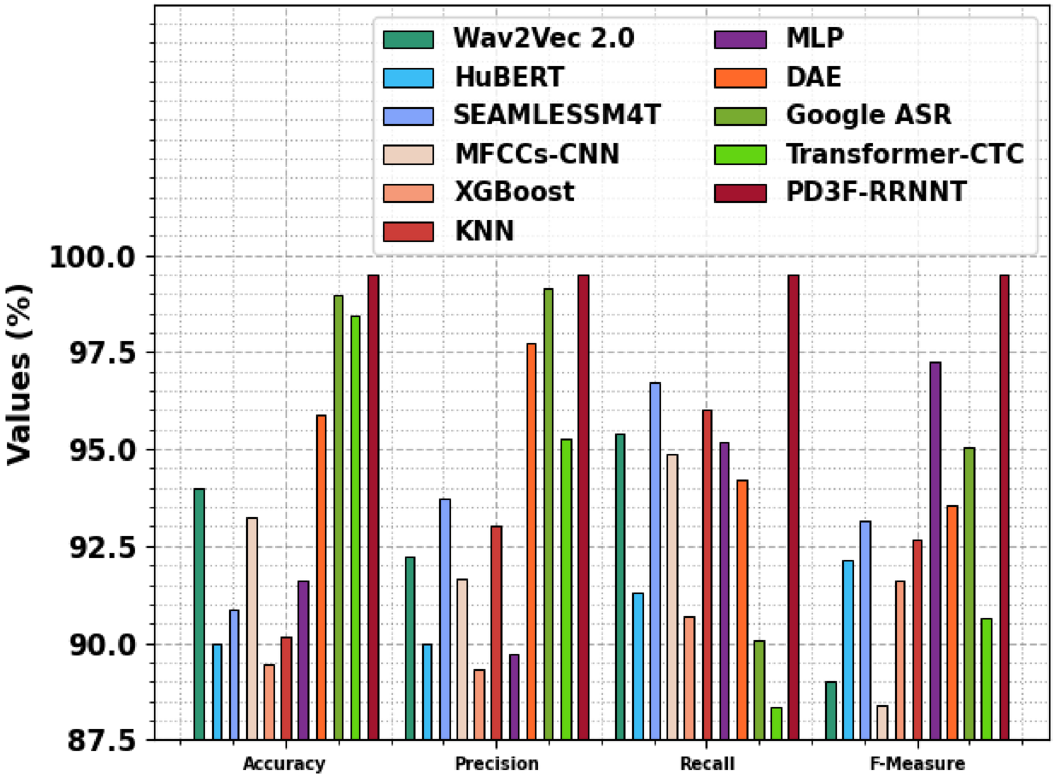


Figure 11. Comparative analysis of PD3F-RRNNT methodology with existing techniques.

In Table 5 and Figure 12, the execution time (EXT) of the PD3F-RRNNT method with existing models is illustrated. The PD3F-RRNNT model gives a lower EXT of 8.10sec, while the Wav2Vec 2.0, HuBERT, SEAMLESSM4T, MFCCs-CNN, XGBoost, KNN, MLP, DAE, Google ASR, and Transformer-CTC methodologies attained a superior EXT of 16.99sec, 18.09sec, 17.98sec, 23.88sec, 14.82sec, 14.28sec, 16.29sec, 24.45sec, 18.81sec, and 13.49sec, respectively.

Table 5. EXT outcome of PD3F-RRNNT model with existing methods.

Model	EXT (sec)
Wav2Vec 2.0	16.99
HuBERT	18.09
SEAMLESSM4T	17.98
MFCCs-CNN	23.88
XGBoost	14.82
KNN	14.28
MLP	16.29
DAE	24.45
Google ASR	18.81
Transformer-CTC	13.49
PD3F-RRNNT	8.10

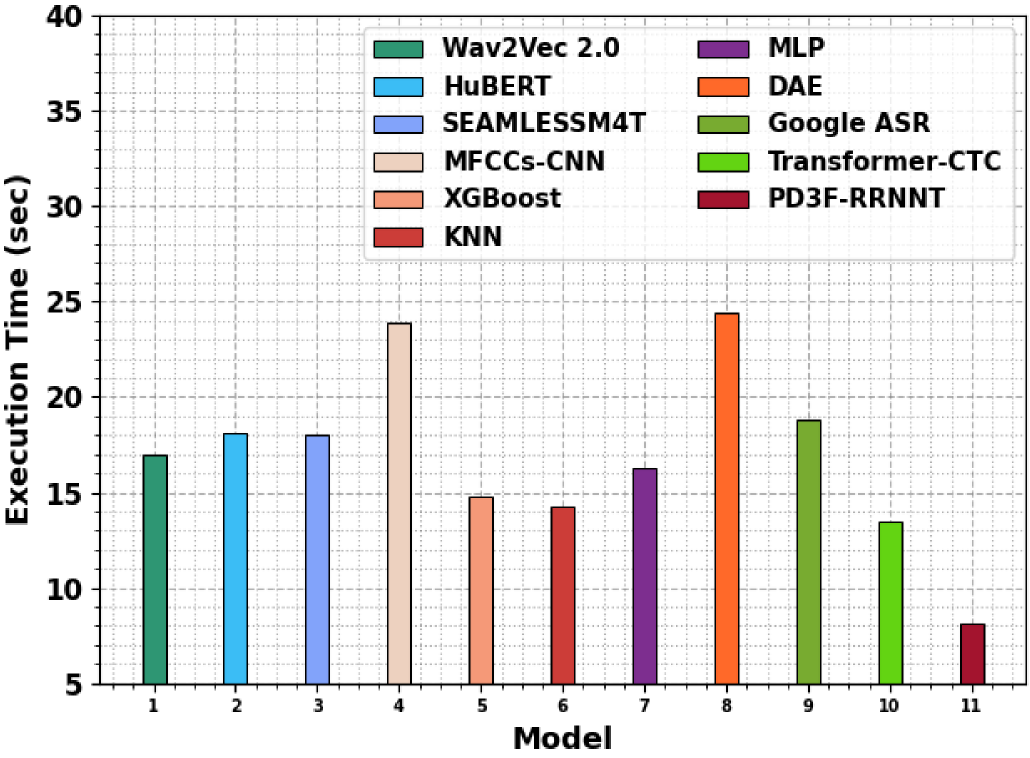


Figure 12. EXT outcome of PD3F-RRNNT model with existing methods.

Table 6 depicts the ablation study of the PD3F-RRNNT methodology. The PD3F-RRNNT methodology attained the highest performance with an $accu_y$ of 99.50%, $prec_n$ of 99.50%, $reca_l$ of 99.50%, and $F_{Measure}$ of 99.50%. By eliminating the transformer attention module in the TransAttUnet, the performance is slightly mitigated to an $accu_y$ of 98.67%, $prec_n$ of 98.72%, $reca_l$ of 98.72%, and $F_{Measure}$ of 98.79%. Furthermore, removing multi-scale skip connections additionally decreases the metrics, resulting in an $accu_y$ of 97.95%, $prec_n$ of 97.86%, $reca_l$ of 97.84%, and $F_{Measure}$ of 97.99%. By replacing the residual BiGRU (RBG) with a standard BiGRU, there is a significant performance drop, with an $accu_y$ of 97.23%, $prec_n$ of 97.01%, $reca_l$ of 97.06%, and $F_{Measure}$ of 97.29%. These results clearly emphasize the efficiency and necessity of each proposed component in achieving optimal model performance.

Table 6. Result analysis of the ablation study of PD3F-RRNNT methodology.

Methodology	$Accu_y$	$Prec_n$	$Reca_l$	$F_{Measure}$
PD3F-RRNNT (full model)	99.50	99.50	99.50	99.50
PD3F-RRNNT (without transformer attention module in TransAttUnet)	98.67	98.72	98.72	98.79
PD3F-RRNNT (without multi-scale skip connections)	97.95	97.86	97.84	97.99
PD3F-RRNNT (with a standard BiGRU instead of the residual BiGRU (RBG))	97.23	97.01	97.06	97.29

Table 7 specifies the computational efficiency and inference speed of the PD3F-RRNNT model against existing approaches [36]. The PD3F-RRNNT model illustrates the most balanced performance, attaining a low FLOPs count of 0.83G, GPU memory usage of 4540MB, and a fast inference time of 10.02 seconds. On the contrary, methods such as EfficientNet-B7 and EfficientNet-LEVIT required significantly higher computational resources, with FLOPs of 17.20 G and 18.50 G, respectively, and inference times of 29.57 seconds and 28.87 seconds. Likewise, MobileNet variants, while relatively lightweight, still illustrate higher FLOPs and slower inference compared to the PD3F-RRNNT approach. These results highlight the superior computational efficiency of the PD3F-RRNNT approach, making it an ideal choice for real-time and resource-constrained environments.

Table 7. Comparison of computational efficiency across methods, illustrating FLOPS, GPU memory usage, and inference time.

Technique	FLOPS (in Giga)	GPU (M)	Inference time (sec)
PD3F-RRNNT	0.83	4540	10.02
SDD-DLT	1.40	2534	19.51
EfficientNet-B7	17.20	4690	29.57
MobileNet-V2	15.80	2494	17.72
MobileNet-V1	19.50	4463	22.00
EfficientNet-B0	17.80	3607	22.96
EfficientNet-LEVIT	18.50	1004	28.87

5. Conclusions

In this manuscript, a novel PD3F-RRNNT approach is proposed. The PD3F-RRNNT approach aims to develop a real-time recognition method for accurately detecting dysarthria speech disorders in children, supporting early diagnosis and intervention. To achieve this, the PD3F-RRNNT technique involves audio data preparation, deep feature representation using TransAttUnet, and DL-based speech classification. Initially, the audio processing phase is employed at various levels, including VAD, noise elimination, pre-emphasis, framing and windowing, and normalization, to transform and extract significant data from audio signals. Furthermore, the PD3F-RRNNT method employs TransAttUnet for the feature extraction process. Ultimately, the RBG method is used to detect and classify speech disorders accurately. The experimental validation of the PD3F-RRNNT model is performed under the dysarthria and non-dysarthria speech dataset. The comparison analysis of the PD3F-RRNNT model revealed a superior accuracy value of 99.50% compared to existing techniques. The limitations of the PD3F-RRNNT model involve threats to model explainability, as the existing approaches lack techniques to clearly interpret decisions for clinical validation, which is crucial for gaining trust from healthcare professionals. Furthermore, the computational complexity of the model may impact its deployment on resource-constrained edge devices, such as smartphones, thereby limiting the feasibility of real-time applications and energy efficiency. The study also encountered limitations due to a relatively small dataset, which may restrict the generalizability of the results across diverse patient populations. Moreover, future work may focus on incorporating interpretable methods to enhance clinical transparency and optimizing the model for lower power consumption. Expanding the dataset with more varied speech samples and conducting extensive validation will additionally improve robustness and applicability in real-world scenarios

Author contributions

Ala Saleh Alluhaidan: Conceptualization, methodology, validation, investigation, writing-original draft preparation; Eman M Alanazi, Nasser Aljohani: Conceptualization, methodology, writing-original draft preparation, writing-review and editing; Amani A Alneil: Software, data creation, visualization, validation, writing-review and editing. All authors have read and agreed to the published version of the manuscript.

Use of Generative-AI tools declaration

The author declares that he has not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors extend their appreciation to the King Salman Center For Disability Research for funding this work through Research Group (No. KSRG-2024-243).

Data availability statement

The data that support the findings of this study are openly available in the Kaggle repository at https://www.kaggle.com/datasets/poojag718/dysarthria-and-nondysarthria-speech-dataset, reference number [32].

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] M. Shahin, U. Zafar, B. Ahmed, The automatic detection of speech disorders in children: challenges, opportunities, and preliminary results, *IEEE J.-STSP*, **14** (2020), 400–412. https://doi.org/10.1109/JSTSP.2019.2959393

[2] G. A. Attwell, K. E. Bennin, B. Tekinerdogan, A systematic review of online speech therapy systems for intervention in childhood speech communication disorders, *Sensors*, **22** (2022), 9713. https://doi.org/10.3390/s22249713

[3] A. Bhardwaj, M. Sharma, S. Kumar, S. Sharma, P.C. Sharma, Transforming pediatric speech and language disorder diagnosis and therapy: the evolving role of artificial intelligence, *Health Sciences Review*, **12** (2024), 100188. https://doi.org/10.1016/j.hsr.2024.100188

[4] A. Mytsyk, M. Pryshliak, Telepractice in the system of providing correctional and developmental services to children with speech disorders: interaction at a distance, *Journal of History Culture and Art Research*, **9** (2020), 94–105. http://doi.org/10.7596/taksad.v9i3.2674

[5] T. Sunderajan, S. V. Kanhere, Speech and language delay in children: prevalence and risk factors, *Journal of Family Medicine and Primary Care*, **8** (2019), 1642–1646. https://doi.org/10.4103/jfmpc.jfmpc_162_19

[6] A. N. Bhat, Motor impairment increases in children with autism spectrum disorder as a function of social communication, cognitive and functional impairment, repetitive behavior severity, and comorbid diagnoses: A SPARK study report, *Autism Res.*, **14** (2021), 202–219. https://doi.org/10.1002/aur.2453

[7] M. Jefferson, Usability of automatic speech recognition systems for individuals with speech disorders: past, present, future, and a proposed model, *University Digital Conservancy*, 2019. https://hdl.handle.net/11299/202757

[8] S. S. Liu, M. Z. Geng, S. K. Hu, X. R. Xie, M. Y. Cui, J. W. Yu, et al., Recent progress in the CUHK dysarthric speech recognition system, *IEEE-ACM T. Audio Spe.*, **29** (2021), 2267–2281. https://doi.org/10.1109/TASLP.2021.3091805

[9] M. Ur Rehman, A. Shafique, Q.-U.-A. Azhar, S. S. Jamal, Y. Gheraibia, A. B. Usman, Voice disorder detection using machine learning algorithms: An application in speech and language pathology, *Eng. Appl. Artif. Intel.*, **133** (2024), 108047. https://doi.org/10.1016/j.engappai.2024.108047

[10] I. El-Henawy, M. Abo-Elazm, Handling within-word and cross-word pronunciation variation for Arabic speech recognition (knowledge-based approach), *Journal of Intelligent Systems and Internet of Things*, **1** (2020), 72–79. https://doi.org/10.54216/JISIoT.010202

[11] Y. V. Sravya, K. Charishma, J. V. Narayana, K. Umasri, K. Sanjana, K. A. Pallavi, Autism spectrum disorder detection using hybrid machine learning and deep learning techniques, *Frontiers in Collaborative Research*, **3** (2025), 51–61. https://doi.org/10.70162/fcr/2025/v3/i1/v3i105

[12] V. R. Prabha, C. H. Bindu, K. R. Devi, An interpretable deep learning approach for autism spectrum disorder detection in children using NASNet-mobile, *Biomed. Phys. Eng. Express*, **11** (2025), 045006. https://doi.org/10.1088/2057-1976/addbe7

[13] J. J. Gao, S. T. Song, A hierarchical feature extraction and multimodal deep feature integration-based model for autism spectrum disorder identification, *IEEE J. Biomed. Health*, **29** (2025), 4920–4931. https://doi.org/10.1109/JBHI.2025.3540894

[14] C. B. Hu, J. Thrasher, W. Q. Li, M. D. Ruan, X. X. Yu, L. K. Paul, et al., Exploring speech pattern disorders in autism using machine learning, 2024, arXiv:2405.05126. https://doi.org/10.48550/arXiv.2405.05126

[15] M. Dia, G. Khodabandelou, A. Q. M. Sabri, A. Othmani, Video-based continuous affect recognition of children with autism spectrum disorder using deep learning, *Biomed. Signal Proces.*, **89** (2024), 105712. https://doi.org/10.1016/j.bspc.2023.105712

[16] H. A. Mengash, H. Alqahtani, M. Maray, M. K. Nour, R. Marzouk, M. A. Al-Hagery, et al., Automated autism spectral disorder classification using optimal machine learning model, *CMC-Comput. Mater. Con.*, **74** (2023), 5251–5265. https://doi.org/10.32604/cmc.2023.032729

[17] K. C. Raja, S. Kannimuthu, Deep learning-based feature selection and prediction system for autism spectrum disorder using a hybrid meta-heuristics approach, J. *Intell. Fuzzy Syst.*, **45** (2023), 797–807.

[18] A. Almadhor, R. Irfan, J. C. Gao, N. Saleem, H. T. Rauf, S. Kadry, E2E-DASR: End-to-end deep learning-based dysarthric automatic speech recognition, *Expert Syst. Appl.*, **222** (2023), 119797. https://doi.org/10.1016/j.eswa.2023.119797

[19] O. Klempir, A. Skryjova, A. Tichopad, R. Krupicka, Ranking pre-trained speech embeddings in Parkinson's disease detection: Does Wav2Vec 2.0 outperform its 1.0 version across speech modes and languages, *Comput. Struct. Biotec.*, **27** (2025), 2584–2601. https://doi.org/10.1016/j.csbj.2025.06.022

[20] K. Attaluri, A. Chvs, S. Chittepu, Empowering dysarthric speech: Leveraging advanced LLMs for accurate speech correction and multimodal emotion analysis, 2024, arXiv:2410.12867. https://doi.org/10.48550/arXiv.2410.12867

[21] R. Mahum, I. Ganiyu, L. Hidri, A. M. El-Sherbeeny, H. Hassan, A novel Swin transformer based framework for speech recognition for dysarthria, *Sci. Rep.*, **15** (2025), 20070. https://doi.org/10.1038/s41598-025-02042-7

[22] S. I. Ng, C. W. Y. Ng, J. R. Wang, T. Lee, Automatic detection of speech sound disorder in Cantonese-speaking pre-school children, *IEEE-ACM T. Audio Spe.*, **32** (2024), 4355–4368. https://doi.org/10.1109/TASLP.2024.3463503

[23] S. R. Shahamiri, K. Mandal, S. Sarkar, Dysarthric speech recognition: an investigation on using depthwise separable convolutions and residual connections, *Neural Comput. & Applic.*, **37** (2025), 7991–8005. https://doi.org/10.1007/s00521-024-10870-3

[24] S. S. Sung, J. So, T. J. Yoon, S. Ha, Automatic detection of speech sound disorder in children using automatic speech recognition and audio classification, *Phonetics Speech Sci.*, **16** (2024), 87–94. https://doi.org/10.13064/KSSS.2024.16.3.087

[25] M. Manoswini, B. Sahoo, A. Swetapadma, A novel speech signal feature extraction technique to detect speech impairment in children accurately, *Comput. Biol. Med.*, **195** (2025), 110681. https://doi.org/10.1016/j.compbiomed.2025.110681

[26] G. Kim, Y. Eom, S. S. Sung, S. Ha, T. J. Yoon, J. So, Automatic children speech sound disorder detection with age and speaker bias mitigation, *Proc. Interspeech*, **2024** (2024), 1420–1424. https://doi.org/10.21437/Interspeech.2024-1799

[27] J. Mun, S. Kim, M. Chung, A cascaded multimodal framework for automatic social communication severity assessment in children with autism spectrum disorder, *Proc. Interspeech*, 2025 (2025), 3055–3059. https://doi.org/10.21437/Interspeech.2025-726

[28] A. Ziani, A. Adouane, M. N. Amiri, S. Smail, New proposed solution for speech recognition without labeled data: tutoring system for children with autism spectrum disorder, *Informatica*, **48** (2024), 109–122. https://doi.org/10.31449/inf.v48i18.5204

[29] M. Labied, A. Belangour, M. Banane, A. Erraissi, An overview of automatic speech recognition pre-processing techniques, *2022 International Conference on Decision aid Sciences and Applications (DASA)*, Chiangrai, Thailand, 2022, 804–809. https://doi.org/10.1109/DASA54658.2022.9765043

[30] B. Z. Chen, Y. S. Liu, Z. Zhang, G. M. Lu, A. W. K. Kong, Transattunet: Multilevel attention-guided u-net with Transformer for medical image segmentation, *IEEE Transactions on Emerging Topics in Computational Intelligence*, **8** (2024), 55–68. https://doi.org/10.1109/TETCI.2023.3309626

[31] X. W. Mou, Y. F. Song, X. P. Xie, M. X. You, R. J. Wang, Res-RBG facial expression recognition in image sequences based on dual neural networks, *Sensors*, **25** (2025), 3829. https://doi.org/10.3390/s25123829

[32] *Dysarthria and non-dysarthria speech dataset*, 2022. Available from: https://www.kaggle.com/datasets/poojag718/dysarthria-and-nondysarthria-speech-dataset.

[33] F. Javanmardi, S. R. Kadiri, P. Alku, Pre-trained models for detection and severity level classification of dysarthria from speech, *Speech Commun.*, **158** (2024), 103047. https://doi.org/10.1016/j.specom.2024.103047

[34] A. Al-Ali, S. Al-Maadeed, M. Saleh, R. C. Naidu, Z. C. Alex, P. Ramachandran, The detection of dysarthria severity levels using AI models: A review, *IEEE Access*, **12** (2024), 48223–48238. https://doi.org/10.1109/ACCESS.2024.3382574

[35] R. Vinotha, D. Hepsiba, L. D. V. Anand, P. M. Bruntha, L. Dinh, H. Dang, Empowering dysarthric communication: hybrid transformer-CTC based speech recognition system, *IEEE Access*, **13** (2025), 82479–82491. https://doi.org/10.1109/ACCESS.2025.3568342

[36] N. A. Aljarallah, A. K. Dutta, A. R. W. Sait, Image classification-driven speech disorder detection using deep learning technique, *SLAS Technol.*, **32** (2025), 100261. https://doi.org/10.1016/j.slast.2025.100261