AIMS Mathematics, 11(8): 25100–25122.
DOI: 10.3934/math.20261009
Received: 30 June 2026
Revised: 22 July 2026
Accepted: 31 July 2026
Published: 14 August 2026

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

Research article

Artificial intelligence with vision transformer based hybrid deep neural network for paediatric speech disorder recognition using dysarthric voice data

Fahd N. Al-Wesabi[1,*] and Wafi Bedewi[2,3]

[1] Department of Computer Science, Applied College at Mahayil, King Khalid University, Saudi Arabia
[2] Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
[3] King Salman Centre for Disability Research, Riyadh, Saudi Arabia

* **Correspondence:** Email: falwesabi@kku.edu.sa.

Abstract: Speech disorders among children will change their intelligibility and fluency. Dysarthria is one of the speech disorders connected to problems that control the muscles required to speak. Children with dysarthria frequently have low, abnormal breath that initiates problems in producing adequate breath to help speaking. They have lower-pitched, harsh, nasal speech, and extremely poor pronunciation. Furthermore, these problems make the speech of children harder to recognize. Dysarthria is generated by neurological loss and occurs earlier in the lives of children, from neurological loss prolonged earlier, like in cerebral palsy, or earlier childhood over brain damage or nervous disorders. Currently, deep learning (DL)-based methods are applied to recognize dysarthria speech disorder. In this study, we developed an Efficient Framework for the Dysarthria Speech Disorder Recognition Technique of Children Using a Hybrid Deep Learning Model (EFDSDRTC-HDLM). Our aim of the EFDSDRTC-HDLM model was to design an effective recognition technique for identifying dysarthria speech disorders in children using advanced speech processing and classification methods. The proposed framework first converted speech signals into time-frequency images through speech activity detection, framing, windowing, and time-frequency image generation. These visual representations were processed by a Vision Transformer to extract discriminative global features, which were subsequently refined by a hybrid Graph Convolutional Network–Bidirectional Long Short-Term Memory with Multi-Head Attention architecture to capture structural relationships,

temporal dependencies, and salient feature representations for accurate dysarthria classification. The efficiency of the proposed framework was evaluated through extensive experiments on a benchmark dataset. Experimental results demonstrated that the proposed model achieves superior scalability and classification performances, outperforming traditional DL models as well as recent transformer-based speech recognition methods across evaluation metrics, accuracy, precision, recall and f-score of 98.73%, 98.4%, 98.73%, and 98.56%, respectively, confirming its effectiveness for pediatric dysarthria speech recognition.

Keywords: dysarthria; speech disorder; hybrid deep learning; multi-head attention mechanism; vision transformer; spectrogram
Mathematics Subject Classification: 68T07, 68T45

1. Introduction

Dysarthria refers to speech disorders that result from neuromuscular deficiencies. Individuals with dysarthria struggle to control and coordinate the speed, extent, force, and timing of movements necessary for speaking, making their speech difficult to understand. Individuals with serious physical conditions often exhibit motor impairments and dysarthria [1]. In children, dysarthria is linked to congenital conditions like cerebral palsy (CP) and acquired causes, namely brain tumors and traumatic brain damage. Nearly 20% of children diagnosed with CP show speech that is influenced by dysarthria. In children, the speech disorders impact their clarity and fluency [2]. A delay in identifying and treating these disorders raises the possibility of social difficulties and academic challenges. Comparatively, dysarthria is a neuromuscular issue of motor performance caused by irregularities in strength, tone, mobility, or accuracy of movements needed for suitable regulation of the speech systems [3]. Sadly, the varied causes don't produce distinct speech symptom profiles, and there is a resemblance among the speech traits connected to each condition. As a result, diagnosing dysarthria accurately is highly difficult.

Although several speech pathologists can effortlessly recall dysarthria's theoretical definitions, a structured method for accurately distinguishing these disorders in children is lacking [4]. Moreover, dysarthria is identified in 4.9% of children with intricate neurodevelopmental conditions and appears over 4 times more frequently in children with specific conditions, like Down syndrome [5]. Dysarthria symptoms can differ based on the root cause and intensity, possibly leading to speech that is extremely unclear or mildly slurred as the disease advances. Because dysarthric speech can be highly unclear, an average listener may struggle to communicate with dysarthric individuals unless they have previous experience with such speakers [6]. This greatly limits the communication abilities of dysarthric individuals with typical speakers. Additionally, dysarthria frequently co-occurs with neurological disorders; therefore, many individuals with dysarthria also experience physical impairments, making it difficult or unfeasible to use digital tools and computers through mouse, keyboard, or touch-based input [7].

There is rising attention on the employment of deep learning (DL) for the automatic detection of speech and voice disorders. DL is a kind of machine learning (ML), which deploys artificial neural networks (ANNs) to attain understanding from data [8]. They can be utilized for learning complex data patterns, like the connections among various features. Moreover, DL models have attained major

interest and excelled in several domains [9]. DL approaches can be trained to acquire the acoustic features that differentiate disordered and normal speech. Additionally, DL models can transform the speech sound disorder detection domain by providing automatic and unbiased methods for detection [10].

In this study, we develop an Efficient Framework for the Dysarthria Speech Disorder Recognition Technique of Children Using a Hybrid Deep Learning Model (EFDSDRTC-HDLM). Primarily, the pre-processing phase is applied to audio files converted into time-frequency images by detecting active speech regions, applying framing and windowing, and using a time-frequency image generator to provide visual inputs for further analysis. Afterward, the extraction of the feature process is executed by the Vision transformer (ViT) method. Last, the proposed EFDSDRTC-HDLM model implements a hybrid of graph convolutional network and bidirectional long short-term memory with a multi-head attention mechanism (G-BiL-MHA) for the classification method. The mathematical outcome signifies that the EFDSDRTC-HDLM system has improved scalability and performance in terms of several evaluations over the recent methodologies.

Despite the advances in DL for speech disorder recognition, identifying the pediatric dysarthric speech accurately remains challenging because of variations in characteristics of speech, limited feature representation, temporal dependencies, and the difficulties in simultaneously capturing the global contextual information. Approaches usually rely on a single DL architecture, resulting in reduced classification performance and robustness. Therefore, our primary aim of this study is to present an effective hybrid DL framework for pediatric dysarthria speech recognition by combining Multi-Head Attention mechanism, Bidirectional Long Short-Term Memory network and Vision Transformer-based feature extraction with a Graph Convolutional Network.

2. Literature review on dysarthria speech analysis

Aljarallah et al. [11] introduced an image classification-based automatic SDD approach for Mel-Spectrogram classification to recognize numerous speech illnesses. At first, the Wavelet Transform (WT) hybridization method has been used to produce a Mel-Spectrogram utilizing the voice samples. Zhang et al. [12] proposed a DSS paradigm called the Long-range and Non-stationary Variational Autoencoder (LNVAE) model. This model predicts dysarthric speech's acoustic parameters by encoding the long-range phonemes' dependencies in frame-level dysarthric speech representations. Furthermore, the LNVAE deploys the Gaussian noise perturbation in the latent variables to extract non-stationary variations in dysarthric speech. In [13], an affordable, portable, and tailored augmentative and another speech communication help personalized according to the dysarthria speaker's needs was designed. The dysarthria speech detection method has been trained through transfer learning (TL) method, with normal speakers' speech data as the resource system. The data augmentation is executed through a virtual microphone and a multiresolution-related feature extractor methodology (VM-MRFE). Despite these methods having revealed promising performance in dysarthria speech recognition, they rely on single model architectures or conventional feature extraction, which may not efficiently capture temporal dependencies and global contextual information. Consequently, their classificational performance and generalization capacity remains limited for complicated pediatric dysarthric speech patterns.

Usha [14] presented a new technique using IoT and sophisticated data mining methods for detecting speech disorders in children. Here, the incorporation of data mining and IoT technology for timely speech disorder identification in children is examined by implementing cutting-edge models.

The motive is to design an effective device for initial identification, improving speech therapy results by practical analysis and monitoring. In [15], an EL enables the combination of several predictions of the model to produce an optimum result. Therefore, an EL-driven SDD approach is presented in this work. The feature engineering relies on ResNet 18 architecture to extract vital features from the MS. XGBoost and CatBoost techniques are leveraged for feature classification. The results of these techniques are utilized for training the support vector machine (SVM) framework for final prediction. Studies have demonstrated that integrating complementary DL architectures and multi-modal feature fusion methods can significantly enhance representation learning for complicated prediction tasks. Spatiotemporal DL models employing multi source data fusion have shown superior capability in getting local structural and global contextual information. These findings motivate the development of hybrid frameworks that combine transformer-based feature extraction with structural and temporal learning mechanisms for enhanced speech disorder recognition.

Mittal et al. [16] focused on dysarthria, a speech impairment typically seen in people with illnesses like amyotrophic lateral sclerosis (ALS) or CP. Initial recognition of dysarthria is vital for efficient interventions and enhanced patient outcomes because it considerably affects communication capabilities. Utilizing the TORGO database, encompassing a vast quantity of samples from individuals with and without dysarthria, and comprising different genders, this research exploits a Convolutional Neural Network (CNN) to classify speech. Lin et al. [17] presented a cognition-related feature decomposition and recombination network (CFDRN) for dysarthria ASR. The slower and faster temporal processors were created for feature decomposition into changeable and stable features. Furthermore, this research employed an adaptation method that depends on unsupervised pre-training models for reducing the data scarcity problems. The CFDRNs were included, and the complete model was modified from normal speech to disordered speech. In [18], an effective dysarthria speech recognition technique was presented utilizing the designed Competitive Crow Search Algorithm-driven Hierarchical Attention Network (CCSA-HAN). The spectral subtraction scheme was deployed to remove unnecessary noise. Following that, feature extraction is performed, and to enhance performance, the data augmentation process takes place. Then, it is executed by inserting several noises, such as train, street, and party crowd noise into an input signal.

Wu et al. [19] demonstrated Semi-supervised learning (SSL) construct classifiers utilizing labeled and unlabeled data. It gets information from labeled samples, which is often costly or labor intensive, to enhance prediction performance. This defines an incomplete data problem, which can be statistically formulated within the framework for finite mixture models, which can be fitted using the expectation-maximization (EM) algorithm. Xu et al. [20] highlighted the effectiveness of integrating ViT and GCN architectures in medical image diagnosis, especially in addressing challenges in small datasets. This method can lead to more efficient and accurate pneumonia diagnoses, especially in resource constrained settings where the small datasets are common.

Comparative evaluation on dysarthria speech disorder recognition is shown in Table 1.

Table 1. A comparison study of advanced methods.

Author	Objective	Method	Dataset	Performance Analysis
Aljarallah et al. [11]	To improve speech disorder classification and provide innovative features to develop accurate, accessible, and effectual diagnostic tools	LEVIT, CatBoost, XGBoost, STFT, and WT	VOICED and LANNA datasets	Accuracy of 99.1
Zhang et al. [12]	To utilize the Gaussian noise perturbation inside the latent variables to acquire the non-stationary fluctuation in dysarthric speech	Variational Autoencoder	CDSD Chinese and UASpeech English datasets	Objective metrics of 29 and 28
Vijayalakshmi et al. [13]	To improve the intelligibility of speech, specifically for individuals with impairment of speech like dysarthria, to employ speech synthesis and speech recognition methods	Hidden Markov Model	-	Speech delivery rate of 4.4 s
Usha [14]	To advance an effectual tool for intervention and quick identification, improving speech therapy solutions over real-world analysis and monitoring	Internet of Things (IoT)	-	-
Dutta and Wahab Sait [15]	To increase diagnostic models often based on subjective assessments through speech-language pathologists and regarding complexity in scalability, consistency, and availability	ResNet 18, CatBoost, XGBoost, and Support Vector Machine (SVM).	VOICED dataset	Accuracy of 99.7%
Mittal et al. [16]	To develop diagnostic proficiencies and projects a promising way to increase the qualities of life for individuals dealing with speech difficulty	CNN and ResNet18	-	Accuracy of 96%
Lin et al. [17]	To advance speaker-independent dysarthric speech detection. To lessen the scarcity of data concerns, to employ unsupervised pre-training models for initializing the method and modify it from normal to dysarthric speech	Gated fusion modules (GFM), CNN	TORGO and UASpeech datasets	Word error rates of 13.73%∼16.23% and 4.50%∼13.20%
Jolad and Khanai [18]	An efficient dysarthria speech recognition model based on the Hierarchical Attention Network and the spectral subtraction model has been employed to extract unwanted noises to enhance the performance of data augmentation	Competitive Swarm Optimiser (CSO) and Crow Search Algorithm (CSA)	-	Accuracy of 0.9141, sensitivity of 0.9208, and specificity of 0.9172

2.1 Research gap

Although Aljarallah et al. [11] presented the effectiveness of Mel-spectrogram-based image classification, the given method mainly concentrated on image representations without utilizing complementary temporal dependencies. Zhang et al. [12] discussed long range phoneme dependencies by a variational autoencoder; still, structural feature relationships were not examined. Vijayalakshmi et al. [13] enhanced dysarthria speech detection using feature augmentation, but its performance highly depended on pre-trained speech representations. Usha [14] emphasized IoT-based speech disorder detection, while [15] used an ensemble learning method based on SVM, CatBoost, XGBoost, and ResNet-18; however, these methods did not combine transformer-based feature learning with temporal sequence modeling. Similarly, Mittal et al. [16] employed a CNN for dysarthria classification, Lin et al. [17] implemented CFDRN for feature decomposition, and Jolad and Khanai [18] implemented hierarchical-attention for speech recognition; still, these approaches did not integratedly exploit temporal dependencies, graph-structured feature relationships, and global contextual information within a unified architecture. Therefore, there remains a demand for an integrated hybrid framework that efficiently combines Multi-Head Attention, BiLSTM, Vision Transformer-based feature extraction with Graph Convolutional Networks to enhance paediatric dysarthria speech recognition robustness and accuracy.

3. Proposed system architecture

In this study, we develop an EFDSDRTC-HDLM. The intention of the EFDSDRTC-HDLM technique is to design an effective recognition technique for identifying dysarthria speech disorders in children using advanced speech processing and classification methods. It contains three kinds of processes involved pre-processing, extraction of features, and classification method. Figure 1 shows the workflow of the EFDSDRTC-HDLM model.

3.1. Speech pre-processing

In the pre-processing stage, the audio files are transformed into time-frequency images by detecting active speech regions, applying framing and windowing, and using a time-frequency image generator to provide visual inputs for further analysis. In pre-processing speech signals, speech areas are recognized inside the raw speech WAV file [21]. It is fed with a sampling frequency (Fs), speech file, a Hamming window with 0.04xFs dimension, and a combined 0.05xFs distance. Afterward, the speech part is recognized around the file of WAV. These specified segments remove the required factors of speech filter out unnecessary segments of speech signal and identify speech of dysarthria signal. Figure 2 specifies the flow of speech preprocessing.

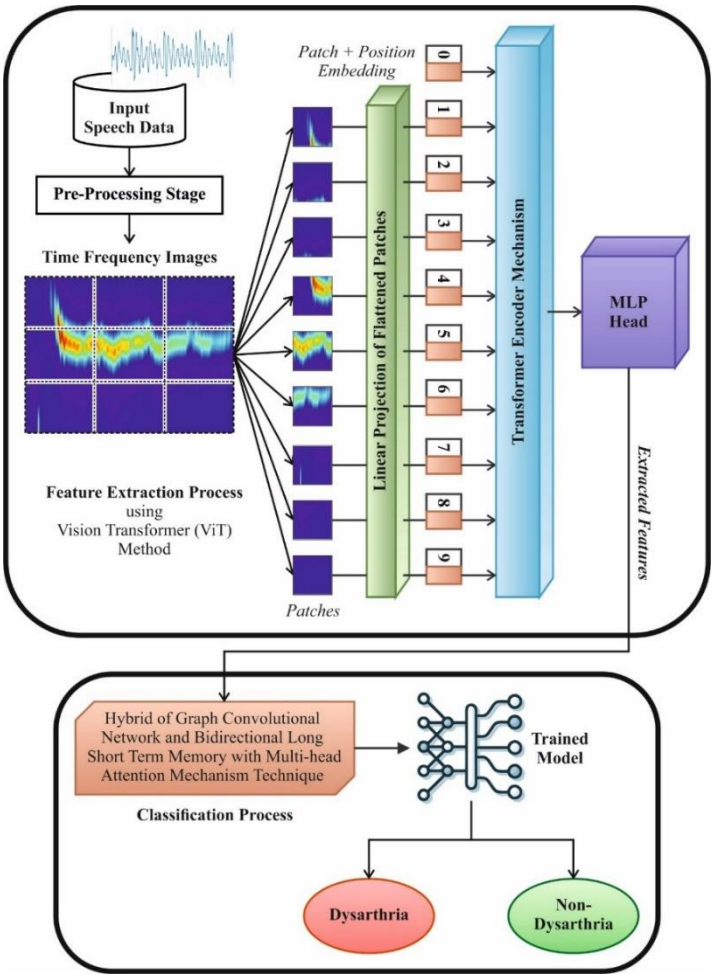

Figure 1. Workflow of the EFDSDRTC-HDLM model.

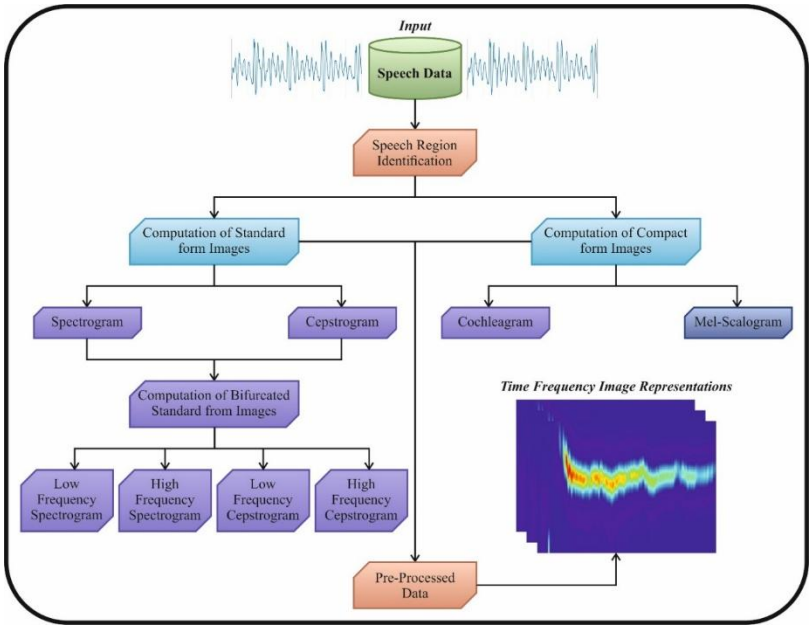

Figure 2. Flow of speech pre-processing.

3.1.1. Calculation of the spectrogram

Spectrograms have two-dimensional visual, which represents an order of spectra. They are frequently created to employ Short-Time Fourier Transform (STFT) with time revealed to single axis, frequency among others, and color or brightness for signifying the sturdiness of an element with specified frequency at every time frame. It comprises a sequence of FTs utilized to window signal and offers time-localized frequency data to examine the frequency elements of signal modification through time.

$$X_{STFT}[a,b] = \sum_{k=1}^{L-1} x\,[k]g[k-m]e^{-\frac{j2\pi nk}{L}}. \tag{1}$$

Here, $g[k]$ indicates a window function of L-point and $x[k]$ signifies signal. The STFT of $x[k]$ interprets FT of mathematical products, $x[k]g[k-m]$. Thus, spectrogram recommends particular CNN frameworks intended for relevant sound. A speech spectrogram depicts the FT of signal and evolutions. The spectrogram is originated utilizing a window dimension of 0.016 times Fs with a 0.08 times Fs, and image dimension of 256xN, where N indicates the overall speech frame counts with overlapping. To examine the behavior of spectrum, its image is bi-furcated into dual elements: high- and low-frequency components. The bi-furcation method dimension of spectrogram images is 128xN pixels.

3.1.2. Cepstrogram calculation of the speech signal

This is a graph model of a speech signal exhibiting the cepstrum modifications through frequency and time. To employ DCT, the spectrogram outcomes in the cepstrogram. An intrinsic function is employed for computing the cepstrogram with a window dimension of 0.016 times Fs, the periodical overlapping of 0.08 times Fs, and a FFT length of 512, which creates an image dimension of 256xN. This is an effectual mode, which employs homomorphic data to accomplish challenges, comprising speaker recognition, speech recognition, pitch detection, and speech emotion recognition. To emphasize the advantages of cepstral investigation, we analyze the condition that the signal includes an echo.

$$y[n] = k[n] + \alpha x[n-\tau]. \tag{2}$$

This depicts the magnitude spectrum of the overall signal.

$$Y(\omega) = K(\omega)(1 + 2\alpha + 2\alpha cos(\omega\tau))^{\frac{1}{2}}. \tag{3}$$

To capture the log of magnitude spectrum, it is probable to partition the spectrum into individual elements.

$$log\big(Y(\omega)\big) = log\big(K(\omega)\big) + \frac{1}{2log\big(1 + 2\alpha + 2\alpha cos(\omega\tau)\big)}. \tag{4}$$

Here, $\log\big(Y(\omega)\big)$ signifies interpreted primary signal $\log(K(\omega))$, with $1/2$ $\log\big(1+\alpha 2+2\alpha\cos(\omega\tau)\big)$. The cepstrogram of signal $y[n]$ has been acquired to employ iscrete cosine transform (DCT) to $\log(Y(\omega))$. The bi-furcation procedure of cepstrogram is similar to the spectrogram process of bi-furcation.

3.1.3. Evaluation of mel-scalogram

A Mel-scalogram is a scalogram methodology in the scale of Mel-frequency. To calculate it, the spectrogram is attained, and a 40 Mel scale using the Mel-filter bank is multiplied by the gained spectrogram. Employing CWT in the Mel-spectrogram results in assessing the Mel-scalogram images in the dimension 40xN.

3.1.4. Assessment of the cochleagram

A cochleagram employs the time-frequency signal domain. Nevertheless, the only dissimilarity is that the spectrogram employs a continuous bandwidth of channel. Conversely, the cochleagram employs a diverse bandwidth, specifically cochlea.

3.2. Feature extraction using vision transformer

The extraction of the feature process is executed by the ViT method. A transformer learns meaning and context through sequential information [22] and utilizes the ideas of self-attention or attention for identifying the relations among components. Before the development of transformers, consumers are essential for training neural networks utilizing huge, labeled databases. Transformers extract this requirement by recognizing patterns among components. Transformers have accomplished significant success in the tasks of natural language processing (NLP). The ViT method employs attention to smaller patches of images instead of individual pixels. The ViT method separates an image into 16x16 pixel patches as an input to the transformer encoding. The encoder includes Multi-Head Self-Attention (MSA) and Multi-Layer Perceptron (MLP). The model is trained and adjusted to classify images.

ViTs have dual benefits over CNN. Primarily, they have input-adaptive weight. Its attention weight is dynamic and can be altered based on input. The 2nd benefit is a global receptive area where ViT can monitor the overall images. Nevertheless, transformers hold few restrictions. Primarily, image attention networks frequently struggle with translational invariance, while objects are repositioned or shifted. Additionally, the ViT-based method aims to global aspects and underachieve CNN. Figure 3 indicates the architecture of the ViT model.

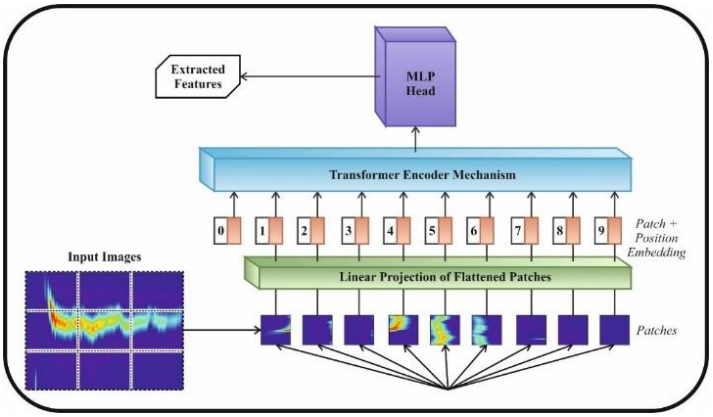

Figure 3. Architecture of the ViT method.

3.3. Multi-model diagnostic framework

At last, the proposed EFDSDRTC-HDLM model implements a hybrid of the G-BiL-MHA method for the classification. A GCN is a type of CNN appropriate for handling graph-structured data [23]. The main objective is to utilize graph convolution for removing the spatial features using a non-Euclidean framework. For $G = (V, E, A)$, an output signal Y, input signal X, and process f are implemented by the GCN network is described as

$$f(X, A) = Y. \tag{5}$$

Moreover, A, V, and E signify the connectivities among nodes V_i and V_j, the adjacency matrix, and node counts, respectively, within graph G.

$$H^{(l+1)} = \sigma\left(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}H^l W^l\right). \tag{6}$$

Here, $\tilde{A}$ denotes the matrix of adjacency after linking the closed loop; parameter $\widetilde{D}$ means diagonal matrix; H^l refers to the initial layer's output value; X stands for identity matrix, $H^0 = X$; $\sigma(\cdot)$ symbolizes the function of activation; and parameter W^l designates the parameter values of initial layers.

A BiLSTM network is a modified version of RNN structure, which includes a forward and a backward LSTM, either linked to a similar layer of input. By combining the information from either direction, it additionally widely explains and represents the sequence data. The output and time t can be determined by

$$\vec{h}_t = \overrightarrow{LSTM}\left(\vec{h}_{t-1}, X_{t'}, \vec{C}_{t-1}\right), t \in [1, T], \tag{7}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM(\overleftarrow{h_{t-1}}, \overleftarrow{C}_{t-1})}, t \in [1, T], \tag{8}$$

$$h_t = \sigma\left(W_h\left[\vec{h}_t, \overleftarrow{h_t}\right] + b_h\right). \tag{9}$$

Here, $\overleftarrow{h}$ and $\vec{h}_{t-1}$, signify the cell and hidden states of an input point at instant $t\text{-}l$, $\overleftarrow{C}_{t-1}$and $\overleftarrow{h}_{t-1}$, characterize the cell and hidden states of the input point at instant t, and b_h and W_h represent the biased term and weighted matrix, correspondingly.

An attention mechanism (AM) is applied for feature extraction and sequence models. A multi-head AM can learn rich and different data models by combining numerous parallel single-head self-attention units.

$$s_i = F(Q, k_i), \tag{10}$$

$$\alpha_i = softmax(s_i) = \frac{\exp(s_i)}{\Sigma_{j=1}^{N} exp\left(s_j\right)}, \tag{11}$$

$$Attention = \left((K, V), Q\right) = \sum_{i=1}^{N} \alpha_i \mu_i. \tag{12}$$

Initially, we compute the similarity s_i among the key and the query. We utilize the function of softmax to mathematically convert the similarity s_i into weighted coefficients α_i. Last, we apply these coefficients, and α_i is used to compute the weighted amount of the values.

Spatial extraction of the features module includes dual layers of GCNConv tailored for building an adjacency matrix, thus allocating a link-weighted matrix to all data points. This distance-based matrix of adjacency offers an insightful depiction of the spatial relation that aids the method in seizure spatial features. To build the GCN, the data features and matrix of adjacency are input across the networks, and data on nodes and their nearby nodes are handled by dual-layer GCNConv. The functions of ReLU are applied to present non-linearity and improve the generalizability and expressiveness of the method. The computation procedure of the feature extraction can be revealed by

$$\tilde{A} = A + I, \tag{13}$$

$$\tilde{A} = \widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}, \tag{14}$$

$$D = \sum_{j} \tilde{A}_{ij}, \tag{15}$$

$$H^{(i+1)} = \sigma\left(\hat{A}H^{(i)}W^{(i)}\right). \tag{16}$$

Moreover, $w \in [0,1]$ characterizes the weight co-efficients; a_{ij} epitomizes the impact of node i; $A \in R^{N\times N}$, which includes the components a_{ij}; A represents a novel matrix of adjacency that signifies the connectivity amongst each node within the graph; and parameter I characterizes the matrix of units that can be applied to include the closed loop in A.

The multi-head attention (MHA) module is employed to process the higher nonlinearity by concentrating on dissimilar parts of the sequences of the input, therefore removing attributes at different levels. For all attention heads, the matrix of input feature $x \in R^{n\times d}$ is applied. Primarily, the scores must be calculated by capturing the dot product of X and the weighted vector $\omega \in R^{d}$: scores $= \omega \times X$. Then, the softmax is used to standardize weights=softmax (scores). Last, the weighted amount has been computed to give the head's output: output= $\Sigma(weight \times X)$. Later handling by each head, the outputs can be connected to create complete feature representations.

The novelty of this study lies in their unified combination of single framework particularly designed for pediatric dysarthric speech recognition. The proposed framework enables complementary learning of global contextual representations through the temporal dependencies via BiLSTM, adaptive feature refinement using Multi-Head Attention, Vision Transformer, and graph-structured feature relationships using GCN resulting in enhanced recognition performance.

Though researchers have examined GCN-LSTM and Vision Transformer-based hybrid networks, they focus on either transformer based or graph structured feature representation independently. In contrast, the proposed framework demonstrates a unified architecture in which the Vision Transformer gets global contextual representations from time frequency speech images, the Graph Convolutional Network takes structural relationships over the extracted features, the BiLSTM models long term temporal dependencies, and the Multi-Head Attention strategy selectively emphasizes the high discriminative feature representations before the classification. This integration enables concurrent learning of temporal, structural, and global characteristics, making the proposed architecture particularly suited for pediatric dysarthria speech recognition.

4. Results and discussion

The method runs on Python 3.6.5 with an i5-8600k CPU, 4GB GPU, 16GB RAM, 250GB SSD,

and 1TB HDD. The experimental evaluation of the EFDSDRTC-HDLM algorithm is inspected by employing the Dysarthria and Non-Dysarthria Speech dataset [24] and Noise Reduced UASPEECH Dysarthria Dataset [25]. The proposed model is evaluated by a 70:30 data split; 70% for training and 30% for testing. This dataset comprises overall numbers of 8214 samples based on the two classes. Table 2 shows the parameter setting of the proposed methodology. The complete details are shown below in Table 3.

Table 2. Parameter setting.

Method	Parameter	Value
Speech Preprocessing	Sampling Frequency	16 kHz
	Frame Length	0.04 × Fs
	Frame Shift	0.05 × Fs
Windowing	Window Function	Hamming
STFT	Window Length	0.016 × Fs
	FFT Length	512
Spectrogram	Image Size	256 × N
	Frequency Split	128 × N
Mel-Scalogram	Mel Filters	40
Vision Transformer	Patch Size	16 × 16
	Embedding Dimension	768
	Attention Heads	12
	Transformer Layers	12
GCN	Graph Convolution Layers	2
	Activation	ReLU
BiLSTM	Hidden Units	256
Multi-Head Attention	Number of Heads	8
Classification	MLP Head	Softmax
Optimizer	Adam	Learning Rate = 0.001
Training	Batch Size	32
	Epochs	3000

Table 3. Details of dataset.

Classes	No. of Samples
Dysarthria	2995
Non-Dysarthria	5219
Total Samples	**8214**

Figure 4 shows the confusion matrices generated by the EFDSDRTC-HDLM model on various epochs. The experimentation outcome values denote that the EFDSDRTC-HDLM method effectively identifies and categorizes every class.

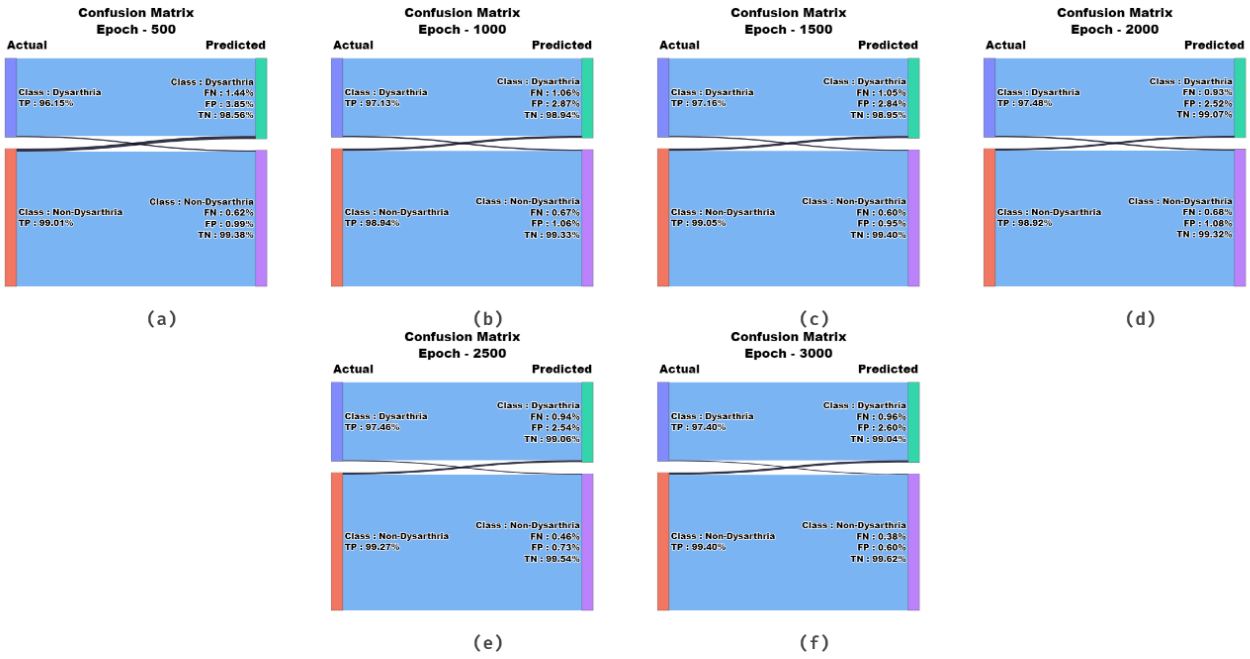

Figure 4. Confusion matrices of EFDSDRTC-HDLM method (a-f) Epochs 500 to 3000.

The classifier outcome of the EFDSDRTC-HDLM system is accomplished with different epochs in Table 4 and Figure 5. The simulation outcome values denote that the EFDSDRTC-HDLM system properly detects all instances. According to 500 epochs, the EFDSDRTC-HDLM method gives an average accuracy of 98.02%, precision of 97.58%, recall of 98.02%, f-score of 97.79%, and MCC of 95.60%. Similarly, at 1500 epochs, the EFDSDRTC-HDLM system presents an average accuracy of 98.36%, precision of 98.11%, recall of 98.36%, f-score of 98.23%, and MCC of 96.47%. As well, for 2500 epochs, the EFDSDRTC-HDLM technique offers an average accuracy of 98.63%, precision of 98.36%, recall of 98.63%, f-score of 98.49%, and MCC of 96.99%. Additionally, at 3000 epochs, the EFDSDRTC-HDLM algorithm gives an average accuracy of 98.73%, precision of 98.40%, recall of 98.73%, f-score of 98.56%, and MCC of 97.13%.

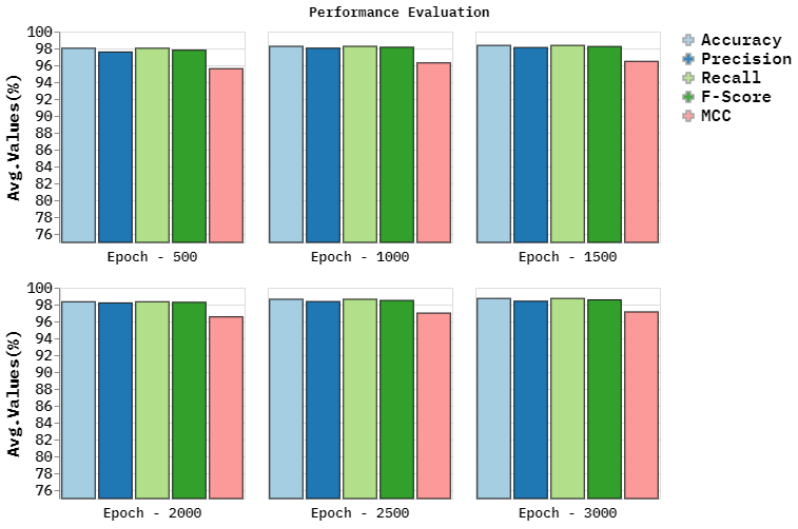

Figure 5. Average values of EFDSDRTC-HDLM system with various epochs.

Table 4. Classifier results of EFDSDRTC-HDLM model under different epochs.

Class Labels	$Accur_y$	$Preci_n$	$Recal_l$	F_{Score}	MCC
Epoch - 500					
Dysarthria	98.30	96.15	98.30	97.21	95.60
Non-Dysarthria	97.74	99.01	97.74	98.37	95.60
Average	**98.02**	**97.58**	**98.02**	**97.79**	**95.60**
Epoch - 1000					
Dysarthria	98.16	97.13	98.16	97.64	96.28
Non-Dysarthria	98.33	98.94	98.33	98.64	96.28
Average	**98.25**	**98.03**	**98.25**	**98.14**	**96.28**
Epoch - 1500					
Dysarthria	98.36	97.16	98.36	97.76	96.47
Non-Dysarthria	98.35	99.05	98.35	98.70	96.47
Average	**98.36**	**98.11**	**98.36**	**98.23**	**96.47**
Epoch - 2000					
Dysarthria	98.13	97.48	98.13	97.80	96.54
Non-Dysarthria	98.54	98.92	98.54	98.73	96.54
Average	**98.34**	**98.20**	**98.34**	**98.27**	**96.54**
Epoch - 2500					
Dysarthria	98.73	97.46	98.73	98.09	96.99
Non-Dysarthria	98.52	99.27	98.52	98.89	96.99
Average	**98.63**	**98.36**	**98.63**	**98.49**	**96.99**
Epoch - 3000					
Dysarthria	98.96	97.40	98.96	98.18	97.13
Non-Dysarthria	98.49	99.40	98.49	98.94	97.13
Average	**98.73**	**98.40**	**98.73**	**98.56**	**97.13**

Figure 6 represents the training accuracy and validation accuracy of an EFDSDRTC-HDLM method by employing Epoch 3000. The training accuracy is incessantly improved and signifies proficient learning and appropriate convergence of the technique. The validation accuracy is followed closely the training curve with the minor variances, offering reduced overfitting and higher generalized outcomes.

Figure 7 demonstrates the training and validation losses of the EFDSDRTC-HDLM approach at Epoch 3000. These losses must be minimized strongly within the primary epochs, designating faster learning, and after progressively decreasing in smooth and continual means, are considered efficient convergence. The closer arrangements among training and validation loss, with only a smaller reliable gap, indicates lower overfitting and stronger generalization capabilities.

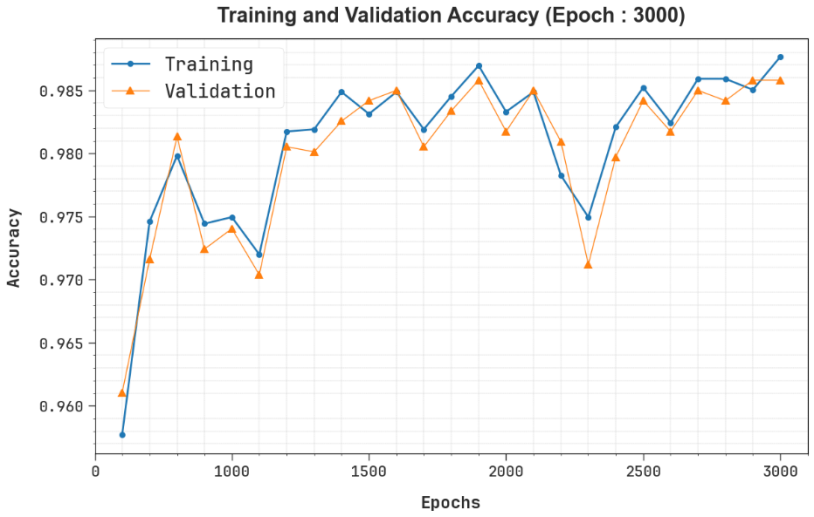

Figure 6. $Accur_y$ curve of the EFDSDRTC-HDLM algorithm at Epoch 3000.

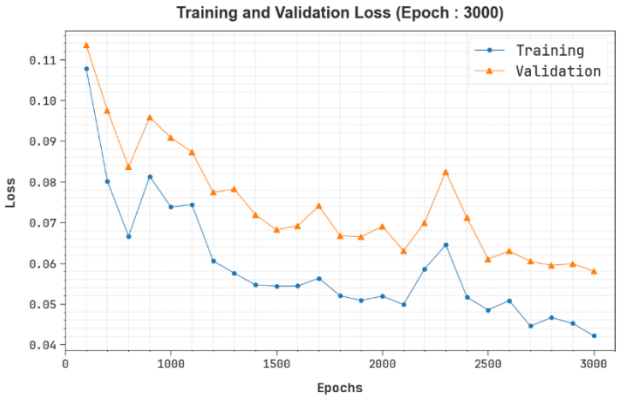

Figure 7. Loss curve of the EFDSDRTC-HDLM technique at Epoch 3000.

In Figure 8, the precision-recall (PR) curve analysis of the EFDSDRTC-HDLM system with Epoch 3000 delivers valuable insights into its effectiveness by plotting Precision against Recall for each class. This figure depicts that the EFDSDRTC-HDLM algorithm determinedly realizes elevated values of PR in numerous classes, signifying that it is possible to uphold a major part of true positive predictions among all the positive predictions. The continual development shows the ability of the EFDSDRTC-HDLM methodology.

In Figure 9, the ROC curve of EFDSDRTC-HDLM system using Epochs 3000 is inspected. The experimentational outcomes denote that the EFDSDRTC-HDLM system realizes increased ROC values over all classes, showing the important abilities to differentiate the classes. This continuous pattern of increased values of ROC over diverse classes indicates the efficient performance of the EFDSDRTC-HDLM system with the class prediction, underscoring the classification process.

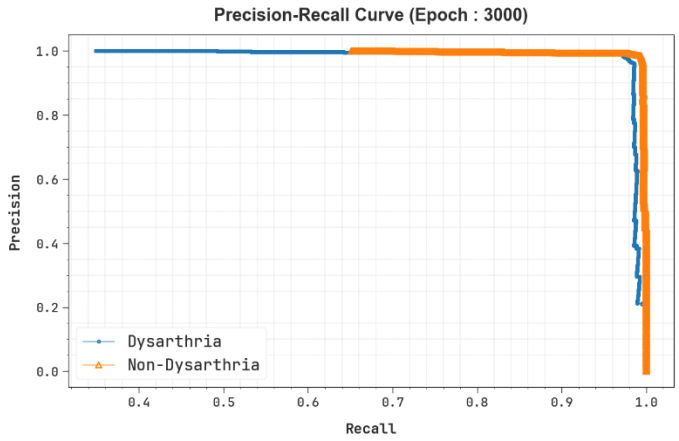

Figure 8. PR curve of the EFDSDRTC-HDLM method under Epoch 3000.

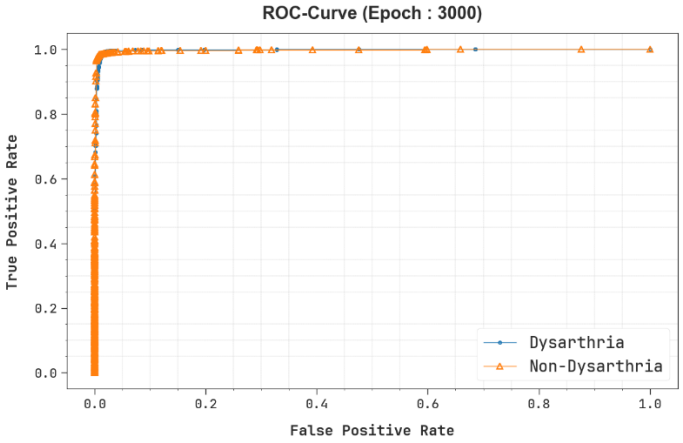

Figure 9. ROC curve of the EFDSDRTC-HDLM method under Epoch 3000.

Table 5 and Figure 10 portray the comparison evaluation of the EFDSDRTC-HDLM model with current approaches with respect to diverse measures under the Dysarthria and Non-Dysarthria Speech Dataset [26–28]. The experimental findings highlight that the proposed EFDSDRTC-HDLM system achieves a higher accuracy, precision, recall, and f-score of 98.73%, 98.4%, 98.73%, and 98.56%, respectively. These methodologies, such as DNN-HMM, CNN, wac2vec2-LARGE-4, Hubert-12, and EfficientNet-LEVIT, obtain somewhat closer performance, while the MFCC+SFIT, EfficientNet B7, MobileNet V2, and VGGish-SVM models have the worst performance. The performance metrics steadily improve with increasing iterations of training due to the progressive optimization of model parameters. Minor fluctuations during intermediate iterations are featured to the random nature of the parameter updates and optimization process. As this increases, the model approaches to a stable solution, which results in consistent classification performance and the robustness of the proposed architecture. To provide an extensive performance evaluation, the comparative study has been extended to include speech recognition models besides conventional CNN-based approaches. This broader comparison enables a fair assessment of the proposed architecture against contemporary state-of-the-art methods and further validates its efficacy for pediatric dysarthria speech recognition.

Table 5. Comparative study of the EFDSDRTC-HDLM algorithm with recent methods under Dysarthria and Non-Dysarthria Speech Dataset.

Methodology	Accuracy	Precision	Recall	F1-Score
DNN-HMM	96.34	92.61	94.40	93.48
MFCC+SFIT	94.72	90.54	96.92	98.10
CNN	96.05	94.84	91.82	91.03
EfficientNet B7	91.56	93.76	97.51	90.36
MobileNet V2	92.39	95.16	97.18	89.57
EfficientNet-LEVIT	97.93	90.42	97.42	91.71
VGGish-SVM	90.21	98.70	90.26	97.80
wac2vec2-LARGE-4	95.17	94.82	92.00	95.00
Hubert-12	95.50	95.36	93.00	96.00
EFDSDRTC-HDLM	98.73	98.40	98.73	98.56

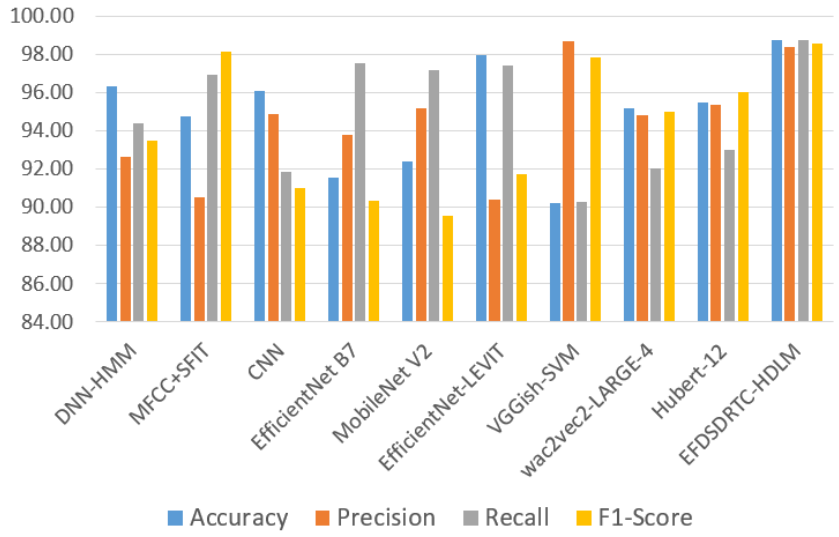

Figure 10. Comparative analysis of the EFDSDRTC-HDLM algorithm with existing approaches.

Table 6 and Figure 11 show the computational time (CT) of the EFDSDRTC-HDLM method with existing methods. Based on CT, the EFDSDRTC-HDLM methodology gives a lower value of 5.00 sec, while the DNN-HMM, MFCC+SFIT, CNN, EfficientNet B7, MobileNet V2, EfficientNet-LEVIT, and VGGish-SVM methodologies obtain a superior CT of 9.48 sec, 8.51 sec, 8.90 sec, 8.43 sec, 10.64 sec, 11.13 sec, and 13.91 sec, correspondingly.

Table 6. CT outcome of the EFDSDRTC-HDLM method with recent methods.

Methodology	Computational Time (sec)
DNN-HMM	9.48
MFCC+SFIT	8.51
CNN	8.90
EfficientNet B7	8.43
MobileNet V2	10.64
EfficientNet-LEVIT	11.13
VGGish-SVM	13.91
EFDSDRTC-HDLM	5.00

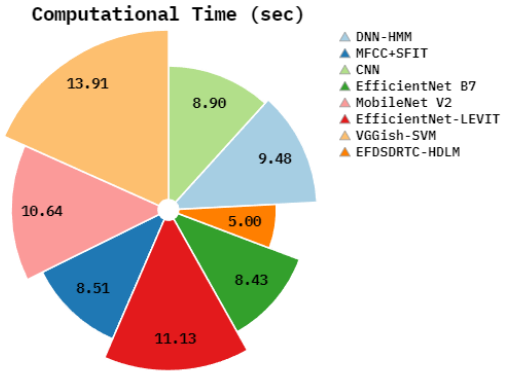

Figure 11. CT outcome of the EFDSDRTC-HDLM system with recent systems.

Table 7 depicts the computational efficiency of the presented EFDSDRTC-HDLM model with existing methods. Based on inference time, the EFDSDRTC-HDLM method provides lesser values of 3.68 ms, Flops of 0.052G and GPU of 14.8 GB, whereas the DNN-HMM, MFCC+SFIT, CNN, EfficientNet B7, MobileNet V2, EfficientNet-LEVIT, VGGish-SVM, wav2vec2-LARGE-4, and Hubert-12 techniques attain greater times of 18.74 ms, 21.58 ms, 12.96 ms, 34.71 ms, 8.43 ms, 16.25 ms, 24.18 ms, 30.82 ms, and 28.67 ms, correspondingly.

Table 7. Computational efficiency of the EFDSDRTC-HDLM method.

Methodology	Inference Time (ms)	Params (M)	FLOPs (G)	GPU (GB)
DNN-HMM	18.74	14.62	5.84	68.5
MFCC+SFIT	21.58	28.94	8.46	73.9
CNN	12.96	9.82	3.15	56.8
EfficientNet B7	34.71	66.35	37.02	91.4
MobileNet V2	8.43	3.54	0.32	39.7
EfficientNet-LEVIT	16.25	19.83	4.91	63.8
VGGish-SVM	24.18	62.47	15.74	79.6
wav2vec2-LARGE-4	30.82	317	95.64	96.1
Hubert-12	28.67	316.2	92.18	94.8
EFDSDRTC-HDLM	**3.68**	**0.73**	**0.052**	**14.8**

Table 8 displays the comparative analysis of the EFDSDRTC-HDLM model with existing methods under a Noise Reduced UASPEECH Dysarthria Dataset [25]. The proposed EFDSDRTC-HDLM model demonstrates the promising performance, accuracy, precision, recall, and f1-score of 97.63%, 97.12%, 97.04%, and 97.62%, respectively, compared to the other models, such as wav2vec2-BASE-1, Hubert-8, eGeMAPS, OpenSMILE, and MFCCs, showing the worst performances.

Table 8. Comparative analysis of the proposed EFDSDRTC-HDLM approach under the Noise Reduced UASPEECH Dysarthria Dataset.

Methodology	Accuracy	Precision	Recall	F1-Score
wav2vec2-BASE-1	93.96	94.18	92	94
Hubert-8	95.25	95.41	93	95
eGeMAPS	87.33	88.14	87	88
OpenSMILE	92.52	92.18	90	92
MFCCs	86.58	86.21	84	87
EFDSDRTC-HDLM	**97.63**	**97.12**	**97.04**	**97.62**

Table 9 represents the ablation study of the proposed EFDSDRTC-HDLM model. The Vision Transformer (ViT) achieves an accuracy of 91.84%, precision of 91.32%, recall of 91.71%, and F1-Score of 91.51%, and the ViT + GCN provides an accuracy of 93.42%, precision of 93.08%, recall of 93.27%, and F1-score of 93.17%. Moreover, the ViT + BiLSTM attains a better accuracy of 94.37%, precision of 94.02%, recall of 94.21%, and F1-score of 94.11%. Furthermore, the ViT + GCN + BiLSTM increases accuracy to 96.05%, precision to 95.71%, recall to 95.96%, and F1-score to 95.83%. Besides, the ViT + GCN + Multi-Head Attention model attains a moderate performance. Last, the EFDSDRTC-HDLM (ViT + GCN + BiLSTM + Multi-Head Attention) model achieves the highest outputs with an accuracy of 98.73%, precision of 98.4%, recall of 98.73%, and F1-score of 98.56%, emphasizing the efficiency of the components.

Table 9. Ablation study of the EFDSDRTC-HDLM model.

Methodology	Accuracy	Precision	Recall	F1-Score
Vision Transformer (ViT)	91.84	91.32	91.71	91.51
ViT + GCN	93.42	93.08	93.27	93.17
ViT + BiLSTM	94.37	94.02	94.21	94.11
ViT + GCN + BiLSTM	96.05	95.71	95.96	95.83
ViT + GCN + Multi-Head Attention	97.14	96.86	97.08	96.97
EFDSDRTC-HDLM (ViT + GCN + BiLSTM + Multi-Head Attention)	98.73	98.4	98.73	98.56

Table 10 represents the 10-fold cross-validation performance of the proposed model using Accuracy, Precision, Recall, and F1-Score. In k-1, the model achieves an accuracy of 96.84%, precision of 96.05%, recall of 96.31%, and F1-score of 96.18%. However, in k-2, the performance improves with an accuracy of 97.42%, precision of 96.88%, recall of 97.05%, and F1-score of 96.96%. Moreover, in k-3, the model records an accuracy of 96.93%, precision of 96.41%, recall of 96.52%, and F1-score of 96.46%. Likewise, in k-9, the model obtains an accuracy of 97.05%, precision of

96.55%, recall of 96.71%, and F1-score of 96.63%, maintaining balanced classification results. Finally, in k-10, the model achieves an accuracy of 97.71%, precision of 97.08%, recall of 97.19%, and F1-score of 97.13%. Overall, the mean performance reaches 97.32% ± 0.46% for accuracy, 96.73% ± 0.49% for precision, 96.89% ± 0.45% for recall, and 96.81% ± 0.48% for F1-score, demonstrating reliable and robust classification performances across all validation folds.

Table 10. K-fold cross validation of the EFDSDRTC-HDLM technique.

K-Fold	Accuracy	Precision	Recall	F1-Score
k-1	96.84	96.05	96.31	96.18
k-2	97.42	96.88	97.05	96.96
k-3	96.93	96.41	96.52	96.46
k-4	98.01	97.45	97.62	97.56
k-5	97.18	96.62	96.84	96.73
k-6	97.56	96.97	97.11	97.04
k-7	96.65	95.98	96.22	96.10
k-8	97.83	97.26	97.33	97.29
k-9	97.05	96.55	96.71	96.63
k-10	97.71	97.08	97.19	97.13
Mean ± SD	**97.32 ± 0.46**	**96.73 ± 0.49**	**96.89 ± 0.45**	**96.81 ± 0.48**

The high classification performance attained by the proposed framework shows its potential applicability in the clinical practice for pediatric dysarthria screening. Accurate identifications of dysarthric speech can help speech language pathologists in the early detection of speech disorders, which enables continuous monitoring of children's speech development and timely therapeutic intervention. The improved Accuracy, Precision, Recall, and F1-scoredemonstrates the framework's capability to reliably identify dysarthric speech from the normal speech, thereby reducing the misclassification and supporting more dependable clinical decision making. To assess the generalization capability of the proposed architecture, experiments are performed for two publicly available dysarthria speech datasets. The dataset division strategy is carefully designed for maintaining independent training and testing subsets, minimizing the possibility of speaker overlap among the two sets. In addition, K-Fold cross-validation is performed to obtain a reliable and comprehensive assessment of the proposed method across data partitions.

5. Conclusions

In this study, we develop an EFDSDRTC-HDLM method. The aim of the EFDSDRTC-HDLM methodology is to develop an effective recognition technique for identifying dysarthria speech disorders in children using advanced speech processing and classification methods. In the pre-processing stage, the audio files can be transformed into time-frequency images by detecting active speech regions, applying framing and windowing, and using a time-frequency image generator to provide visual inputs for further analysis. This is followed by the extraction of a feature process, which is executed by the ViT model. Finally, the proposed EFDSDRTC-HDLM technique implements a hybrid of the G-BiL-MHA model for the classification. The proficiency of the EFDSDRTC-HDLM approach is confirmed by the wide-ranging analysis utilizing the standard dataset. The mathematical

outcome signifies that the EFDSDRTC-HDLM system has improved scalability and performance in terms of several evaluations over the methodologies. Despite the dataset consisiting of pediatric dysarthric speech recordings, detailed age-group annotations are not available in the publicly released dataset. Consequently, age-stratified performance analysis could not be performed. In future work, we will evaluate the proposed framework using datasets containing comprehensive demographic information to investigate model performance across pediatric age groups and on speaker-independent data partitioning to further enhance the model's generalization capability.

Author contributions

Fahd N. Al-Wesabi: Conceptualization; Methodology; Funding Acquisition; Writing – Original Draft, Writing – Review & Editing, Data Curation; Validation, Formal Analysis. Wafi Bedewi: Software; Visualization, Formal Analysis, Original Draft; Investigation, Validation, Original Draft, Writing – Review & Editing.

Use of Generative-AI tools declaration

The authors declare that no generative AI or AI-assisted technologies were used in the preparation of this manuscript. All aspects of the research, including the study design, data analysis, interpretation of results, and manuscript writing, were carried out solely by the authors.

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 [24] and https://www.kaggle.com/datasets/aryashah2k/noise-reduced-uaspeech-dysarthria-dataset [25].

Acknowledgments

The authors extend their appreciation to the King Salman center for Disability Research for funding this work through Research Group Number KSRG-2026-269. The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Project under grant number RGP2/199/46.

Conflict of interest

All authors declare no conflicts of interest in this paper.

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. J. Iuzzini-Seigel, K. M. Allison, R. Stoeckel, A tool for differential diagnosis of childhood apraxia of speech and dysarthria in children: A tutorial, *Lang. Speech Hear. Ser.*, **53**(2022), 926–946. https://doi.org/10.1044/2022_LSHSS-21-00164
3. G. Tartarisco, R. Bruschetta, S. Summa, L. Ruta, M. Favetta, M. Busà, Artificial intelligence for dysarthria assessment in children with ataxia: A hierarchical approach, *IEEE Access*, **9** (2021), 166720–166735. https://doi.org/10.1109/ACCESS.2021.3135078
4. S. H. Lee, M. Kim, H. G. Seo, B. M. Oh, G. Lee, J. H. Leigh, Assessment of dysarthria using one-word speech recognition with hidden Markov models, *J. Korean Med. Sci.*, **34** (2019), e108. https://doi.org/10.3346/jkms.2019.34.e108
5. 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
6. A. A. Anthony, C. M. Patil, J. Basavaiah, A review on speech disorders and processing of disordered speech, *Wireless Pers Commun*, **126** (2022), 1621–1631. https://doi.org/10.1007/s11277-022-09812-w
7. H. P. Rowe, S. E. Gutz, M. F. Maffei, K. Tomanek, J. R. Green, Characterizing dysarthria diversity for automatic speech recognition: A tutorial from the clinical perspective, *Front. Comput. Sci.*, **4** (2022), 770210. https://doi.org/10.3389/fcomp.2022.770210
8. M. Kooi-van Es, C. E. Erasmus, B. J. de Swart, N. B. Voet, P. J. van der Wees, I. J. de Groot, et al., Dysphagia and dysarthria in children with neuromuscular diseases, a prevalence study, *J. Neuromuscul. Dis.*, **7** (2020), 287–295. https://doi.org/10.3233/JND-190436
9. P. McCabe, J. Korkalainen, D. Thomas, Diagnostic uncertainty in childhood motor speech disorders: A review of recent tools and approaches, *Curr. Dev. Disord. Rep.*, **11** (2024), 105–112. https://doi.org/10.1007/s40474-024-00295-x
10. H. L. Fouad, H. A. Abdulmohsin, A hybrid speech recognition system using deep learning methods, *J. Intell. Syst. Internet Things*, **15** (2025), 105–121. https://doi.org/10.54216/JISIoT.150109
11. 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
12. D. Zhang, H. Zhang, W. Lu, W. Li, J. Wang, J. Wei, Long-range and non-stationary encoding for dysarthric speech data augmentation, *IEEE J. STSP*, **19** (2025), 767–782. https://doi.org/10.1109/JSTSP.2025.3562417
13. A. T. Celin Mariya, P. Vijayalakshmi, T. Nagarajan. K. Mrinalini, Augmentative and alternative speech communication (AASC) aid for people with dysarthria, *Comput. Speech Lang.*, **92** (2025), 101777. https://doi.org/10.1016/j.csl.2025.101777
14. M. Usha, An IoT and data mining-based tool for early identification of speech disorders in children using advanced algorithms. In: *Artificial intelligence based smart and secured applications*, ASCIS 2024, 2024, 375–384. https://doi.org/10.1007/978-3-031-86290-8_27
15. A. K. Dutta, A. R. W. Sait, A speech disorder detection model using ensemble learning approach, *J. Disabil. Res.*, **3** (2024), 1–8. https://doi.org/10.57197/JDR-2024-0026

16. K. Mittal, K. S. Gill, K. Rajput, V. Singh, Enhancing the diagnosis of speech disorders: An in-depth investigation into dysarthria classification Using the ResNet18 Model. In: *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*, 2024. https://doi.org/10.1109/ICITEICS61368.2024.10625627
17. Y. Lin, L. Wang, Y. Yang, J. Dang, CFDRN: A cognition-inspired feature decomposition and recombination network for dysarthric speech recognition, *IEEE/ACM T. Audio SPE*, **31** (2023), 3824–3836. https://doi.org/10.1109/TASLP.2023.3319276
18. B. Jolad, R. Khanai, Competitive crow search algorithm-based hierarchical attention network for dysarthric speech recognition, *Int. J. Wirel. Mobi. Comput.*, **25** (2023), 340–352. https://doi.org/10.1504/ijwmc.2023.135384
19. J. Wu, Y. G. Wang, G. J. McLachlan, Informative missingness and its implications in semi-supervised learning, *Innovation Inform.*, **2** (2026), 100033. https://doi.org/10.59717/j.xinn-inform.2026.100033
20. N. Xu, J. Wu, F. Cai, X. A. Li, H. B. Xie, ViT-GCN: A novel hybrid model for accurate pneumonia diagnosis from x-ray images, *Biomed. Phys. Eng. Expr.*, **11** (2025), 045034. https://doi.org/10.1088/2057-1976/adebf4
21. S. Aurobindo, R. Prakash, M. Rajeshkumar, Comparative analysis of different time-frequency image representations for the detection and severity classification of dysarthric speech using deep learning, *Results Eng.*, **25** (2025), 104561. https://doi.org/10.1016/j.rineng.2025.104561
22. E. Alattas, J. Clark, A. Al-Aama, S. K. Jarraya, Evaluating features and variations in deepfake videos using the CoAtNet model, *J Imaging*, **11** (2025), 194. https://doi.org/10.3390/jimaging11060194
23. W. Zhou, W. Wang, X. Wang, Research on lightning prediction based on GCN-LSTM model integrating spatiotemporal features, *Atmosphere*, **16** (2025), 447. https://doi.org/10.3390/atmos16040447
24. Dysarthria and Non-Dysarthria Speech Dataset. Available from: https://www.kaggle.com/datasets/poojag718/dysarthria-and-nondysarthria-speech-dataset.
25. Noise Reduced UASPEECH Dysarthria Dataset. Available from: https://www.kaggle.com/datasets/aryashah2k/noise-reduced-uaspeech-dysarthria-dataset
26. S. V. Vishnika, S. Chandrakala, Investigation of DNN-HMM and lattice-free maximum mutual information approaches for impaired speech recognition, *IEEE Access*, **9** (2021), 168840–168849. https://doi.org/10.1109/ACCESS.2021.3129847
27. 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
28. 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