ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Adaptive attention and pyramid pooling for high-performance dysarthric speech recognition

Zahraa Tarek[1], Tarek Abd El-Hafeez[2], Esraa Hassan[3]*, Ahmed Fahim[4], Khaled Alnowaiser[4], and Mahmoud Y. Shams[3]

[1]Department of Computer Engineering and Information, College of Engineering, Wadi Ad Dwaser, Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia, Al-Kharj, Saudi Arabia

[2]Department of Computer Science, Faculty of Science, Minia University, Minia, Egypt

[3]Department of Machine Learning and Information Retrieval, Faculty of Artificial Intelligence, Kafrelsheikh University, Kafrelsheikh, Egypt

[4]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia

Abstract

Speech pathology is a multidisciplinary field that assesses, diagnoses, and manages communication disorders caused by neurological, developmental, structural, and degenerative conditions. Among these disorders, dysarthria is a motor speech impairment characterized by reduced speech intelligibility and abnormal articulatory patterns resulting from neuromuscular dysfunction. Automatic dysarthria assessment remains challenging because of substantial inter-speaker variability, diverse severity levels, and the complex temporal–spectral characteristics of pathological speech signals. To address these challenges, this paper proposes an attention-enhanced convolutional neural network for automatic dysarthria detection and severity classification. The proposed architecture integrates a convolutional block attention module to enhance channel-wise and spatial feature representations, a coordinate attention module to capture long-range contextual dependencies while preserving positional information, and spatial pyramid pooling to extract robust multi-scale representations and accommodate input variability. Hierarchical convolutional layers with rectified linear unit activation and max-pooling operations progressively learn discriminative speech representations, followed by fully connected layers and a softmax classifier for prediction. The proposed framework was comprehensively evaluated on two publicly available datasets: TORGO Dataset 1, containing dysarthric and healthy control speech recordings, and TORGO Severity Dataset 2, comprising four dysarthria severity classes. Mel-frequency cepstral coefficients, together with their first- and second-order temporal derivatives and complementary acoustic descriptors, were extracted to characterize the speech signals. Extensive experiments, including cross-validation, repeated runs, ablation studies, Shapley additive explanations-based interpretability analysis, and statistical significance analysis, demonstrate the effectiveness and robustness of the proposed framework. On TORGO Dataset 1, the proposed model achieved 95.00% accuracy, 94.50% precision, 95.25% recall, and a 94.75% F1-score, outperforming conventional machine learning and deep learning baselines. On TORGO Severity Dataset 2, the proposed framework achieved 99.10% classification accuracy, an F1-score of 0.9910, and an area under the curve of 0.9999, demonstrating strong performance in multi-class dysarthria severity classification. These findings highlight the proposed

***Corresponding author:**
Esraa Hassan
(esraa.hassan@ai.kfs.edu.eg)

Citation: Tarek Z, Abd El-Hafeez T, Hassan E, Fahim A, Alnowaiser K, Shams MY. Adaptive attention and pyramid pooling for high-performance dysarthric speech recognition. *Artif Intell Health*. doi: 10.36922/AIH026280086

Received: July 10, 2026

Revised: July 27, 2026

Accepted: August 6, 2026

Published online: September 4, 2026

model's strong discriminative capability, robustness, and interpretability, indicating its potential as a reliable computer-aided decision-support system for dysarthria detection and severity assessment.

Keywords: Dysarthria detection; TORGO database; Convolutional neural networks; Attention mechanisms; Speech pathology; Machine learning

1. Introduction

Speech is the main mode of verbal communication and involves coordinated neural control, language, and motor execution. Dysarthria is a group of motor speech disorders caused by neural disorders that affect the speech production mechanism, resulting in inaccurate articulation and reduced speech clarity. It is frequently associated with neurological diseases, such as cerebral palsy (CP) and amyotrophic lateral sclerosis (ALS), in which both the severity and acoustic features of speech impairment have significant inter-speaker variability. Thus, reliably identifying and assessing dysarthria remains a significant clinical challenge. This study uses the TORGO speech corpus and dysarthria severity dataset, a well-established dataset of almost 2,000 speech samples from patients with CP or ALS, and a selection of neurologically healthy speakers. The corpus provides a diverse set of pathological and non-pathological speech examples which can serve as a useful reference for examining dysarthric speech patterns and as a basis for creating data-driven diagnostic tools. In recent years, speech signal analysis has improved significantly with advances in machine learning and deep learning techniques. Traditional learning algorithms such as ensemble methods, including random forests, have proven competitive in classification tasks because they are robust and easy to interpret. However, architectures that handle complex temporal dynamics and spectral variations in speech signals more effectively, such as convolutional neural networks (CNNs) and sequence-based models like long short-term memory (LSTM) networks, have shown greater success.[1-17] Despite these advances, modeling the significant acoustic variability and long-range dependency that is typical of dysarthric speech is a problem yet to be solved. To address these drawbacks, we propose an attention-based CNN architecture for dysarthria detection. The proposed framework is based on the convolutional block attention module (CBAM), which selectively strengthens informative channel and spatial representations, and on coordinate attention (CA) to model extended contextual relationships within speech features. The attention mechanisms are applied at multiple network layers, and spatial pyramid pooling (SPP) is used to obtain a compact, discriminative global representation. The last layer is fully connected, followed by a softmax classifier. The proposed method is extensively tested on the TORGO dataset and benchmarked against conventional machine learning and deep learning baselines. Clinically meaningful measures, such as accuracy, precision, recall, and F1-score, were used to evaluate the performance. The results show that the attention-based CNN model outperforms the other models. Furthermore, applying Shapley additive explanations (SHAP) improves model transparency by revealing feature relevance and increasing confidence in automated diagnostic decisions. This research provides a comprehensive, clinically viable framework for early, accurate, and reliable dysarthria detection, combining state-of-the-art attention mechanisms and interpretability principles, with potential applications in real-world speech pathology and healthcare systems.

1.1. Problem statement

We propose a CNN architecture with attention mechanisms: CBAM and CA to improve the ability to discriminate features and to capture contextual relationships between features in speech representation. These attention components allow the model to highlight salient temporal and spectral features that are important for distinguishing dysarthric from non-dysarthric speech.

1.2. Research question

What are the improvements that can be achieved with an attention-enhanced CNN-based model for detecting dysarthria in the TORGO database for CP and ALS patients?

1.3. Contributions

This study makes the following contributions:

(i) We introduced a novel CNN model combined with a CBAM and a CA module, which enhances feature extraction and captures long-range dependencies in dysarthric speech.
(ii) Our model performs better than traditional machine learning or deep learning methods in detecting dysarthria in the TORGO database, and the TORGO Severity Dataset can outperform both of these methods in dysarthria detection.

(iii) SPP is implemented, leading to global feature aggregation that boosts the model's ability to generalize over various dysarthric speech patterns.
(iv) For clinical use, we employ SHAP analysis to give insights into the model's decision-making process, which is transparent and trustworthy.

Our proposed model balances performance and interpretability and can be a viable approach for real-world dysarthria detection in speech pathology and assistive speech technologies.

2. Related work

Despite high variability in dysarthric speech, speaker-specific features, and the scarcity of large, diverse datasets, automatic identification and severity estimation of dysarthric speech remain challenging. While new-generation machine learning and deep learning techniques for assessing dysarthria are gaining popularity, challenges remain: limited discriminative feature learning, low speaker independence, and low robustness on heterogeneous datasets.[18] Narendra and Alku[19] conducted early research on dysarthria recognition using hybrid deep learning-based radiomics with CNNs, multilayer perceptrons, and LSTM. They used both speech signals and glottal-flow waveforms, and their results showed that glottal-flow representations performed better, with an accuracy of 81.12% on the TORGO dataset. Joshy and Rajan[20] also investigated severity classification using deep neural networks, CNNs, and gated recurrent units with mel-frequency cepstral coefficients (MFCCs) and their derivatives as input features. They achieved good recognition accuracy under speaker-dependent conditions (93.97%) but a large drop in speaker-independent evaluation (49.22%), highlighting the difficulty of building generalized models. Recent research has increasingly used self-supervised learning techniques or transformer-based representations. Javanmardi *et al.*[21] compared the performance of wav2vec2-BASE, wav2vec2-LARGE, and HuBERT models for detecting and assessing dysarthria severity on the UA-Speech and TORGO datasets. The results showed that HuBERT features consistently outperformed the traditional baselines and improved accuracy by up to 10.46% in classifying event severity; however, the maximum accuracy reported on TORGO was still 76.12%. Similarly, Stumpf *et al.*[22] introduced the Speaker-Agnostic Latent Regularisation approach, which combined multi-task learning with contrastive learning in a transformer network. They achieved classification accuracy of up to 70.48% and an F1-score of 59.23% on UA-Speech. Competitive performance has also been achieved with alternative methods using signal processing. Shabber and Sumesh[23] used Fourier–Bessel series expansion to obtain amplitude envelope and instantaneous frequency information and found accuracy of 85%–97.8% on their TORGO, UA-Speech, and Parkinson's speech datasets. Radha and Bansal[24] proposed a SincNet-based framework to detect and estimate dysarthria severity, achieving 95.7% and 99.6% accuracy on TORGO and UA-Speech, respectively. Later, Radha *et al.*[25] improved classification accuracy using CNN architectures with multiple short-time Fourier transforms and pre-emphasis filtering, reaching a peak accuracy of 94.08% on TORGO. There are also positive results in the application of transfer learning techniques. Shanmugapriya and Mohan[26] converted speech into scalogram features and evaluated various pre-trained CNN architectures, achieving 99.22% accuracy on TORGO. Similarly, Sekhar *et al.*[27] used Mel-spectrogram representations with transfer-learning-based CNN models and achieved 97.73% classification accuracy. A few studies have examined hybrid approaches that integrate acoustic and linguistic information. Alharbi *et al.*[28] introduced a CNN–bidirectional LSTM network with ASR-based textual features, which gave 90% accuracy on TORGO. Likewise, S *et al.*[29] proposed a frequency-based bifurcation method for binary dysarthria classification and multi-class severity prediction using TORGO and UA-Speech, achieving over 99% accuracy on both datasets and strong cross-dataset generalization. In recent years, multi-stage learning strategies in severity estimation have been emphasized. Al-Ali *et al.*[30] proposed a two-stage framework that used a binary classifier; they extracted sentence-level representations, clustered them with *k*-means, and achieved the best accuracy of 91%. Shabber *et al.*[31] also used more complex preprocessing methods and wavelet-based scalogram representations, coupled with deep learning, to further improve classification effectiveness, with accuracies ranging from 90% to 99% and strong F1-scores across all datasets. Meanwhile, Wang *et al.*[32] adapted automatic speech recognition for dysarthric speech using a Conformer-based framework, yielding a significant reduction in word error rate on both UA-Speech and TORGO. Table 1 summarizes and compares recent dysarthria classification methods evaluated on the TORGO dataset. Despite significant advances, many approaches still rely on traditional feature extraction techniques or pre-trained architectures that fail to model long-range contextual relations in dysarthric speech signals and the complex interactions between channels and space. To address these limitations, the proposed framework uses an attention-enhanced CNN architecture to enrich feature representations with complementary attention mechanisms. Two modules are added to enhance spatial- and channel-wise learning: CA and CBAM. Furthermore, SPP is incorporated for multi-scale feature aggregation to alleviate speaker variation and the diversity effect

Table 1. Summary of recent approaches related to the work

Author	Methodology	Dataset used	Results	Remarks (Limitations & potential solutions)
Narendra & Alku[19]	CNN-MLP, CNN-LSTM using glottal flow & raw speech	TORGO	81.12%	Glottal flow improved performance, but dataset size limited generalizability. Larger datasets could enhance model robustness
Joshy & Rajan[20]	DNN, CNN, GRU with MFCCs	UA-Speech, TORGO	93.97% (Speaker-dependent), 49.22% (Speaker-independent)	Speaker-dependent performance is high, but generalization remains weak. More diverse training data is needed
Javanmardi *et al.*[21]	Pre-trained wav2vec2 & HuBERT	UA-Speech, TORGO	76.12%	HuBERT outperforms baselines, but improvements in severity classification are needed
Stumpf *et al.*[22]	Transformer-based SALR framework	UA-Speech	70.48% accuracy, 59.23% F1-score	Addressed intra-class variability but requires further optimization for real-world deployment
Shabber & Sumesh[23]	Fourier–Bessel series expansion	TORGO, UA-Speech, Parkinson	85%–97.8%	Effective on multiple datasets, but computational complexity is high
Radha & Bansal[24]	SincNet model	TORGO, UA-Speech	95.7%, 99.6%	High accuracy but limited interpretability
Shanmugapriya & Mohan[26]	Pre-trained CNNs on scalogram images	TORGO	99.22%	Requires validation on larger datasets
Sekhar *et al.*[27]	TL–CNN with Mel-spectrograms	TORGO	97.73%	Limited dataset coverage may hinder generalization
Alharbi *et al.*[28]	Hybrid CNN–BLSTM with ASR textual features	TORGO	90%	Improved transcription consistency, but still sensitive to ASR errors
S *et al.*[29]	DCNN with frequency-based bifurcation	TORGO, UA-Speech	99.53%, 97.85%	Robust results but needs evaluation on spontaneous speech
Wang *et al.*[32]	Conformer-based ASR adaptation	TORGO, UA-Speech	21.5%, 12.7% WER	ASR accuracy improved but still challenges with severely dysarthric speech

Abbreviations: ASR: Automatic speech recognition; BLSTM: Bidirectional long short-term memory; CNN: Convolutional neural network; DCNN: Deep convolutional neural network; DNN: Deep neural network; GRU: Gated recurrent unit; LSTM: Long short-term memory; MFCC: Mel-frequency cepstral coefficients; MLP: Multilayer perceptron; SALR: Speaker-adaptive learning and recognition; SincNet: Sinc-based convolutional neural network; TL: Transfer learning; WER: Word error rate.

of speech impairment. The experimental results show that the proposed approach achieves 95% classification accuracy and delivers strong precision, recall, and F1 scores compared with the CNN–LSTM, transformer-based, and hybrid baselines on the TORGO dataset. The results showed better generalization capabilities and higher interpretability, which is useful for implementing the proposed model in real-world dysarthria assessment scenarios.

3. Proposed work

Dysarthria is a motor speech disorder arising from neurological impairments that disrupt the mechanisms responsible for speech production. In recent years, automated dysarthria detection using deep learning techniques has attracted growing interest due to its potential to support early diagnosis, monitoring, and rehabilitation processes.[33] In this study, we present a CNN framework enhanced with attention mechanisms, namely the CBAM[34] and CA, to improve discriminative feature learning for dysarthria classification. The proposed approach leverages MFCCs as input representations, enabling effective modeling of both spectral and temporal characteristics of dysarthric speech.[35]

3.1. Data preprocessing

The datasets consist of several sets of speech recordings, organized by different speech conditions. The main

acoustic features used in this work are MFCCs, extracted from each audio sample for their effectiveness in representing the spectral properties of speech signals.[36] MFCC features are temporally aggregated by averaging features across time frames to produce a fixed-dimensional representation for neural network input. The dataset was split into 80% for training and 20% for testing, maintaining class proportions for a balanced representation of speech categories in both sets. Feature scaling involves normalizing the MFCC features using standardization, scaling them to zero mean and unit variance before model training. This normalization scales features to a common scale and helps the model converge quickly and reliably. In addition, the speech condition label was transformed into categorical speech conditions using one-hot coding, which enables effective multi-class classification by using a binary vector suitable for learning by a neural network for each class. CBAM is an attention mechanism that improves CNNs by adaptively refining feature maps through learned spatial and channel attention. The attention module is lightweight, has minimal computational overhead, and applies channel attention first, then spatial attention, for better performance.[37-40] This module improves feature representation by helping CNNs focus on important features while suppressing less relevant ones.

3.2. Coordinate attention

Coordinate attention enhances the model's ability to capture long-range dependencies by encoding spatial information into channel attention. It decomposes attention into two parallel transformations along the horizontal and vertical axes as shown in **Equation 1**.

$$Z_c^h=\sigma\left(W_h\left(GAP_x\left(F\right)\right)\right),\quad Z_c^v=\sigma\left(W_v\left(GAP_y\left(F\right)\right)\right) \quad (1)$$

where $\left(GAP_x\left(F\right)\right)$ and $\left(GAP_y\left(F\right)\right)$ are global average pooling operations along horizontal and vertical dimensions, and $\left(W_h\right)$ and $\left(W_v\right)$ are transformation functions.

These enhanced representations were combined to generate attention-enhanced features that improve positional information in speech signals. A deep CNN framework was developed to extract powerful and discriminative representations from the input data. The architecture consisted of a series of convolutional layers with different kernel sets that progressively learn spatial patterns at multiple levels of abstraction. After each convolutional layer, a rectified linear unit (ReLU) activation function helps model complex feature interactions, enables smooth gradient propagation through the network, and introduces nonlinearity. Max pooling was applied at various stages to control complexity, extract model-level features, and reduce spatial dimensions. These operations improve computational efficiency and increase invariance to local spatial variations. Too much pooling, however, can cause the loss of fine-grained spatial details. To mitigate this effect, attention mechanisms were introduced into the CNN pipeline to selectively preserve informative features. The CBAM was added to the network at various layers to dynamically adjust feature representations. CBAM uses a two-stage attention mechanism: first, it computes global context information via average and max pooling across channels and applies channel-wise attention to highlight informative feature channels; second, it applies spatial attention to these channels to enhance their information. Then, the network focuses on important structures and suppresses background information through spatial attention, identifying and highlighting salient regions in the feature maps. To further boost the performance of the network in modeling context information on different spatial scales, SPP was used, as shown in Figure 1. SPP pools features at several levels and extracts both local and global features without imposing a hard limit on input size. This multi-scale representation makes the model more flexible and robust, making it more applicable to real-world applications with inputs of different dimensions. The computed high-level feature representations were then passed through a sequence of fully connected layers where features were mapped to the target classification space. At the output layer, a softmax function was used to generate normalized probabilities for each class, ensuring reliability and understandable decision-making. The proposed CNN-based approach with CBAM effectively uses hierarchical convolutional feature learning, attention-guided refinement, and adaptive multi-scale pooling to boost the representation ability and classification accuracy of the CNN. The full structure of the proposed framework is shown in Figure 2.

4. Results and discussion

4.1. Dataset description

4.1.1. TORGO Dataset 1

The TORGO speech corpus is a multimodal dataset that integrates synchronized acoustic recordings with three-dimensional articulatory motion data collected from speakers affected by dysarthria and from neurologically healthy control participants. The dysarthric speech samples originate from individuals diagnosed with CP or ALS, which are among the most prevalent neurological conditions associated with motor speech impairments. The dataset was developed through a joint research initiative involving the Department of Computer Science and the Department of Speech-Language Pathology at the

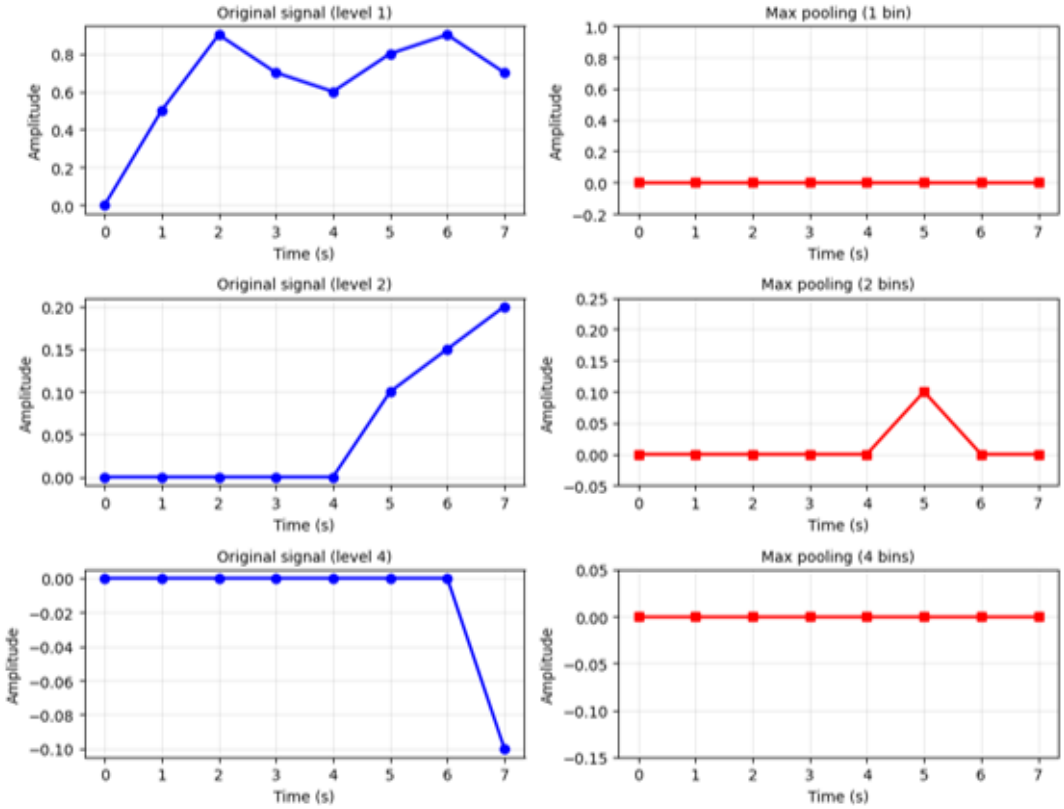

Figure 1. The spatial pyramid pooling operations

University of Toronto, in collaboration with the Holland Bloorview Kids Rehabilitation Hospital in Toronto. It contains about 2,000 speech recordings organized into four groups of speakers: female speakers with dysarthria, male speakers with dysarthria, female control speakers (without speech impairment), and male control speakers. Each category has the same number of recordings, and each category includes 500 audio samples from multiple recording sessions. This balanced composition provides uniform representation of both genders and clinical conditions, making the TORGO corpus suitable for exploring the characteristics of dysarthric speech and for creating data-driven speech analysis models.

4.1.2. TORGO Severity Dataset 2

This study was conducted using the TORGO Severity dataset, a publicly available multimodal database designed for research on dysarthric speech and speech disorders. TORGO Severity contains speech recordings from both speakers with dysarthria and neurologically healthy control speakers. It includes synchronized audio and articulatory information, making it one of the most widely used benchmark datasets for automatic dysarthric speech analysis. In this work, only the audio recordings were utilized for speech classification, while the articulatory and video modalities were not considered. To facilitate the multi-class dysarthria severity classification, the speech recordings were organized into four severity categories: normal, low, medium, and very low. The final dataset contains 1,776 audio recordings, divided into predefined training and testing subsets. The dataset consists of 1,776 FLAC audio files, equally distributed between the .flac and .FLAC file formats. The predefined dataset organization contains two subsets: training set: 1,418 recordings (79.8%), testing set: 358 recordings (20.2%).

As shown in Table 2, the dataset is imbalanced, with the most recordings in the very low severity class and fewer recordings in the low severity class. This imbalance reflects the natural distribution of the available speech recordings and was accounted for when training and evaluating the model. Each recording has a fixed length of about 3 seconds, with a consistent temporal window for feature extraction and classification. Several acoustic descriptors were calculated for each of the audio recordings to describe the spectral and

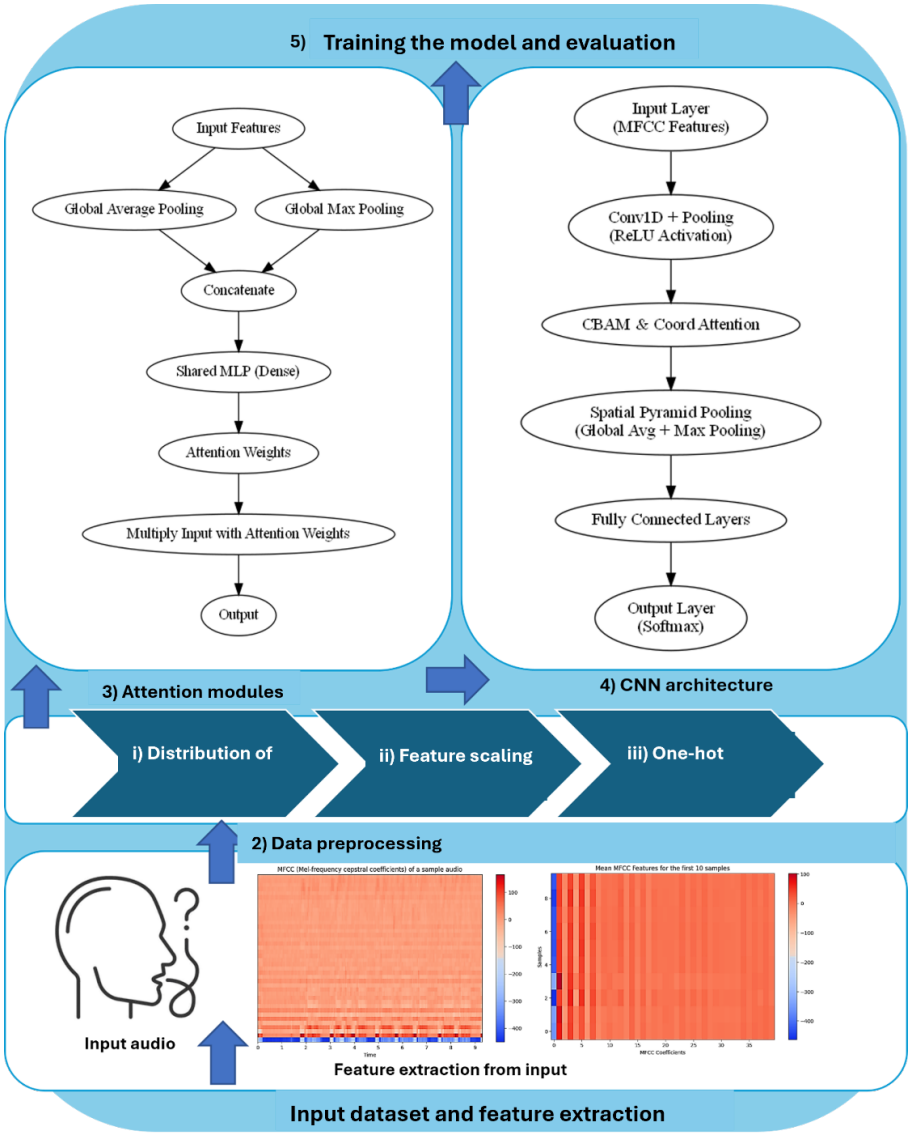

Figure 2. The general structure of the proposed framework
Abbreviations: CBAM: Convolutional block attention module; CNN: Convolutional neural network; Conv1D: One-dimensional convolution; Coord: Coordinate; MFCC: Mel-frequency cepstral coefficients; MLP: Multilayer perceptron; ReLU: Rectified linear unit.

temporal properties of speech: MFCCs, delta MFCCs, delta-delta MFCCs, root mean square (RMS) energy, zero-crossing rate (ZCR), spectral centroid, and spectral rolloff. These acoustic properties reflect complementary information on speech articulation, energy distribution, and spectral dynamics relevant to discriminating different severity levels of dysarthria. The statistics indicate significant acoustic differences between the severity groups. For instance, the average ZCR (0.3777) and the average spectral centroid (3,200.83 Hz) were greater in the case of the very low class than in the case of the low class, suggesting that the very low class has higher spectral variability, while the low class has higher RMS energy (0.2160) but less spectral centroid. These variations indicate that the chosen acoustic characteristics contain significant discriminatory information for dysarthria severity classification, as shown in Table 3 and Table 4. Table 5 compares performance with baseline methods over time, and Table 6 shows the top 10 most important SHAP features for each dysarthria severity class for Dataset 2. Figure 3 illustrates representative spectrograms from the four dysarthria severity categories in the TORGO Severity Dataset. Figure 4 presents representative speech waveforms from the four severity classes of the TORGO Severity dataset.

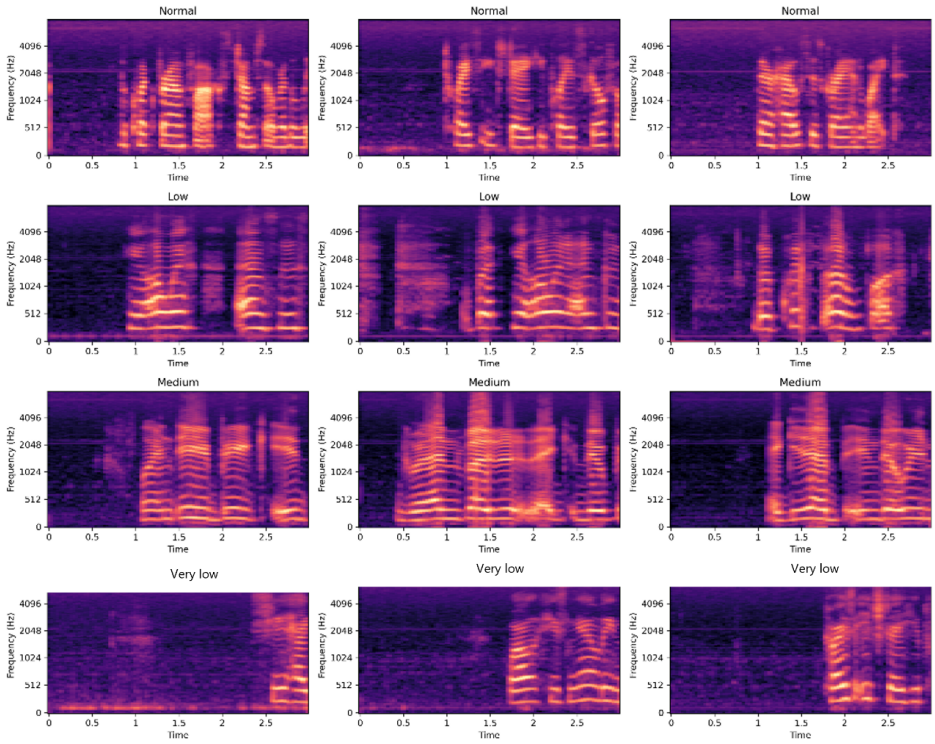

Figure 3. Sample spectrograms for TORGO Severity Dataset 2

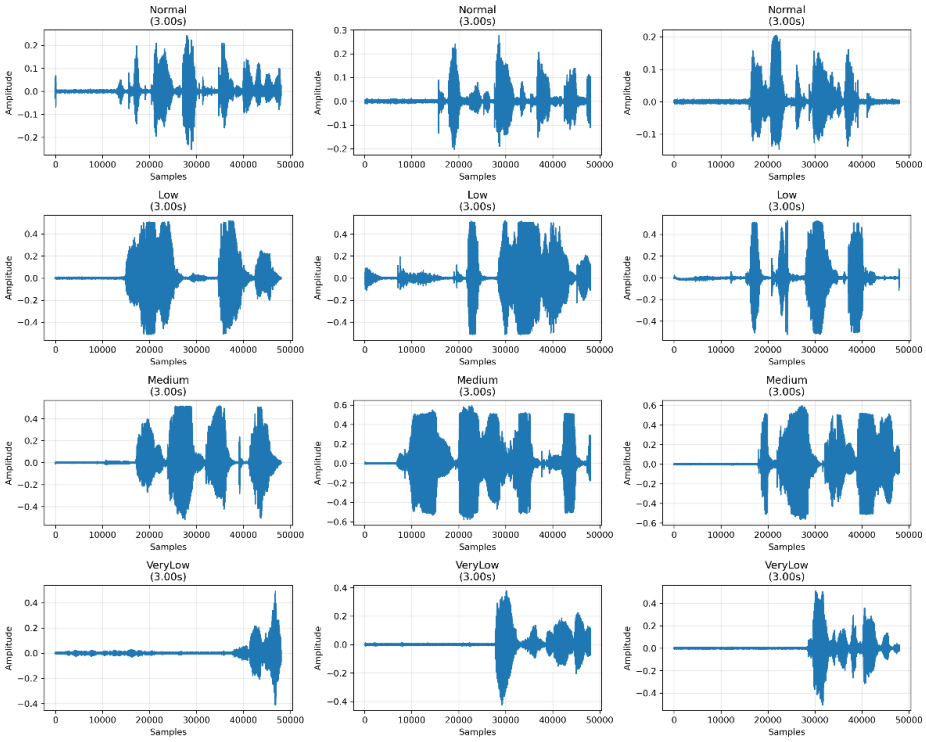

Figure 4. Sample waveforms for TORGO Severity Dataset 2

Table 2. The TORGO Severity Dataset 2 splitting

Class	Training	Testing	Total
Normal	400	100	500
Low	58	16	74
Medium	426	108	534
Very low	534	134	668
Total	1,418	358	1,776

4.2. Model evaluation

The proposed classification structure was evaluated with a controlled evaluation setup with an 80/20 split between training and testing data after completing the training process. To provide a comprehensive and objective assessment of the model's predictive power, a wide range of performance indicators was used. Accuracy was introduced as the first performance measure: the proportion of instances correctly classified out of the total number of samples. To discuss the reliability of class prediction in a more detailed way, the precision value was computed, which is defined as the number of correctly predicted positive instances as a fraction of the total number of positive instances predicted. Complementarily, sensitivity was used to assess how well the model identified true positive cases among all positive cases. The dataset was divided into training and testing sets using a stratified utterance-level split with an 80:20 ratio, maintaining proportional representation across severity classes. This approach ensures sufficient training data for each severity level while preventing class-imbalance artifacts during evaluation. However, we recognize that this split does not guarantee speaker independence, which may lead to optimistic performance estimates. To partially address this limitation, we also conducted 5-fold utterance-level cross-validation to assess model stability. The cross-validation results show consistent performance across folds, suggesting our model's robustness to data variability.

4.3. Experimental setup

To evaluate the performance of the proposed machine learning framework, a set of controlled experiments was conducted. All experiments were executed on a workstation featuring an Intel Core i5 processor operating at 3 GHz, 8 GB of system memory, and a 64-bit Windows 10 operating environment. The complete implementation, including model development, training, and evaluation, was carried out in Python (version 3.10.12).

4.4. Experimental results

Mel-frequency cepstral coefficients are widely used in speech and audio analysis because they capture the spectral qualities of sound in a way that resembles human auditory perception. Generally, MFCC extraction converts the time-domain signal to the frequency domain via a Fourier transform, projects it onto the perceptually motivated Mel scale, applies logarithmic compression, and finally applies a discrete cosine transform to obtain a sparse, decorrelated feature representation. The three MFCC feature plots are shown in Figure 5, representing three different audio samples. Figure 5 shows the distribution of MFCCs over the various coefficients. The lower-order coefficients have larger amplitudes in all samples, indicating substantial spectral energy in these coefficients. The higher-order coefficients, by contrast, show significant differences between samples, reflecting differences in spectral and acoustic properties. The third sample shows a larger drop in the first coefficient, which could correspond to variations in timbral or tonal quality, compared to the other samples.

Figure 6 presents two subplots comparing MFCC features before and after scaling. Figure 6A displays the raw MFCC values, where the x-axis represents the MFCC coefficients, and the y-axis represents their values. The first few coefficients exhibit a steep drop, followed by stabilization, with significant variations in magnitude that could introduce bias in machine learning models. In contrast, Figure 6B shows the MFCC features after scaling, likely through standardization or normalization, ensuring a more balanced distribution of values. The transformation removes dominant magnitude differences, making all coefficients contribute more equally in machine learning models. Unlike unscaled features, where the first few coefficients dominate, scaled features show more variability while maintaining relative differences. This preprocessing step is crucial for improving model performance and ensuring consistent feature importance across all coefficients.

Figure 7 shows a comparative visualization of the difference between the two feature representations learned by a CNN with two different attention mechanisms, CA and CBAM. The dimensions of the features extracted from the model are shown on the x-axis, and the time steps for the sequence of input signals are shown on the y-axis.

Table 3. Summary of representative acoustic statistics for the four severity classes

Class	Duration (s)	RMS energy (a.u.)	Zero-crossing rate (crossings/s)	Spectral centroid (Hz) Spectral centroid (Hz)
Normal	3.00 ± 0.00	0.0317 ± 0.0133	0.3090 ± 0.0742	2,755.97 ± 415.88
Low	3.00 ± 0.00	0.2160 ± 0.1168	0.1542 ± 0.0566	2,360.46 ± 110.71
Medium	3.00 ± 0.00	0.1214 ± 0.0450	0.2186 ± 0.0519	2,630.87 ± 472.36
Very low	3.00 ± 0.00	0.0316 ± 0.0149	0.3777 ± 0.0972	3,200.83 ± 519.15

Table 4. Training configuration summary

Parameter	Value	Description
Optimizer	AdamW	Adaptive moment estimation with weight decay
Learning rate	1e−3	Initial learning rate
Learning rate schedule	CosineAnnealingWarmRestarts	T_0 = 20 epochs, T_mult = 2
Batch size	32	Training and validation batch size
Maximum epochs	200	Maximum number of training epochs
Early stopping patience	30	Stop if no improvement for 30 epochs
Dropout rate	0.4	Dropout probability for regularization
Weight decay	1e−4	L2 regularization coefficient
Label smoothing	0.1	Smoothing factor for cross-entropy loss
Gradient clipping	1.0	Maximum gradient norm
Loss function	CrossEntropyLoss	With label smoothing
Device	CUDA/CPU	NVIDIA GPU when available

Table 5. Comparison with baseline methods +time for Dataset 2

Model	Accuracy	F1	Area under the curve	Inference speed
Proposed model	99.10%	0.9910	0.9999	~0.5–1.0 ms
Random forest	98.65%	0.9865	0.9997	~0.1 ms
Support vector machine (Radial basis function)	98.65%	0.9865	0.9999	~0.2 ms
Logistic regression	96.85%	0.9686	0.9982	~0.05 ms
K-nearest neighbor	95.95%	0.9597	0.9971	~0.1 ms
Decision tree	95.50%	0.9556	0.9641	~0.05 ms
Naive Bayes	78.15%	0.7828	0.9494	~0.05 ms

Table 6. Top 10 most important SHAP features for each dysarthria severity class for Dataset 2

Rank	Normal	SHAP value	Low	SHAP value	Medium	SHAP value	Very low	SHAP value
1	MFCC: Mean_13	0.020462	MFCC: Std_29	0.004268	MFCC: Mean_13	0.011032	MFCC: Mean_13	0.016625
2	MFCC: Std_27	0.009026	MFCC: Mean_17	0.003467	Delta: Std_34	0.010876	Delta: Mean_27	0.015835
3	MFCC: Mean_33	0.008366	RMS	0.001940	Delta: Std_0	0.008968	Delta2: Mean_23	0.012553
4	Spec_Rolloff	0.007146	Delta: Mean_33	0.001605	MFCC: Mean_0	0.008891	Delta: Std_4	0.011152
5	MFCC: Mean_14	0.007058	Delta2: Std_9	0.001568	MFCC: Mean_38	0.008505	ZCR	0.009776
6	Delta2: Std_11	0.006666	MFCC: Std_8	0.001562	Delta2: Std_6	0.008373	Delta: Std_0	0.009070
7	Delta2: Mean_1	0.005624	Delta2: Mean_35	0.001418	MFCC: Mean_33	0.007303	Delta Std_34	0.008629
8	Delta2: Std_10	0.005572	MFCC: Std_27	0.001243	Delta2: Std_7	0.007119	MFCC_: Mean_26	0.006963
9	Delta: Std_7	0.005494	MFCC: Mean_6	0.001125	ZCR	0.006614	MFCC: Std_37	0.006497
10	MFCC: Mean_4	0.005428	Delta: Std_0	0.001012	MFCC: Mean_12	0.006418	Delta2: Std_35	0.006404

Abbreviations: MFCC: Mel-frequency cepstral coefficients; RMS: Root mean square; SHAP: Shapley additive explanations; ZCR: Zero-crossing rate.

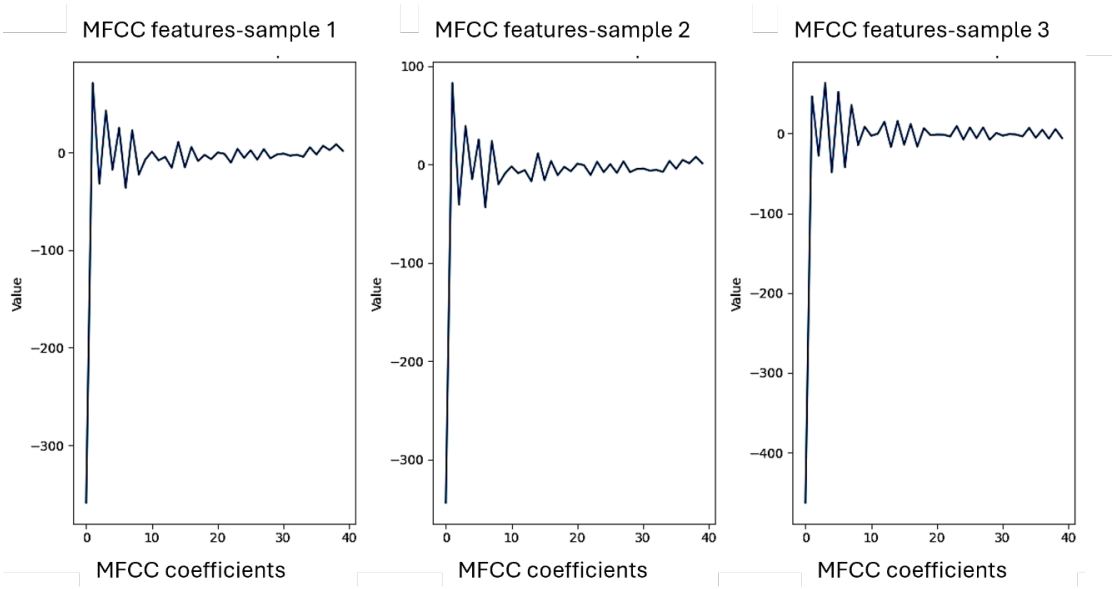

Figure 5. The resulting Mel-frequency cepstral coefficient (MFCC) features of the three applied audio samples for Dataset 1. (A) Sample 1. (B) Sample 2. (C) Sample 3.

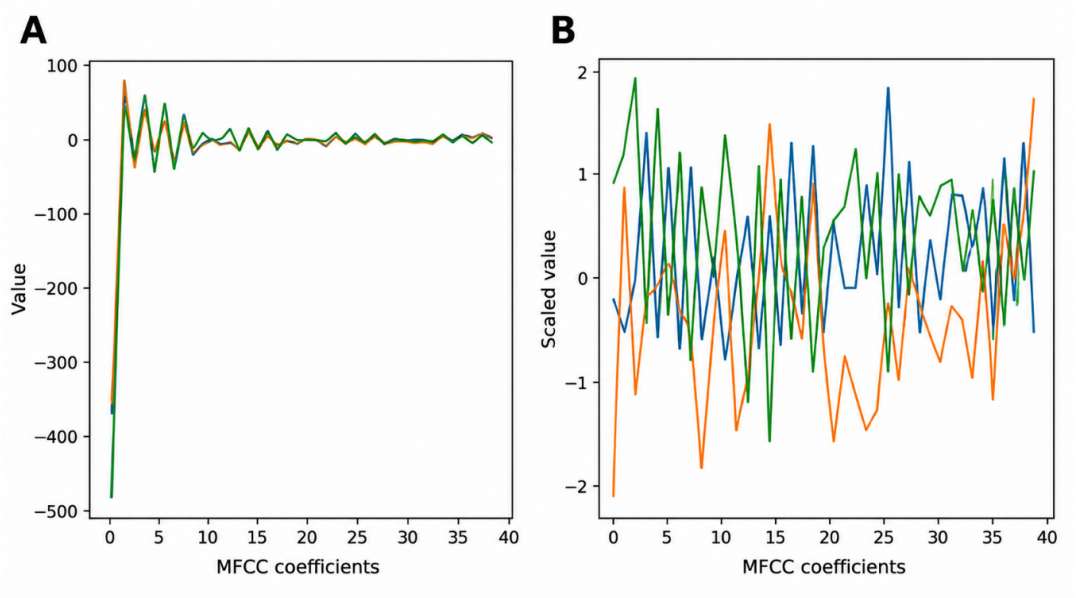

Figure 6. The Mel-frequency cepstral coefficient (MFCC) features (A) before and (B) after scaling for Dataset 1

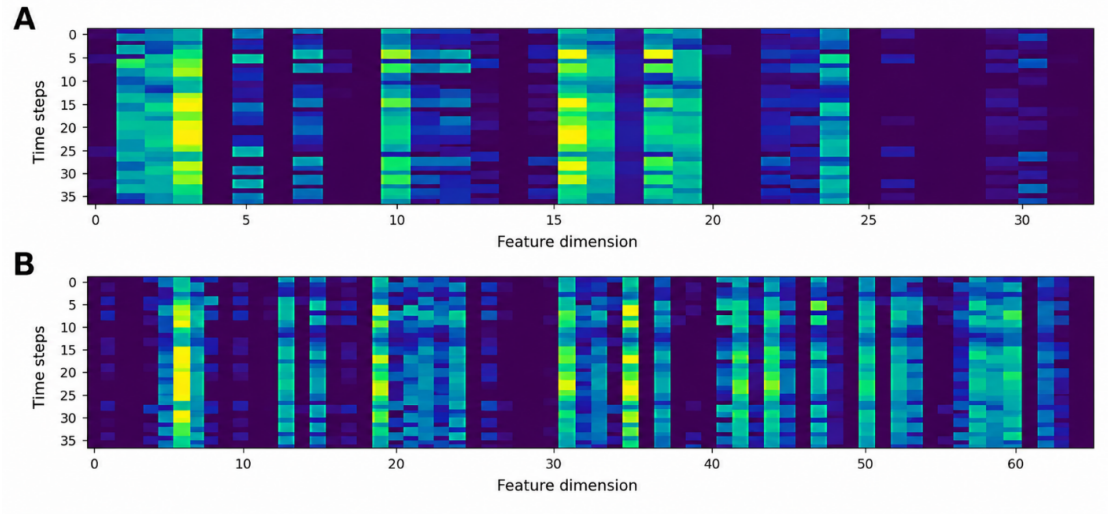

Figure 7. The heatmap of the applied dysarthria audio signal for the (A) coordinate attention and (B) convolutional block attention modules for Dataset 1

The color intensity represents heatmap activation, and brighter colors indicate higher attention weights. The CA mechanism improves spatial attention by incorporating location information, resulting in a more structured, well-separated feature representation, as shown in the top heatmap, where activations are more localized to distinct feature regions. CBAM, which processes channel and spatial attention separately, on the other hand, activates features in a much more spread-out distribution, as shown in the bottom heatmap, with features being activated more evenly among multiple dimensions. Comparative analysis of these attention mechanisms reveals their respective roles in feature learning in CNN-based models, which suggests that choosing a suitable attention mechanism is crucial for the specific audio classification task.

Figure 8 presents the training performance of a CNN model using accuracy and loss curves over 30 epochs.

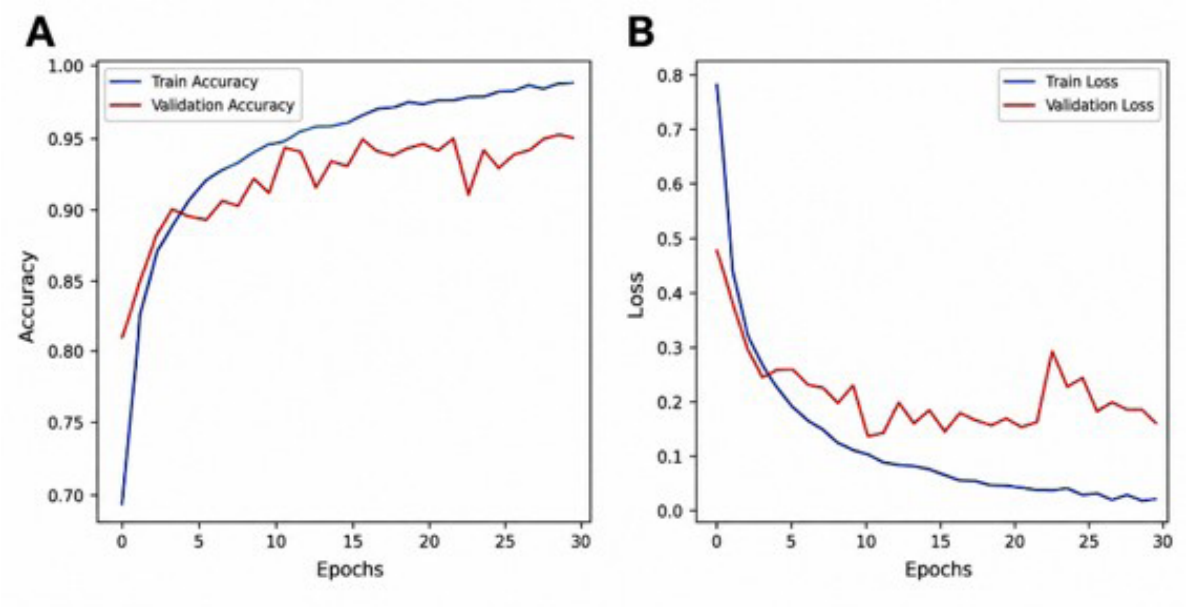

Figure 8. Performance of the proposed convolutional block attention module and convolutional neural network method on Dataset 1: (A) training and validation accuracy and (B) training and validation loss

Figure 8A shows the accuracy trends for both training and validation sets, while the right graph depicts the corresponding loss curves. Initially, both training and validation accuracy increased rapidly, with validation accuracy stabilizing around 95% while training accuracy continued to improve, reaching nearly 100%. This suggests effective learning but also a potential risk of overfitting. The loss curves in Figure 8B indicate a steady decrease in training loss, whereas validation loss fluctuates after an initial decline. The divergence between training and validation loss in later epochs suggests that the model might be overfitting; these results highlight the importance of regularization techniques such as dropout or early stopping to maintain generalization performance. The receiver operating characteristic curves shown in Figure 9 illustrate the classification performance of a model across different classes.

The proposed dysarthria classification framework, based on a CNN augmented with CA and the CBAM, demonstrates strong and consistent performance across gender-specific speech categories.

The SHAP values represent the impact of each feature on the model's prediction of each posture quality class. Feature_1 has the highest overall importance, with strong contributions to Class_0, Class_2, and Class_3, indicating it is a strong discriminative feature across posture categories, as shown in Figure 10. Feature_3 also plays an important role, with a significant influence on Class_2, whereas Feature_100 strongly influences differentiation between intermediate posture conditions in Class_3 and Class_1. Features such as Feature_703, Feature_1102, and Feature_701 are important within classes but less important overall. Figure 11 shows the scatter plots of the evaluation metrics collected over five folds. High performance is observed across all data partitions, as indicated by the small differences between folds. The detailed quantitative results are summarized in the classification performance information presented in Table 7, which shows the performance of each fold and the mean performance across all the folds. SHAP values of applied features for four-class classification of the proposed method for Dataset 1 are shown in Figure 12. Accuracy, precision, recall, F1-score, and AUC remain consistently high, ensuring the stability and reliability of the proposed architecture. Figure 11 shows the cross-validation performance for each fold, providing further evidence of the robustness and repeatability of the proposed framework on Dataset 2. Multiple statistical hypothesis tests were conducted to see if the performance improvements that were observed were statistically significant. Table 8 shows the results of the paired *t*-test, Wilcoxon signed-rank test, and McNemar's test performed

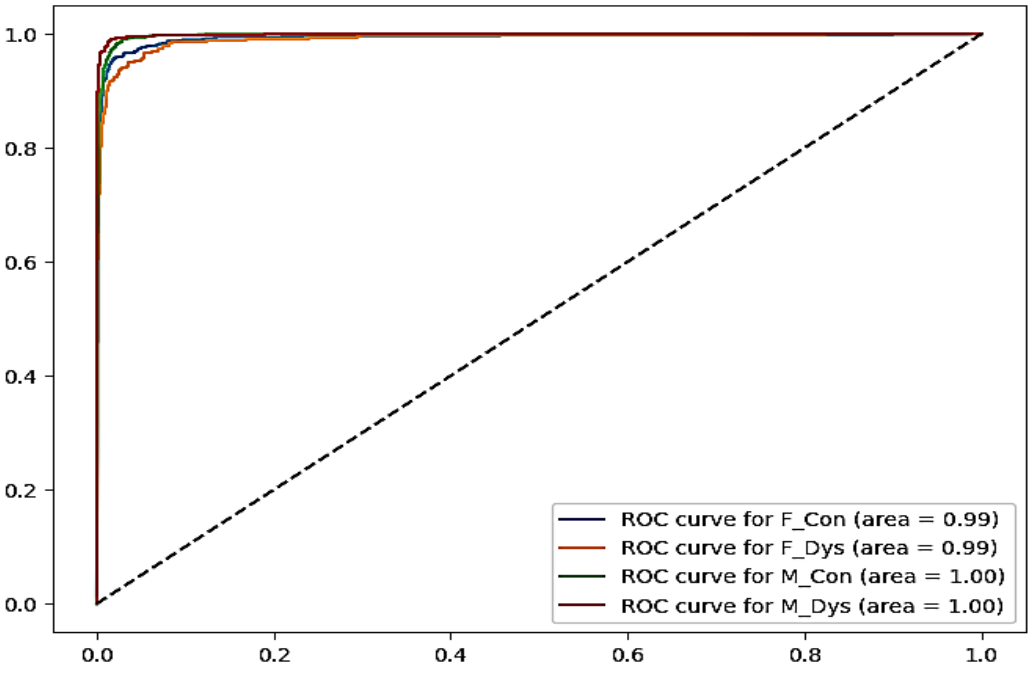

Figure 9. The receiver operating characteristic (ROC) curves for dysarthria classification across gender groups

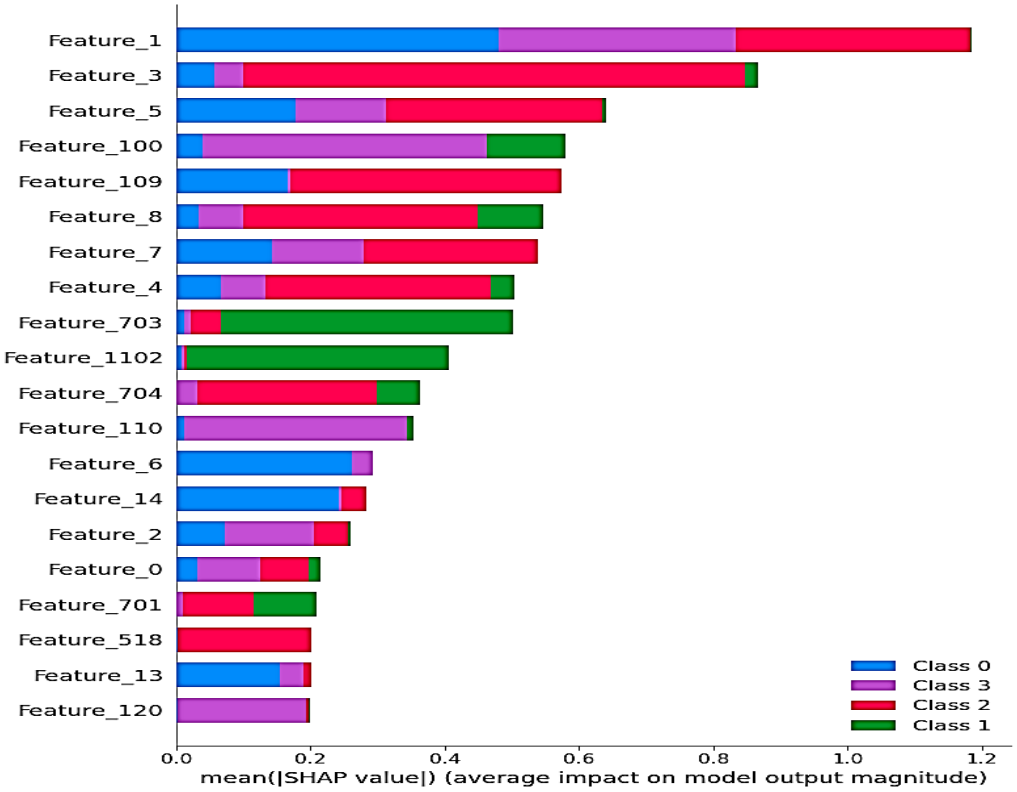

Figure 10. Shapley additive explanations (SHAP) values of applied features for four-class classification using the proposed method for Dataset 1

between the proposed framework and the baseline model. The *p*-values indicate that improvements in the proposed framework are statistically significant compared with the baseline methods, validating the effectiveness of the proposed attention-enhanced architecture. Table 9 shows the mean values, standard deviations, and 95% confidence intervals for all evaluation metrics, providing further analysis of the model's overall robustness. Narrow confidence intervals and low standard deviations indicate the stability of classification performance obtained by the proposed framework. The proposed model is compared with several baseline classifiers, such as traditional machine learning algorithms, in Table 10. The proposed multi-attention CNN consistently outperforms all classification evaluation metrics. Each architectural element was further explored using an ablation study (Table 11). Table 11 compares the baseline CNN with various combinations of CBAM, CA, and SPP. The results clearly show that individual components improve classification accuracy, and the full architecture achieves the highest overall accuracy. Table 12 presents the quantitative results, and Figure 12 compares the evaluation metrics across various random seeds. All experiments show negligible differences, indicating that the proposed framework is robust and random weight initialization has minimal impact. In this section, we show the ROC curves for multi-class dysarthria severity classification with the proposed attention-enhanced CNN (Dataset 2) in Figure 13. Table 13 summarizes the top 15 most influential acoustic features, and we consider the resulting feature importance rankings. The analysis shows that some MFCC features, along with their first-order and second-order temporal derivatives (Delta and Delta-delta features), play a significant role in the learned feature representations. Other conventional acoustic parameters, such as ZCR, spectral rolloff, and RMS, also aid classification. These results show the proposed model's ability to learn meaningful temporal and spectral speech characteristics that are highly discriminative for automatic dysarthria classification (Figure 14).[41-44]

Table 14 compares the accuracy results of recent studies and the proposed model on the TORGO dysarthric speech database. The studies exhibit a range of accuracies, from 70.48% (Stumpf *et al.*[22]) to 93.97% (Joshy and Rajan[20]), reflecting diverse methodologies and their effectiveness in dysarthric speech analysis. The proposed model achieves 95.00% accuracy, outperforming several existing approaches, including Alharbi *et al.*[28] (90.00%), Al-Ali *et al.*[30] (91.00%), and Shabber *et al.*[31] (90.00%). This highlights the model's potential to advance dysarthria detection and underscores ongoing progress in the field.

Table 7. Five-fold cross-validation performance of the proposed model (Dataset 2)

Metric	Mean	Standard deviation	95% confidence interval
Accuracy	99.32%	±0.25%	99.10–99.55
Precision	99.34%	±0.24%	99.13–99.55
Recall	99.32%	±0.25%	99.10–99.55
F1-score	99.32%	±0.25%	99.10–99.54
Area under the curve	99.73%	±0.58%	99.21–100.00

Table 8. Statistical significance analysis comparing the proposed framework with the baseline model (Dataset 2)

Statistical test/metric	Value
Paired *t*-test statistic	0.5019
Paired *t*-test *p*-value	0.6160
Wilcoxon signed-rank statistic	21,116.5
Wilcoxon signed-rank *p*-value	0.5529
McNemar test (Class 0)	0.0000
McNemar test (Class 1)	5.43×10^{-9}
McNemar test (Class 2)	0.0000
McNemar test (Class 3)	0.0000
Accuracy (Proposed model)	99.10%
Accuracy (baseline model)	32.66%

Table 9. Performance stability and 95% confidence interval of the proposed framework (Dataset 2)

Metric	Mean (%)	Standard deviation	95% confidence interval
Accuracy	99.10	0.0046	98.20–99.78
Precision	99.11	0.0046	98.20–99.78
Recall	99.10	0.0046	98.20–99.78
F1-Score	99.10	0.0046	98.20–99.78
Area under the curve	99.99	0.0001	99.96–100.00

Table 10. Comparative performance of the proposed framework and baseline machine learning models (Dataset 2)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Area under the curve (%)
Random forest	98.65	98.65	98.65	98.65	99.97
Support vector machine (Radial basis function)	98.65	98.65	98.65	98.65	99.99
Logistic regression	96.85	96.92	96.85	96.85	99.82
Decision tree	95.50	95.93	95.50	95.56	96.41
K-nearest neighbor	95.95	96.03	95.95	95.97	99.71
Naïve Bayes	78.15	78.56	78.15	78.28	94.94

Table 11. Comparative performance of the proposed model and its architectural variants

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Area under the curve (%)
Simple CNN (Baseline)	99.55	99.55	99.55	99.55	99.99
CNN + Attention	99.10	99.10	99.10	99.10	99.98
CNN + Residual	99.10	99.10	99.10	99.10	99.99

Abbreviation: CNN: Convolutional neural network

Table 12. Performance stability of the proposed model across different random seeds

Random seed	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Area under the curve (%)
1	99.55	99.56	99.55	99.55	100.00
7	99.55	99.56	99.55	99.55	100.00
21	99.55	99.55	99.55	99.54	100.00
42	99.55	99.55	99.55	99.55	99.995
84	99.55	99.55	99.55	99.55	99.83
123	100.00	100.00	100.00	100.00	100.00

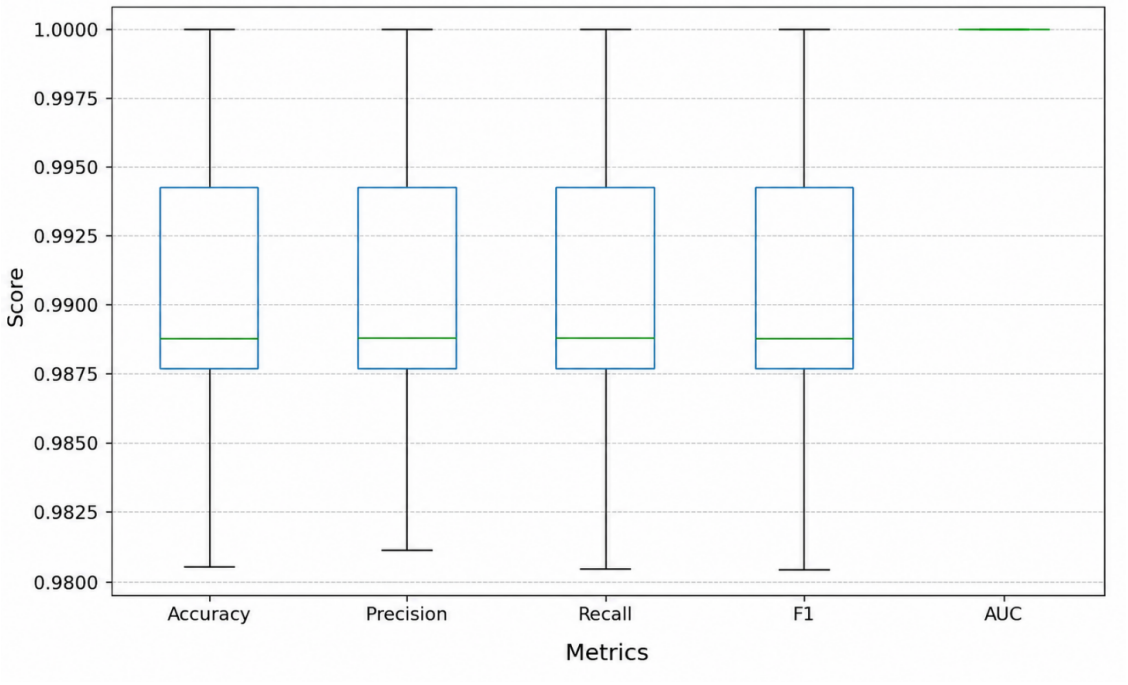

Figure 11. The cross-validation (Dataset 2)

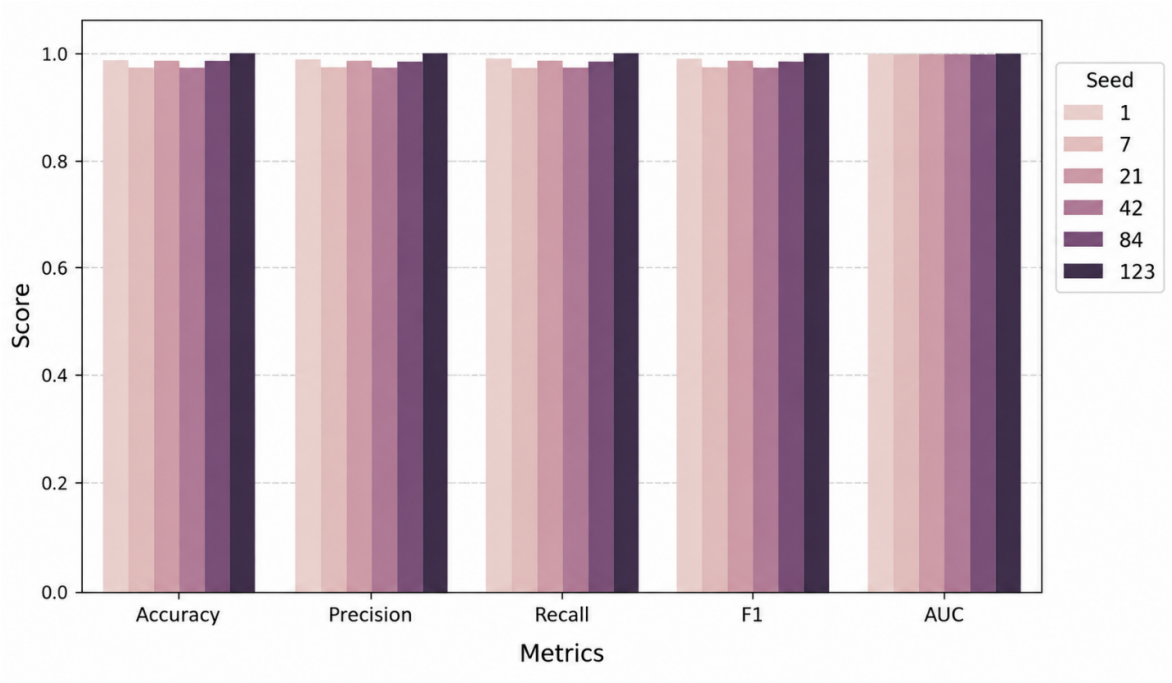

Figure 12. Performance of the proposed attention-enhanced convolutional neural network across different random seeds (Dataset 2)

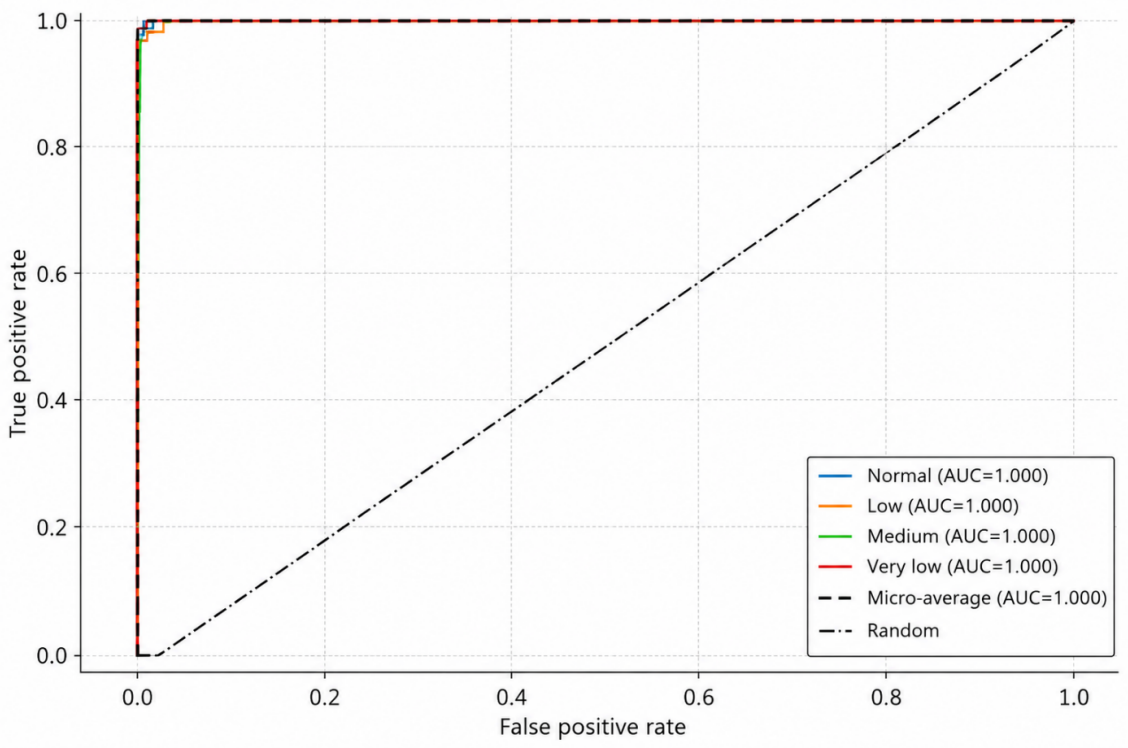

Figure 13. Receiver operating characteristic (ROC) curves for multi-class dysarthria severity classification using the proposed attention-enhanced convolutional neural network (Dataset 2)
Abbreviation: AUC: Area under the curve.

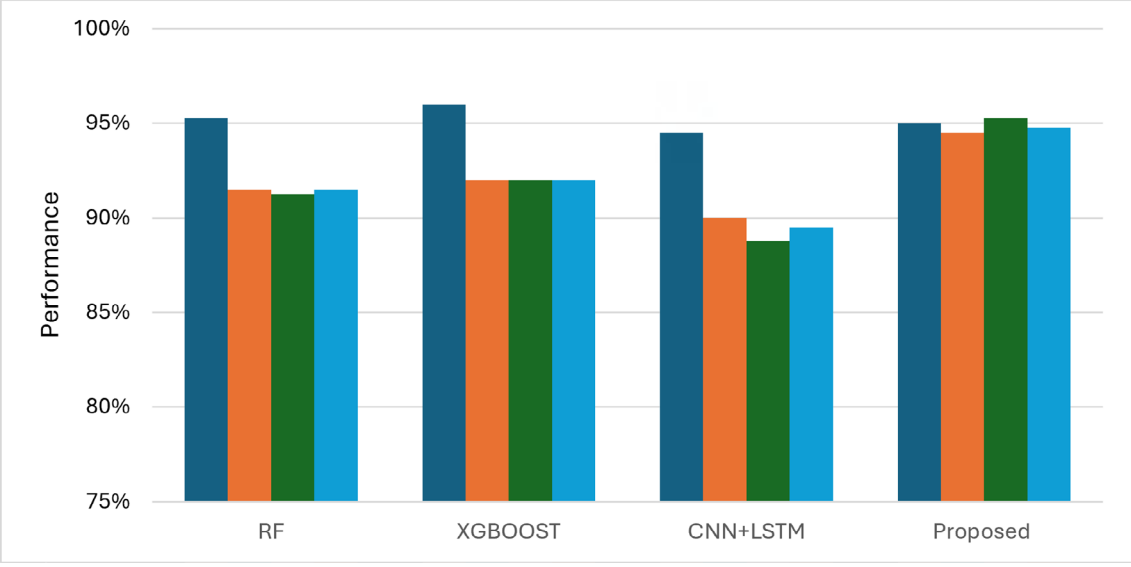

Figure 14. Comparative analysis of the proposed method with the random forest (RF), extreme gradient boosting (XGBOOST), and convolution neural network (CNN) + long short-term memory (LSTM) approaches (Dataset 1)

Table 13. Top 15 most important features identified by the proposed model

Rank	Feature	Global importance
1	F164	0.004188
2	F241	0.004007
3	F233	0.003593
4	F228	0.003538
5	F119	0.003405
6	F23	0.003123
7	F76	0.003063
8	F237	0.002947
9	F50	0.002821
10	F34	0.002547
11	F32	0.002507
12	F136	0.002505
13	F10	0.002477
14	F192	0.002313
15	F24	0.002259

Table 14. Comparative accuracy results of recent studies and the proposed model on the TORGO dysarthric speech database

Author	Results
Joshy and Rajan[20]	93.97%
Javanmardi *et al.*[21]	76.12%
Stumpf *et al.*[22]	70.48%
Alharbi *et al.*[28]	90.00%
Al-Ali *et al.*[30]	91.00%
Shabber *et al.*[31]	90.00%
Proposed work	95.00%

5. Conclusion

The findings demonstrate the strong effectiveness of the proposed CNN-based dysarthria detection framework enhanced with attention mechanisms, specifically CBAM and CA. The proposed model surpasses previously reported approaches, underscoring its robustness in identifying dysarthric speech associated with neurological conditions such as CP and ALS. The consistently high precision, recall, and F1-scores across both dysarthric and control speech categories further confirm the model's capacity to generalize across diverse speaker characteristics. A key strength of the proposed framework lies in its attention-driven feature learning strategy. CBAM enhances representation quality by sequentially modeling channel-wise and spatial attention, allowing the network to prioritize salient acoustic cues while suppressing redundant or less informative features. Complementarily, CA captures long-range dependencies by embedding positional information into channel attention, which is particularly beneficial for modeling the temporal and spectral complexity of dysarthric speech. In addition, incorporating SPP enables multi-scale feature aggregation, improving robustness to variations in speech duration and supporting the extraction of both local and global contextual information. Despite these advantages, several limitations merit consideration. While overall performance is strong, slight variability was observed across specific subject groups. In particular, the F1-score for female dysarthric speech (0.90) was marginally lower than that of other classes, suggesting that gender-related differences in dysarthric articulation patterns may influence classification outcomes. The findings demonstrate the strong effectiveness of the proposed attention-enhanced CNN framework for automatic dysarthria severity classification. By integrating the CBAM, CA, and SPP, the proposed architecture effectively learns discriminative multi-scale acoustic representations while emphasizing informative temporal and spectral speech characteristics. Extensive experiments conducted on two publicly available datasets demonstrated consistently strong performance, achieving an overall classification accuracy of 99.10%, an F1-score of 0.9910, and an AUC of 0.9999, while outperforming conventional machine learning and deep learning baselines. Additional validation through cross-validation, repeated random-seed experiments, ablation studies, statistical significance testing, and SHAP-based explainability further confirms the robustness, stability, and interpretability of the proposed framework.

Despite these encouraging results, several limitations should be acknowledged. The proposed model was evaluated on publicly available datasets, which, although

widely used, remain relatively limited in size and diversity compared with real-world clinical populations. Performance may also be influenced by class imbalance, inter-speaker variability, recording conditions, and differences in dysarthria severity distributions. Moreover, slight variations observed across certain subject groups indicate that demographic and clinical variability may still affect classification performance. Therefore, the proposed framework should be regarded as a computer-aided decision-support system designed to assist clinicians rather than replace expert clinical assessment.

Future work should focus on validating the proposed framework with larger, more diverse datasets collected in real clinical environments, evaluating its integration into routine clinical workflows, assessing its effectiveness as a diagnostic decision-support tool, and conducting usability studies involving clinicians and speech-language pathologists. Additionally, future research should assess the robustness of the model under varying recording conditions and across different healthcare institutions through multi-center clinical validation. Moreover, the computational demands of deep learning models with attention mechanisms may pose challenges for deployment on resource-constrained platforms. Therefore, future work should also explore model compression techniques, lightweight architectures, and edge-optimized implementations to facilitate real-time deployment in practical assistive and clinical applications.

Acknowledgments

None.

Funding

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number PSAU/2025/01/34420.

Conflict of interest

The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions

Conceptualization: Zahraa Tarek, Esraa Hassan, Ahmed Fahim, Khaled Alnowaiser, Tarek Abd El-Hafeez

Investigation: Zahraa Tarek, Esraa Hassan, Tarek Abd El-Hafeez, Mahmoud Y. Shams

Methodology: Zahraa Tarek, Esraa Hassan, Khaled Alnowaiser

Writing–original draft: Zahraa Tarek, Mahmoud Y. Shams, Esraa Hassan, Tarek Abd El-Hafeez

Writing–review & editing: Esraa Hassan, Khaled Alnowaiser, Mahmoud Y. Shams

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The datasets analyzed in this study can be accessed through the following links: Dataset 1: https://www.kaggle.com/datasets/pranaykoppula/torgo-audio; Dataset 2: https://www.kaggle.com/datasets/speechlab01/torgo-severity

References

1. Singh N, Tripathi P. Explainable SI-SAPN: signal imagery-based spatial attention pyramidal network model for Parkinson's disease detection using speech spectrograms. *Int J Inf Technol.* 2025:1-21.

   doi: 10.1007/s41870-025-02693-9

2. Zhang Z, Tang C, Yi W, Xu M, Occhipinti LG. Machine learning-driven smart socks with printed strain sensors for fall detection. In: *2025 IEEE BioSensors Conference (BioSensors)*. San Diego, CA, USA: IEEE; 2025:1-4.

   doi: 10.1109/BioSensors65002.2025.11239179

3. Wen-Tsai S, Hao-Wei K, Sung-Jung H. Speech recognition via CTC-CNN model. *Comput Mater Contin.* 2023;76(3):3833-3848.

   doi: 10.32604/cmc.2023.040024

4. Singh S, Wang Q, Zhong Z, *et al.* Robust Cross-Etiology and Speaker-Independent Dysarthric Speech Recognition. *arXiv*. Preprint posted online 2025.

   doi: 10.48550/arXiv.2501.14994

5. Spencer KA, Eddy B, Papathanasiou I, Summers D, Britton D. Management of velopharyngeal impairment in adults with dysarthria: a systematic review. *Am J Speech Lang Pathol.* 2025;34(1):391-409.

   doi: 10.1044/2024_AJSLP-24-00287

6. Hao J, Pan H. A novel deep transformer-based CvT model for sign language recognition in visual communication. *Sci Rep.* 2025:16(1).

   doi: 10.1038/s41598-025-31558-1

7. Hasan MT, Hossain MAE, Mukta MSH, Akter A, Ahmed M, Islam S. A review on deep learning-based cyberbullying detection. *Future Internet.* 2023;15(5):179.

   doi: 10.3390/fi15050179

8. Azad M, Khan MFK, El-Ghany SA. XAI-enhanced machine

learning for obesity risk classification: a stacking approach with LIME explanations. *IEEE Access.* 2025;13:13847-13865.

doi: 10.1109/ACCESS.2025.3530840

9. Shams MY, Tarek Z, Elshewey AM. A novel RFE-GRU model for diabetes classification using PIMA Indian dataset. *Sci Rep.* 2025;15(1):982.

doi: 10.1038/s41598-024-82420-9

10. Liu S, Geng M, Hu S, *et al.* Recent progress in the CUHK dysarthric speech recognition system. *IEEE/ACM Trans Audio Speech Lang Process.* 2021;29:2267-2281.

doi: 10.1109/TASLP.2021.3091805

11. Hassan E, Elbedwehy S, Shams MY, Abd El-Hafeez T, El-Rashidy N. Optimizing poultry audio signal classification with deep learning and burn layer fusion. *J Big Data.* 2024;11(1):135.

doi: 10.1186/s40537-024-00985-8

12. Qian Z, Xiao K, Yu C. A survey of technologies for automatic dysarthric speech recognition. *EURASIP J Audio Speech Music Process.* 2023;2023(1):48.

doi: 10.1186/s13636-023-00318-2

13. He Y, Seng KP, Lim CS, Ang LM. Robust dysarthric speech recognition with GAN enhancement and LLM correction. *Adv Intell Syst.* 2026;8(2):e202500873.

doi: 10.1002/aisy.202500873

14. Hertrich I, Dietrich S, Blum C, Ackermann H. The role of the dorsolateral prefrontal cortex for speech and language processing. *Front Hum Neurosci.* 2021;15:645209,

doi: 10.3389/fnhum.2021.645209

15. Basilakos A, Fridriksson J. Types of motor speech impairments associated with neurologic diseases. In: *Handbook of Clinical Neurology*. Elsevier; 2022:71-79.

doi: 10.1016/b978-0-12-823384-9.00004-9

16. Chung Y, Hong J, Lee J, Kim E. Diagnosis-aware multitask fine-tuning of Whisper for dysarthric speech recognition. *Speech Commun.* 2026;180,103393.

doi: 10.1016/j.specom.2026.103393

17. Choi J, Moya-Galé G, Hwang K, Hirschberg J, Levy ES. Automatic speech recognition for intelligibility assessment in children with dysarthria. *J Speech Lang Hear Res.* 2026;69(4):1438-1454.

doi: 10.1044/2025_JSLHR-25-00562

18. Kim Y, Kent RD, Weismer G. An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria. *J Speech Lang Hear Res.* 2011;54(2):417-429.

doi: 10.1044/1092-4388(2010/10-0020)

19. Narendra NP, Alku P. Glottal source information for pathological voice detection. *IEEE Access.* 2020;8:67745-67755.

doi: 10.1109/ACCESS.2020.2986171

20. Joshy AA, Rajan R. Automated dysarthria severity classification: a study on acoustic features and deep learning techniques. *IEEE Trans Neural Syst Rehabil Eng.* 2022;30:1147-1157.

doi: 10.1109/TNSRE.2022.3169814

21. Javanmardi F, Kadiri SR, Alku P. Pre-trained models for detection and severity level classification of dysarthria from speech. *Speech Commun.* 2024;158:103047.

doi: 10.1016/j.specom.2024.103047

22. Stumpf L, Kadirvelu B, Waibel S, Faisal AA. Speaker-independent dysarthria severity classification using self-supervised transformers and multi-task learning. *arXiv.* 2024.

doi: 10.48550/arXiv.2403.00854

23. Shabber SM, Sumesh EP. AFM signal model for dysarthric speech classification using speech biomarkers. *Front Hum Neurosci.* 2024;18:1346297.

doi: 10.3389/fnhum.2024.1346297

24. Radha K, Bansal M. Automated detection and severity assessment of dysarthria using raw speech. In: *2023 14th International Conference on Computing, Communication and Networking Technologies (ICCCNT).* Delhi, India: IEEE; 2023:1-7.

doi: 10.1109/ICCCNT56998.2023.10307923

25. Radha K, Bansal M, Dulipalla VR. Variable STFT layered CNN model for automated dysarthria detection and severity assessment using raw speech. *Circuits Syst Signal Process.* 2024;43(5):3261-3278.

doi: 10.1007/s00034-024-02611-7

26. Shanmugapriya P, Mohan V. Comparative analysis of deep learning models for dysarthric speech detection. *Soft Comput.* 2023;28(6):5683-5698.

doi: 10.1007/s00500-023-09302-6

27. Sekhar SM, Kashyap G, Bhansali A, Singh K. Dysarthric speech detection using transfer learning with convolutional neural networks. *ICT Express.* 2022;8(1):61-64.

doi: 10.1016/j.icte.2021.07.004

28. Alharbi GF, Alamri NK, Sabbeh SF. Automatic classification of speech dysarthric intelligibility levels using textual feature. *IEEE Access.* 2025;13:39982-39992.

doi: 10.1109/ACCESS.2025.3547041

29. S A, R P, M R. Comparative analysis of different time-frequency image representations for the detection and severity classification of dysarthric speech using deep learning. *Results Eng.* 2025;25:104561.

doi: 10.1016/j.rineng.2025.104561

30. Al-Ali AS, Haris RM, Akbari Y, Saleh M, Al-Maadeed S, Kumar MR. Integrating binary classification and clustering for multi-class dysarthria severity level classification: a two-stage approach. *Cluster Comput.* 2025;28(2):136.

doi: 10.1007/s10586-024-04748-1

31. Shabber SM, Sumesh EP, Ramachandran VL. Scalogram-based performance comparison of deep learning architectures for dysarthric speech detection. *Artif Intell Rev.* 2025;58(5):128.

doi: 10.1007/s10462-024-11085-7

32. Wang Q, Zhong Z, Singh S, *et al.* Dysarthric Speech Conformer: adaptation for sequence-to-sequence dysarthric speech recognition. In: *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Hyderabad, India: IEEE; 2025:1-5.

doi: 10.1109/ICASSP49660.2025.10889046

33. Enderby P. Disorders of communication. In: *Handbook of Clinical Neurology*. Amsterdam, The Netherlands: Elsevier; 2013:273-281.

doi: 10.1016/B978-0-444-52901-5.00022-8

34. Zhang Z, Wang M. Convolutional neural network with convolutional block attention module for finger vein recognition. *arXiv*. Preprint posted online 2022.

doi: 10.48550/arXiv.2202.06673

35. Zakariah M, Alnuaim A. Recognizing human activities with the use of convolutional block attention module. *Egypt Inform J.* 2024;27:100536.

doi: 10.1016/j.eij.2024.100536

36. Aryal N, Lee SW. Frequency-based CNN and attention module for acoustic scene classification. *Appl Acoust.* 2023;210:109411.

doi: 10.1016/j.apacoust.2023.109411

37. Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional Block Attention Module. In: *Computer Vision – ECCV 2018 (Lecture Notes in Computer Science)*. Cham, Switzerland: Springer International Publishing; 2018:3-19.

doi: 10.1007/978-3-030-01234-2_1

38. Chen L, Yao H, Fu J, Ng CT. The classification and localization of crack using lightweight convolutional neural network with CBAM. *Eng Struct.* 2023;275:115291.

doi: 10.1016/j.engstruct.2022.115291

39. Rokhva S, Teimourpour B. Accurate and real-time food classification through the synergistic integration of EfficientNetB7, CBAM, transfer learning, and data augmentation. *Food Humanit.* 2025;4:100492.

doi: 10.1016/j.foohum.2024.100492

40. Hassan E, Shams MY, Hikal NA, Elmougy S. A novel convolutional neural network model for malaria cell images classification. *Comput Mater Contin.* 2022;72(3):5889-5907.

doi: 10.32604/cmc.2022.025629

41. Sarhan S, Nasr AA, Shams MY. Multipose Face Recognition-Based Combined Adaptive Deep Learning Vector Quantization. *Comput Intell Neurosci*. 2020;2020:1-11.

doi: 10.1155/2020/8821868

42. Hassan E, Shams MY, Hikal NA, Elmougy S. The effect of choosing optimizer algorithms to improve computer vision tasks: a comparative study. *Multimed Tools Appl.* 2023;82(11):16591-16633.

doi: 10.1007/s11042-022-13820-0

43. Joshy AA, Parameswaran PN, Janardan S, TK T, Rajan R. How dysarthria severity influences speech recognition in cerebral palsy. *J Neurolinguistics.* 2026;79:101348,

doi: 10.1016/j.jneuroling.2026.101348

44. Altulyan M. A robust and integrated speech recognition tool for dysarthria patients using lip movement recognition. *Eng Technol Appl Sci Res.* 2026;16(3):35660-35669.

doi: 10.48084/etasr.17799