Ingénierie des Systèmes d'Information
Vol. 31, No. 2, February, 2026, pp. 323-333
Journal homepage: http://iieta.org/journals/isi

Deep Learning Based Dysarthric Speech Recognition Using Improved Chaotic Archimedes Optimization Algorithm for Acoustic Features Selection

Namita Kure[1,2*], Somnath B. Dhonde[3]

[1] Department of E & TC, AISSMS Institute of Information Technology, Savitribai Phule Pune University, Pune 411001, India
[2] Department of E & TC, Dr. D. Y. Patil Institute of Technology, Savitribai Phule Pune University, Pimpri, Pune 411018, India
[3] Department of E & TC, AISSMS College of Engineering, Savitribai Phule Pune University, Pune 411001, India

Corresponding Author Email: npachling@gmail.com

https://doi.org/10.18280/isi.310201

Received: 3 November 2025
Revised: 10 January 2026
Accepted: 17 February 2026
Available online: 28 February 2026

Keywords:
dysarthric speech recognition, deep learning, deep convolutional neural network, bidirectional long short-term memory, gated recurrent unit, speech processing

ABSTRACT

Dysarthria is a neurological speech disorder resulting from weakness in the muscles controlling speech production and articulation, which hinders effective communication. Over the last decade, automatic speech recognition (ASR) has experienced tremendous growth in human-computer interactive systems; however, the effectiveness and reliability of traditional ASR are limited for dysarthric speech recognition (DSR), as these systems are typically developed for everyday speech. Many researchers have presented the DSR system using machine learning (ML) and deep learning (DL) techniques; however, its effectiveness is limited by poor voice intelligibility, class imbalance, limited generalization capability, complex DL architectures, and intra- and inter-class feature variability. This paper presents the hybrid DL framework DBGNet, which combines a deep convolutional neural network (DCNN), bidirectional long short-term memory (BiLSTM), and a gated recurrent unit (GRU) to enhance feature correlation, long-term connectivity, and temporal feature depiction. The DBGNet accepts multiple dysarthric features comprising time-domain features (TDF), frequency-domain features (FDF), and phonatory features (PF). It uses an improved Archimedes optimization algorithm (IAOA) for feature selection, based on a dynamic chaotic tent map (DCTM) and a neighborhood-guided scheme (NGS) to enhance solution diversity, global search exploration, the search space, and the AOA's convergence. The IAOA + DBGNet provides an overall accuracy of 99.33% for the UASpeech and 98.30% for the TORGO dataset that shows a significant boost over the DSR accuracy of traditional techniques.

1. INTRODUCTION

Massive growth in Industry 4.0 leads to the use of artificial intelligence for automation in human-computer interaction systems [1]. An automatic speech recognition (ASR) system converts spoken input into text or control commands. The speech is a natural, reliable, robust, and non-invasive way of communication [2, 3]. The ASR captures the semantic and contextual information from the voice to recognize the speaker and spoken content. However, many disorders limit the efficacy of ASR systems due to reduced voice intelligibility [4, 5].

Dysarthria is a major motor voice disorder that results from vocal muscle weakness, reduced coordination of speech muscles, slurred speech, hyperactivity, hesitation, and uncoordinated behavior [6]. The dysarthria causes rapid or slow speech, slurred speech, a monotone or breathy voice, and loudness [7]. It has a significant impact on voice characteristics, including phonation, articulation, resonance, prosody, naturalness, and intelligibility [8, 9]. Dysarthria is typically diagnosed by a speech-language pathologist (SLP) based on a physical examination, patient history, and an assessment of the patient's ability to produce words, sounds, and sentences. However, ASR is failing due to the voice's lower intelligibility, which lacks the imperative features [10].

Various machine learning (ML) and deep learning (DL) techniques have been presented for dysarthric speech recognition (DSR). Soleymanpour et al. [11] explored dysarthric speech synthesis with controlled severity, pauses, and prosodic adjustments. Augmenting ASR training with both real and synthetic data reduced word error rate (WER) by 12.2%, with further gains from severity-aware synthesis. The main challenge lies in maintaining naturalness and bridging the domain gap between synthetic and real speech. Almadhor et al. [12] advanced visual-speech dysarthric ASR by modeling phoneme shapes with convolutional and attention-based networks. Their visual-based system demonstrated substantial improvements, particularly for very low-intelligibility subsets, although challenges persisted in collecting and integrating visual data. Lin et al. [13] proposed a cognitively inspired cognition-inspired feature decomposition and recombination network (CFDRN) model that decomposed slow and rapid temporal features for fusion on self-supervised learning (SSL) backbones. Tested on TORGO and UASpeech, the system consistently reduced WERs by 13–16% and 4–13%, respectively. The drawback,

however, was the increased complexity of training and sensitivity to backbone choice.

Geng et al. [14] introduced spectro-temporal embeddings via singular value decomposition and integrated them into hybrid ASR systems with speaker adaptation. Results showed WER reductions across multiple corpora, but the approach required a complex feature pipeline and careful adaptation tuning. Rajeswari et al. [15] combined variational mode decomposition and wavelet denoising with Convolutional Neural Networks (CNNs), achieving 95.95% accuracy on UASpeech. The enhancement improved dysarthric speech features, though the setup was limited to isolated-word recognition. Yue et al. [16] and Yue et al. [17] explored source–filter decomposition, separating magnitude spectra into distinct streams for CNN–RNN (Recurrent Neural Network) fusion. Experiments on TORGO and UASpeech showed reductions in WER relative to MFCC baselines, with up to a 1.7% absolute improvement. While source–filter separation normalized speaker style effectively, WERs remained relatively high. Shahamiri et al. [18] later extended Transformer-based ASR with transfer learning and augmentation. Speaker-adaptive models trained on UASpeech demonstrated subject-level accuracy improvements of up to 23%. Although effective, absolute WERs remained high due to data scarcity.

Irshad et al. [19] shifted from recognition to detection, using UTran-DSR with Vision Transformer (ViT) tokenization for dysarthria classification. Their system achieved 97.75% accuracy across UASpeech and TORGO, although the task was limited to binary detection rather than complete ASR, raising concerns about overfitting. In follow-up work, Hsieh and Wu [20] investigated curriculum learning combined with articulatory embeddings, achieving an 11.37% improvement over baselines on UASpeech. This confirmed the value of articulatory cues, though their extraction and generalization remained challenges. Revathi et al. [21] focused on small-scale datasets and applied perceptual and intelligibility-enhancing features with decision-level fusion. Their system achieved 81% accuracy in digit recognition, demonstrating that simple fusion techniques can significantly enhance performance on limited tasks; nevertheless, reliance on a minimal dataset limits scalability for continuous ASR.

Hu et al. [22] combined domain-adapted SSL features with inversion features and multi-pass decoding strategies. Their approach improved WERs and character error rates across multiple corpora, including UASpeech, TORGO, and the Dementia Bank Pitt corpus. Importantly, they demonstrated that hybrid systems integrating SSL features outperform standalone SSL models, although integration remains complex and domain mismatches persist. Another line of research by Wang et al. [23] involved fine-tuning SSL representations such as Wav2Vec2 and HuBERT, complemented with graph neural network (GAN)-based and spectral augmentations. This approach achieved a best WER of 16.53%, particularly benefiting subsets with very low intelligibility. The challenge, however, lies in stabilizing GAN training and computational demands. Kumar et al. [24] leveraged residual CNNs with spectral features and speaker-adaptive transfer learning. Their preprocessing included voice cropping and augmentation for hoarse voices, resulting in accuracy improvements of 8.16% for speakers with very low intelligibility. However, the focus on isolated words limited the applicability to continuous speech.

Hsieh and Wu [25] applied a hierarchical curriculum learning strategy combined with knowledge distillation and tailored augmentation. The system achieved average WERs of 19.44% and 6.12% on UASpeech and TORGO, respectively. This reduced cross-group interference effectively, but the requirement for intelligibility labels increased training complexity. Wang et al. [26] explored phonetic boundary refinement using phone-purity-guided tokens derived from HuBERT. Their hybrid TDNN and Conformer system showed consistent WER improvements, with the best reaching 23.25%. However, reliance on phone supervision and codebook size trade-offs complicates scalability. Singh et al. [27] employed Whisper-based models with speaker-independent representations, achieving strong generalization across speakers in both the SAP-1005 and TORGO datasets. While accuracy on Parkinson's speech was high (WER 10.71%), cross-etiology transfer (e.g., Parkinson's to ALS/CP) led to notable performance drops, highlighting a need for personalization. He et al. [28] introduced a comprehensive augmentation pipeline combining synthetic speech (Tacotron2), tempo perturbation, and GAN-based conversion. Fine-tuning Wav2Vec2-XLSR and Whisper-Tiny models yielded substantial reductions in WER, with the best average at 13.58%. Despite the strong performance, mismatches between synthetic and real speech and the computational costs of GANs remain significant hurdles.

Shahamiri et al. [29] applied depthwise-separable convolutions with residual connections to strengthen acoustic modeling. By addressing vanishing gradients and bottlenecks, their system improved word recognition rates by up to 22.58% compared to baselines. Although effective for mild and moderate dysarthria, this deep architecture increased parameter requirements and posed challenges with per-speaker variability. Wang et al. [30] addressed the problem using a Conformer-based sequence-to-sequence model with selective layer freezing during transfer learning. Their two-phase adaptation strategy achieved a WER of 21.5% on UASpeech and 12.7% on TORGO. While the method offers a relatively straightforward adaptation pipeline, the scarcity of dysarthric data and the variability across speakers limit its generalizability. Yue et al. [31] investigated articulatory acoustic features, mutual information analysis, and multimodal fusion. Their end-to-end acoustic–articulatory ASR system demonstrated relative reductions in WER of up to 7.6% for dysarthric and 12.8% for typical speech on the TORGO dataset. The key strength of their approach lies in the complementarity between articulatory and acoustic signals, though collecting synchronized articulatory data across different speakers remains a practical challenge.

The DL algorithms provide better accuracy than ML-based methods due to their superior feature representation capacity, higher hierarchical features representation, increased connectivity, and correlation between local and global features of voice [32, 33]. The following research gaps are identified from the extensive survey of DSR techniques.

- Lower DSR rate due to noise, artifacts, and low intelligibility of the voice.
- Higher computational complexity due to the intricate structure of DL models.
- Lower long-term correlation and temporal depiction of the voice signal due to redundant and non-discriminative features.
- Higher training time and trainable parameters limit the deployment flexibility of the real-time standalone devices.

- The performance of the feature selection techniques is hugely affected by poor solution diversity and poor convergence.
- Lower generalization capability due to vast variations in the phonetic parameters of dysarthric voice.
- Lower DSR accuracy due to the spectral leakage problem, lower feature variance, and the low-frequency resolution problem.

This paper presents the DL-based DSR using DBGNet to improve the classification accuracy. The significant offering of the paper is summarized as follows:

- Development of a hybrid DL framework (DBGNet) combining CNN, bidirectional long short-term memory (BiLSTM), and gated recurrent unit (GRU) to enhance feature correlation, long-term dependencies, and temporal representation for DSR.
- Utilized multiple dysarthric speech features, including time-domain features (TDF), frequency-domain features (FDF), and PF, to capture comprehensive speech characteristics.
- Implemented an Archimedes optimization algorithm (IAOA) with a dynamic chaotic tent map (DCTM) and neighborhood-guided scheme (NGS) for robust feature selection, improving solution diversity, global exploration, and convergence.
- Improved AOA employs the DCTM for population initialization to prevent clustering of initial solutions and premature convergence. By generating a more uniform distribution of individuals across the search space, the CTM enhances the algorithm's global search capability and strengthens its exploration performance.
- The NGS is employed to enhance local exploration, preserve population diversity, and guide updates across multiple subregions, thereby expanding the search space.
- Addressed challenges in DSR, such as poor intelligibility, class imbalance, inter-class and intra-class variability, and complex DL architectures.

The remaining paper is arranged as follows: Section 2 presents the development and operation of the DSR system. Section 3 offers the simulation results for the UASpeech and the TORGO datasets. Section 4 offers the conclusions and future scopes of the work.

2. METHODOLOGY

The framework of the proposed DBGNet-based DSR system consists of speech pre-processing, data augmentation using a GAN, multiple dysarthric feature extraction, feature selection, and classification using the DGL framework, as depicted in Figure 1.

The speech signal is enhanced using wavelet-based soft thresholding, which minimizes noise, artifacts, and irregularities in the voice while accounting for its sub-bands. Furthermore, GAN-based data augmentation is used to reduce class imbalance by generating synthetic samples. Various SDF, TDF, and PF features are extracted from the dysarthric voice to illustrate the influence of dysarthria on the voice. The improved AOA algorithm is used to select more distinctive features, thereby reducing the computational complexity and the number of trainable parameters of the DBGNet.

2.1 DBGNet model

The hybrid DBGNet combines a 1-D deep convolutional neural network (DCNN), BiLSTM, and GRU to provide connectivity features, long-term correlations, and temporal representation of features [34-36]. The DBGNet consists of three parallel arms. The 1st parallel arm includes a 3-layered 1-D DCNN. The 2nd parallel arm comprises two BiLSTM layers, each with 50 units. The 3rd layer encompasses two GRU layers with 50 units each. The DCNN comprises three sequential convolutional layers: 64 filters in the first, 128 in the second, and 256 in the third. Further ReLU and a batch normalization (BN) layer are used to enhance nonlinearity and accelerate training. The outputs of the DCNN arm, the BiLSTM arm, and the GRU arms are flattened and concatenated together. The concatenated deep features are then provided to the fully connected layer (FCL), which has 50 hidden units, to strengthen correlations among hierarchical features at multiple levels. Furthermore, the FCL with two layers, followed by a softmax classifier, is used for DSR.

2.2 Feature selection using improved Archimedes optimization algorithm

According to Archimedes' principle, when an item is placed inside a liquid, the liquid exerts an upward force on the object that is equal to the weight of the object. This force is sometimes referred to as the Buyont force. When this requirement is met, the Archimedes principle is considered valid under the circumstances.

The Archimedes principle is assumed to be satisfied by the characteristics with the maximum fitness value. The velocity, density, and location of the items (feature set) that do not fulfill the principles are modified. This includes the positioning of the objects. Covariance (CoV), the ratio of inter-intra-class variance (RIIN), and entropy (ENT) of the characteristics that characterize the distinguishing aspects of the emotions are used to calculate the object's fitness. An illustration of the AOA-MAUT method is depicted in Figure 2.

The AOA mathematical modelling under equilibrium is described by Eqs. (1)-(5).

$$F_b = W_o \tag{1}$$

$$p_b v_b a_b = p_o v_o a_o \tag{2}$$

$$a_o = \frac{p_b v_b a_b}{p_o v_o} \tag{3}$$

$$W_b - W_r = W_o \tag{4}$$

$$p_b v_b a_b - p_r v_r a_r = p_o v_o a_o \tag{5}$$

Here, p_b and p_o denote the liquid and object densities and objects, v_b and v_o symbolize liquid volume and object volume, and a_b and a_o represent the fluid and object acceleration, respectively [37]. The steps of AOA are mathematically described as follows:

Step 1: Initialization of AOA population using DCTM

The object's population (O) is initialized randomly in Eq. (6), with the probable best feature set. Here, O_i denotes the objects in the population of size N, lb and ub symbolize the lower and upper bounds of the search space. The volumes (vol), density (den), and accelerations (acc) are initialized arbitrarily using Eqs. (6)-(8), respectively. The traditional AOA suffers from initial population clustering, leading to premature convergence and poor global search. The improved

AOA uses the tent chaotic map to initialize the population, thereby avoiding clustering of initial solutions and premature convergence. The DCTM uniformly distributes the population to boost global search ability and exploration rate.

$$O_i = lb_i + C_t \times (ub_i - lb_i); i = 1, 2, \ldots, N \quad (6)$$

$$\begin{aligned} den_i &= C_t \\ vol_i &= C_t \end{aligned} \quad (7)$$

$$acc_i = lb_i + C_t \times (ub_i - lb_i) \quad (8)$$

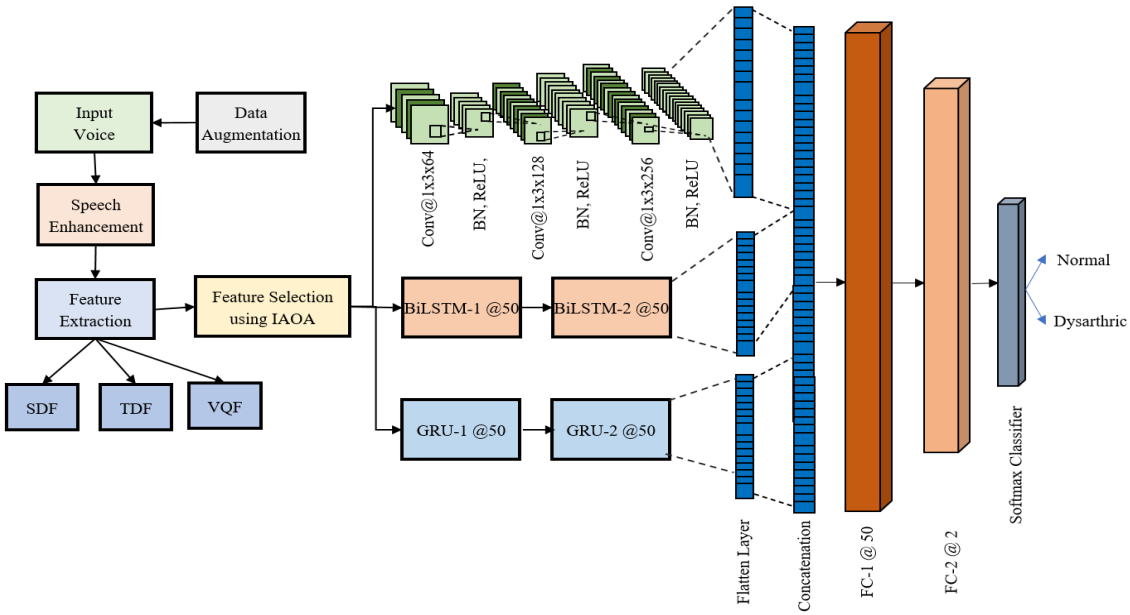

Figure 1. Flow of proposed DBGNet-based dysarthric speech recognition (DSR) system

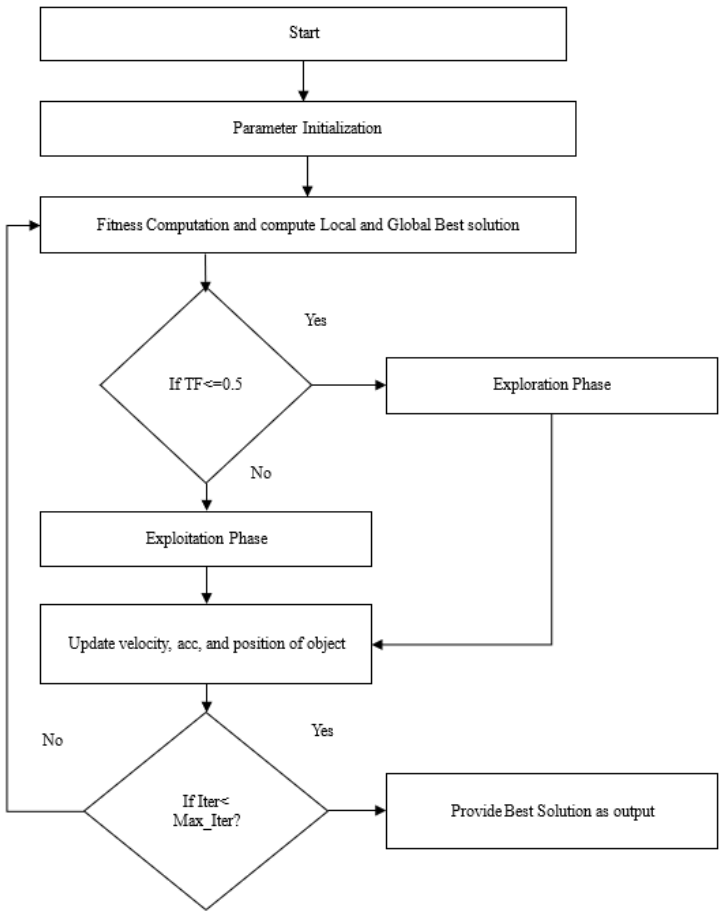

Figure 2. Process of IAOA-based feature selection scheme
Note: IAOA = improved Archimedes optimization algorithm.

The DCTM is computed using Eq. (9), where C_t indicates the DCTM variable and lies between [0,1], and α is a control variation that regulates the distribution and slope of the chaos. The α is dynamically updated based on the progress of the iterations and the population diversity.

$$C_t = \begin{cases} \frac{c_t}{\alpha}, & 0 < C_t < \alpha \\ \frac{1 - c_t}{1 - \alpha}, & \alpha < C_t < 1 \end{cases} \quad (9)$$

The population diversity is computed using Eqs. (10) and (11), which consider the average and current population sizes. The distribution is normalized using Eq. (12) to standardize diversity and avoid dependence on the absolute scale.

$$D^t = \frac{1}{N}\sum_{i=1}^{N} \left\| O_i^t - \overline{O_i^t} \right\| \quad (10)$$

$$\overline{O_i^t} = \frac{1}{N}\sum_{i=1}^{N} O_i^t \quad (11)$$

$$D_{norm}^t = \frac{D^t - D_{min}}{D_{max} - D_{min} + \epsilon} \quad (12)$$

The value of the control variable is updated dynamically based on population diversity, as given in Eq. (13).

$$\alpha^t = \alpha_{min} + (\alpha_{max} - \alpha_{min}).\, D_{norm}^t \quad (13)$$

Higher values of α lead to greater diversity in the population, whereas lower values lead to lower diversity. The value of α_{min} and α_{max} are chosen as 0.3 and 0.7, respectively.

Step 2: Compute the population fitness

The fitness of objects is computed using Eq. (14), which considers *CoV*, *RIIN* and *ENT* to describe correlation, distinctiveness and higher feature depiction.

$$Fitness = w_1 * CoV + w_2 * ENT + w_3 * RIIN \quad (14)$$

Assign den_{best}, vol_{best} and acc_{best} using the best fitness value.

Step 3: Updated density and velocity

Modify the *den* and *vol* of the objects i for t^{th} iteration utilizing Eqs. (15) and (16), respectively. Here, it denotes the arbitrarily distributed random number, $r1$ and $r2$ are random numbers, den_{best} and vol_{best} denotes the density and volume of the best [38].

$$den_i^{t+1} = den_i^t + r1 \times (den_{best} - den_i^t) \quad (15)$$

$$vol_i^{t+1} = vol_i^t + r2 \times (vol_{best} - vol_i^t) \quad (16)$$

Step 4: Update object position using exploration and exploitation using a neighborhood guided scheme

The transfer operator (TF) and the density operator (d) should be computed using Eqs. (17) and (18), respectively [34]. The TF is responsible for determining whether collisions occur when the equilibrium condition is met. This condition depicts the redundancy of characteristics in the population. Using the TF, the exploration (object collision) and exploitation (no-collision) strategies in AOA may be selected with assistance. As the number of iterations grows, the TF increases as well. The efficient use of TF and d may achieve an excellent equilibrium between exploration and exploitation.

$$TF = \exp\left(\frac{ITR - ITR_{max}}{ITR_{max}}\right) \quad (17)$$

Here, ITR_{max} and ITR depicts the maximum number of iterations and the present iteration, respectively.

$$d^{t+1} = exp\left(\frac{ITR_{max} - ITR}{ITR_{max}}\right) - \left(\frac{ITR}{ITR_{max}}\right) \quad (18)$$

When TF ≤ 0.5, the algorithm applies the exploration scheme, taking object collisions into account, and acc is computed using Eq. (19).

$$acc_i^{t+1} = \frac{den_{mr} + vol_{mr} \times acc_{mr}}{den_i^{t+1} \times vol_i^{t+1}} \quad (19)$$

When TF > 0.5, the algorithm applies the exploitation scheme, taking object collisions into account, and acc is computed using Eq. (20).

$$acc_i^{t+1} = \frac{den_{best} + vol_{best \times acc_{best}}}{den_i^{t+1} \times vol_i^{t+1}} \quad (20)$$

The normalized *acc* is computed using Eq. (21), where the upper and lower range of normalization (*uu* and *ll)* are considered 0.9 to 0.1, respectively [34].

$$acc_{i-norm}^{t+1} = uu \times \frac{acc_i^{t+1} - \min(acc)}{\max(acc) - \min(acc)} + ll \quad (21)$$

The object's position is altered utilizing Eq. (22) for collision, where $C_1 = 2$. The value of T increases progressively with time and was depicted by $T = C_3 * TF$.

$$x_i^{t+1} = x_i^t + C_1 \times rand \times acc_{i-norm}^{t+1} \times d \times (O_{rand} - x_i^t) \quad (22)$$

When the population is updated during the non-collision stage, traditional AOA provides the best global guidance for updating the population toward the optimal solution, which may lead to premature convergence and poor solution diversity. To avoid this, the proposed AOA used the neighborhood guidance scheme to improve local exploration, maintain population diversity, and update the population across various subregions to increase the search space. The population is updated using Eq. (23), where $\overline{O_{\aleph}^t}$ denotes average neighborhood guidance, which is computed using Eq. (24). Here, $\aleph_k$ denotes the number of neighbors, which is considered to be 10.

$$O_i^{t+1} = O_{best}^t + F \times C_2 \times rand \times acc_{i-norm}^{t+1} \times d \times (T \times O_{best} - O_i^t) + \hbar \times (\overline{O_{\aleph}^t} - O_i^t) \quad (23)$$

$$\overline{O_{\aleph}^t} = \frac{1}{\aleph_k}\sum_{i \in \aleph_k} O_i^t \quad (24)$$

The object's direction is changed using Eqs. (25) and (26), with +1 for the same direction and -1 for the opposite direction, to generate a more variable population. Here, F denotes the flag, C_4 stands for constant, and PR indicates the random value.

$$F = \begin{cases} +1 \; if \; PR \leq 0.5 \\ -1 \; if \; PR > 0.5 \end{cases} \quad (25)$$

$$PR = 2 * rand - C_4 \quad (26)$$

3. EXPERIMENTAL RESULTS AND SIMULATION

The suggested DSR scheme is implemented in MATLAB R2025a on an NVIDIA GPU with 64 GB of RAM and 512 tensor cores. The training configurations of the DBGNet are provided in Table 1.

Table 1. Parameter specification of DCNN

Parameter	Specification
Batch size	64
Epochs	100
Momentum	0.8
Dropout rate	0.5
Loss function	Cross-entropy
Initial learning rate	0.001
Learning method	Adam

Note: DCNN = deep convolutional neural network.

3.1 Dataset

The experimental results are evaluated on two public datasets to assess the DSR system's generalization capability. The UASpeech dataset consists of 300 words repeated 3 times for computer commands, radio alphabets, digits, and common English words [39]. It consists of 11,700 standard and 14,400 dysarthric samples. The samples are balanced using GAN-based data augmentation, resulting in 15,000 samples per class. The TORGO dataset [40] consists of 1,000 dysarthric and 1,000 standard voice samples. Each class consists of 500 male and 500 female samples. The UASpeech and TORGO datasets are cropped and padded to 1-second duration, and resampled to 1,6000 Hz to maintain uniformity in the data. The results are obtained from 10-fold cross-validation to assess the system's stability.

3.2 Simulation results for UASpeech dataset

Table 2 shows that integrating the IAOA with DBGNet consistently enhances DSR performance on UASpeech compared to the conventional AOA. Across almost all feature set sizes, DBGNet + IAOA achieves higher results. For example, at 350 features, accuracy rises from 98.3% (AOA) to 99.33% (IAOA), with precision, recall, and F1-score all improving from 0.98 to 0.99, indicating a more balanced and robust classification. Similarly, at 450 features, accuracy improves from 96.83% to 97.39%, with precision and recall increasing from 0.96/0.97 to 0.97/0.98, respectively. Even with fewer features, such as 150, the model using IAOA yields 87.54% accuracy compared to 86.83% with AOA, and F1-score improves from 0.87 to 0.88. These gains demonstrate that IAOA provides a better exploration–exploitation balance during feature selection, allowing DBGNet to capture more discriminative cues from dysarthric speech with fewer irrelevant features. The most significant performance boost occurs around mid-range feature sizes (350–400 features), where IAOA achieves nearly 1% higher accuracy and F1-score than AOA, underscoring its effectiveness in identifying the optimal feature subset for DSR tasks. For the overall features (517 features), the DBGNet-IAOA achieves an accuracy of 94.72%, a recall of 0.92, a precision of 0.97, and an F1-score of 0.95.

The visualizations of the various performance metrics for the UASpeech dataset are provided in Figures 3-6, respectively.

Table 2. Comparative analysis of DSR for UASpeech

Number of Features	DBGNet + AOA				DBGNet + IAOA			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
517	94	0.96	0.92	0.94	94.72	0.97	0.92	0.95
500	95.67	0.95	0.95	0.95	96.21	0.96	0.96	0.96
450	96.83	0.96	0.97	0.97	97.39	0.97	0.98	0.98
400	97.1	0.97	0.97	0.97	97.57	0.98	0.98	0.98
350	98.3	0.98	0.98	0.98	99.33	0.99	0.99	0.99
300	94.33	0.96	0.93	0.94	94.98	0.97	0.94	0.95
250	92.33	0.94	0.9	0.92	92.92	0.95	0.91	0.93
200	90.67	0.93	0.88	0.9	91.38	0.94	0.89	0.91
150	86.83	0.88	0.85	0.87	87.54	0.89	0.85	0.88
100	87.33	0.89	0.85	0.87	88.01	0.9	0.86	0.88
50	77.67	0.77	0.78	0.78	78.32	0.78	0.79	0.79
30	79.17	0.77	0.82	0.8	79.13	0.79	0.79	0.79

Note: DSR = dysarthric speech recognition; AOA = Archimedes optimization algorithm; IAOA = improved Archimedes optimization algorithm.

Table 3. Comparative analysis of DSR for the TORGO dataset

Number of Features	DBGNet + AOA				DBGNet + IAOA			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
517	93.06	0.95	0.90	0.93	94.00	0.96	0.92	0.94
500	94.52	0.94	0.95	0.94	95.67	0.95	0.95	0.95
450	96.02	0.95	0.97	0.96	96.83	0.96	0.97	0.97
400	96.06	0.96	0.96	0.96	97.10	0.97	0.97	0.97
350	97.46	0.98	0.98	0.98	98.30	0.98	0.98	0.98

300	93.76	0.95	0.92	0.94	94.33	0.96	0.93	0.94
250	91.47	0.93	0.89	0.91	92.33	0.94	0.90	0.92
200	89.91	0.93	0.88	0.90	90.67	0.93	0.88	0.90
150	86.32	0.87	0.84	0.85	86.83	0.88	0.85	0.87
100	86.33	0.88	0.85	0.87	87.33	0.89	0.85	0.87
50	76.93	0.76	0.78	0.77	77.67	0.77	0.78	0.78
30	77.28	0.77	0.78	0.77	79.17	0.77	0.82	0.80

Note: DSR = dysarthric speech recognition; AOA = Archimedes optimization algorithm; IAOA = improved Archimedes optimization algorithm.

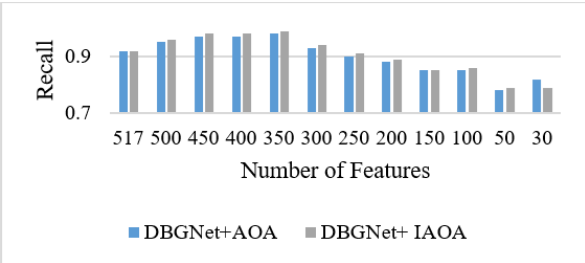

Figure 3. Recall of DSR for different features for the UASpeech dataset
Note: DSR = dysarthric speech recognition.

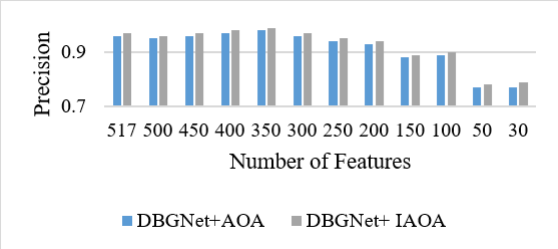

Figure 4. Precision of DSR for different features for the UASpeech dataset
Note: DSR = dysarthric speech recognition.

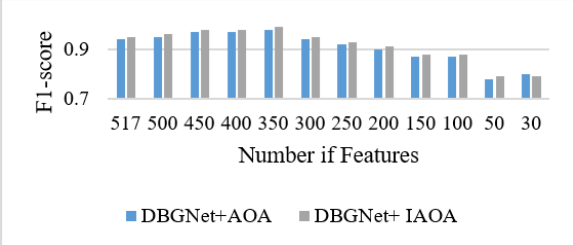

Figure 5. F1-score of DSR for different features for the UASpeech dataset
Note: DSR = dysarthric speech recognition.

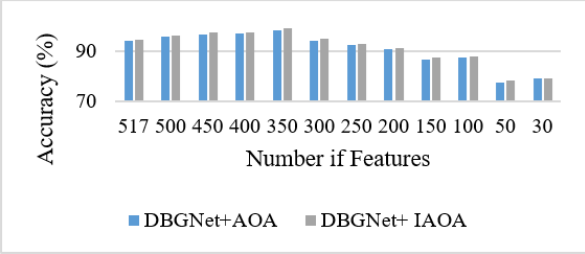

Figure 6. Accuracy of DSR for different features for the UASpeech dataset
Note: DSR = dysarthric speech recognition.

3.3 Simulation results for TORGO dataset

Table 3 presents the DSR results of DBGNet when combined with two metaheuristic feature selection techniques, AOA and IAOA, on the TORGO dysarthric speech dataset. The number of selected features has a significant influence on the classifier's metrics. With DBGNet + AOA, performance improves as the number of features is reduced from 517 to 350, reaching a peak of 98.3% accuracy, 0.98 precision, 0.98 recall, and 0.98 F1-score at 350 features, indicating that AOA effectively removes redundant or irrelevant attributes. Reducing features further (300 down to 30) causes a gradual decline in all metrics, reaching the lowest values of 77.67% accuracy, 0.77 precision, 0.78 recall, and 0.78 F1-score at 50 features. The improved variant (DBGNet + IAOA) consistently outperforms standard AOA at almost every feature set size, achieving its best result of 99.33% accuracy, 0.99 precision, 0.99 recall, and 0.99 F1-score at 350 features, which is approximately 1% higher than DBGNet + AOA. These findings highlight that the hybrid DL model benefits from metaheuristic-driven feature selection and that IAOA identifies more discriminative features than the standard AOA. On the UASpeech dataset, the optimal trade-off between reduced dimensionality and classification accuracy is achieved. The visualizations of the various metrics for DSR for the TORGO dataset are visualized in Figures 7-10, respectively.

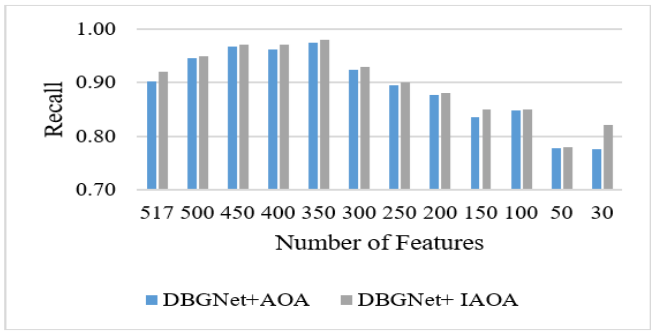

Figure 7. Recall of DSR for different features for the TORGO dataset
Note: DSR = dysarthric speech recognition.

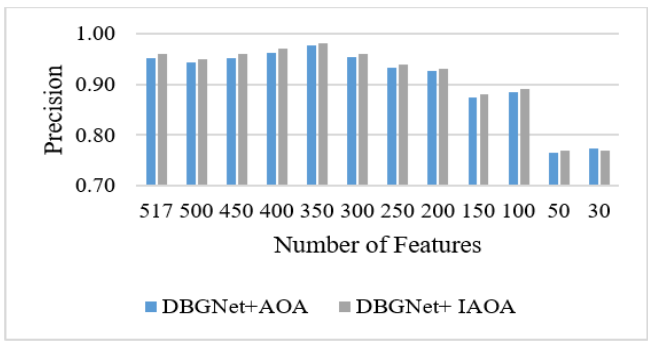

Figure 8. Precision of DSR for different features for the TORGO dataset
Note: DSR = dysarthric speech recognition.

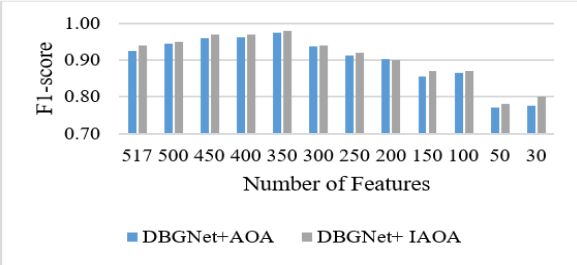

Figure 9. F1-score of DSR for different features for the TORGO dataset
Note: DSR = dysarthric speech recognition.

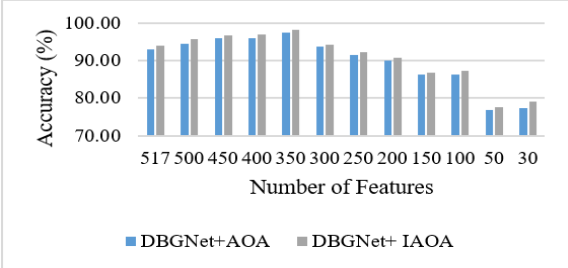

Figure 10. Accuracy of DSR for different features for the TORGO dataset
Note: DSR = dysarthric speech recognition.

3.4 Ablation study

The performance of the suggested system is compared with that of traditional popular transfer learning models, such as VGG16, ResNet18, MobileNetV2, and ViT, as shown in Table 4. Traditional models such as VGG16, ResNet18, and MobileNetV2 achieve accuracies between 92.5–93.5% on UASpeech and 91.8–92.8% on TORGO, with relatively high training times (2376–2632 sec) and moderate testing times (0.42–0.65 sec). ViT improves accuracy to 95.20% (UASpeech) and 94.45% (TORGO) but incurs the highest training cost (2632–2745 sec). DBGNet provides a more efficient alternative, reaching 94% and 94.74% accuracy with reduced training times of 2234 and 2045 sec. The integration of the AOA further enhances performance, boosting accuracy to 98.30% (UASpeech) and 97.46% (TORGO) while significantly lowering training time to 1834–1890 sec. The best results are achieved by IAOA + DBGNet, which delivers state-of-the-art accuracy of 99.33% on UASpeech and 98.30% on TORGO, along with the fastest training (1750–1720 sec) and testing times (0.28–0.31 sec), demonstrating superior efficiency and robustness for DSR applications.

Table 5 compares the trainable parameters of different DL models for DSR, highlighting their computational complexity for DSR. Lightweight architectures such as MobileNetV2 (3.45M) and ResNet18 (11.69M) require significantly fewer parameters than heavy models like ViT (86M) and VGG16 (13.69M). The proposed DBGNet has a moderate complexity of 10.72M parameters. When optimized using AOA and IAOA, the parameter count is further reduced to 7.33M, making the optimized models more compact and computationally efficient while still achieving superior performance.

Table 4. Training time and testing time comparison for various transfer learning models and DBGNet

Method	UASpeech			TORGO		
	Accuracy (%)	Training Time (sec)	Testing Time (sec)	Accuracy (%)	Training Time (sec)	Testing Time (sec)
VGG16	92.50	2457	0.45	91.80	2563	0.61
ResNet18	93.50	2435	0.42	92.35	2341	0.65
MobileNetV2	93.50	2376	0.42	92.80	2532	0.62
ViT	95.20	2632	0.53	94.45	2745	0.56
DBGNet	94	2234	0.44	94.74	2045	0.53
AOA + DBGNet	98.30	1834	0.32	97.46	1890	0.43
IAOA + DBGNet	99.33	1750	0.28	98.30	1720	0.31

Table 5. Trainable parameters for the model (Millions)

Model	Trainable Parameters
VGG16	13.69 M
ResNet18	11.69 M
MobileNetV2	3.45 M
ViT	86 M
DBGNet	10.72 M
AOA + DBGNet	7.33 M
IAOA + DBGNet	7.33 M

Table 6. Comparative analysis of 95% CI for DSR method

Model	UASpeech	TORGO
DBGNet	[93.20%, 94.20%]	[93.95%, 94.97%]
AOA + DBGNet	[97.88%, 98.72%]	[96.95%, 97.97%]
IAOA + DBGNet	[98.50%, 99.70%]	[97.88%, 98.72%]

Note: DSR = dysarthric speech recognition; CI = confidence interval.

Table 6 shows a clear improvement in reliability as optimization is applied to the DSR models for 95% confidence interval (CI). On the UASpeech dataset, DBGNet achieves a confidence interval of [93.20%, 94.20%], which improves substantially with AOA + DBGNet to [97.88%, 98.72%], and reaches its highest stability with IAOA + DBGNet at [98.50%, 99.70%]. A similar trend is observed on the TORGO dataset, where DBGNet achieves [93.95%, 94.97%], AOA + DBGNet improves this to [96.95%, 97.97%], and IAOA+DBGNet further enhances robustness to [97.88%, 98.72%].

3.5 Comparative results analysis

The performance comparison of different DSR methods on the UASpeech and TORGO datasets demonstrates the significant improvement achieved by the proposed IAOA + DBGNet framework, as given in Table 7. Traditional approaches such as Residual CNN [24] achieved accuracies of 75.91% on UASpeech and 74.25% on TORGO, while S-CNN [29] performed lower with 67.00% and 65.80%, respectively. T-GDA [32] attained a high accuracy of 96.30% on UASpeech, showing the potential of statistical feature-based methods, but lacked evaluation on TORGO. DL approaches such as AlexNet [33] further improved recognition

performance, reaching 92.30% on UASpeech and 90.25% on TORGO. The proposed IAOA + DBGNet method outperforms all prior work, achieving 99.33% accuracy on UASpeech and 98.30% on TORGO, representing a substantial 7–32% improvement over conventional CNN-based models. This improvement can be attributed to the hybrid DL architecture that combines CNNs, BiLSTMs, and GRUs for robust feature extraction, along with the IAOA, which effectively selects discriminative features while maintaining population diversity and enhancing global and local search capabilities.

Table 7. Comparative results analysis with traditional techniques

Reference	Algorithm	Accuracy for UASpeech (%)	Accuracy for TORGO (%)
Kumar et al. [24]	Residual CNN	75.91	74.25
Janbakshi et al. [32]	T-GDA	96.30	95.40
Shahamiri et al. [29]	S-CNN	67.00	65.80
Farhadipour et al. [33]	AlexNet	92.30	90.25
Proposed Method	IAOA + DBGNet	99.33	98.30

4. CONCLUSIONS AND FUTURE SCOPES

Thus, this paper provides the DSR based on the ensemble DBGNet using MAFs. The IAOA-based feature selection helps select salient features, minimizing redundancy and enhancing the distinctiveness of the system's features. The MAFs can capture the spectral, prosodic, intonational, and temporal properties of the dysarthric voice. The DGBNet + IAOA achieves 99.33% accuracy, 0.99 recall, 0.99 precision, and a F1-score of 0.99 with 350 acoustic features, demonstrating its superior ability to capture these features. The DBGNet achieves a recall of 0.98, an F1-score of 0.98, a precision of 0.98, and an accuracy of 98.30 on the TORGO dataset. However, the effectiveness of the DBGNet is limited by its lower generalization on real-time datasets and the need for extensive parameter settings. The “black box nature” of the DBGNet leads to poor explainability and the interpretability of the model, which limits the user’s trust in the automatic DSR system.

In the future, the “Interpretability and Explainability” of DL models can be improved to interpret results better and increase system trust by visualizing spectral and temporal changes in the voice. The results of the DBGNet can be enhanced by employing hyperparameter tuning to minimize the tedious task of parameter setting. The focus can be given to dysarthric word correction for a partially recognized word.

REFERENCES

[1] Qian, Z.P., Xiao, K.J., Yu, C.C. (2023). A survey of technologies for automatic dysarthric speech recognition. EURASIP Journal on Audio, Speech, and Music Processing, 2023(1): 48. https://doi.org/10.1186/s13636-023-00318-2

[2] Bhat, C., Strik, H. (2025). Speech technology for automatic recognition and assessment of dysarthric speech: An overview. Journal of Speech, Language, and Hearing Research, 68(2): 547-577. https://doi.org/10.1044/2024_JSLHR-23-00740

[3] Kent, R.D., Weismer, G., Kent, J.F., Vorperian, H.K., Duffy, J.R. (1999). Acoustic studies of dysarthric speech: Methods, progress, and potential. Journal of Communication Disorders, 32(3): 141-186. https://doi.org/10.1016/S0021-9924(99)00004-0

[4] Kain, A.B., Hosom, J.P., Niu, X., Van Santen, J.P., Fried-Oken, M., Staehely, J. (2007). Improving the intelligibility of dysarthric speech. Speech Communication, 49(9): 743-759. https://doi.org/10.1016/j.specom.2007.05.001

[5] Borrie, S.A., McAuliffe, M.J., Liss, J.M. (2012). Perceptual learning of dysarthric speech: A review of experimental studies. Journal of Speech, Language, and Hearing Research, 55(1): 290-305. https://doi.org/10.1044/1092-4388(2011/10-0349)

[6] Alrajhi, T., Ykhlef, M., Alsanad, A. (2024). Recent advances in dysarthric speech recognition: Approaches and datasets. Journal of King Abdulaziz University: Computing and Information Technology Sciences, 13(2): 36-57. https://doi.org/10.4197/Comp.13-2.3

[7] Hamza, A., Addou, D., Hadjadji, I. (2025). Enhancing dysarthric speech intelligibility: A review of techniques. In 2025 3rd International Conference on Electronics, Energy and Measurement (IC2EM), Algiers, Algeria, pp. 1-6. https://doi.org/10.1109/IC2EM63689.2025.11101176

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

[9] Latha, M., Shivakumar, M., Manjula, G., Hemakumar, M., Kumar, M.K. (2023). Deep learning-based acoustic feature representations for dysarthric speech recognition. SN Computer Science, 4(3): 272. https://doi.org/10.1007/s42979-022-01623-x

[10] Liu, S., Geng, M., Hu, S., Xie, X., et al. (2021). Recent progress in the CUHK dysarthric speech recognition system. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29: 2267-2281. https://doi.org/10.1109/TASLP.2021.3091805

[11] Soleymanpour, M., Johnson, M.T., Soleymanpour, R., Berry, J. (2022). Synthesizing dysarthric speech using multi-speaker TTS for dysarthric speech recognition. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, pp. 7382-7386. https://doi.org/10.1109/ICASSP43922.2022.9746585

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

[13] Lin, Y.Q., Wang, L.B., Yang, Y.B., Dang, J.W. (2023). CFDRN: A cognition-inspired feature decomposition and recombination network for dysarthric speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31: 3824-3836. https://doi.org/10.1109/TASLP.2023.3319276

[14] Geng, M.Z., Xie, X.R., Ye, Z., Wang, T.Z., et al. (2022). Speaker adaptation using spectro-temporal deep features for dysarthric and elderly speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30: 2597-2611. https://doi.org/10.1109/TASLP.2022.3195113

[15] Rajeswari, R., Devi, T., Shalini, S. (2022). Dysarthric speech recognition using variational mode decomposition and convolutional neural networks. Wireless Personal Communications, 122(1): 293-307. https://doi.org/10.1007/s11277-021-08899-x

[16] Yue, Z., Loweimi, E., Cvetkovic, Z. (2022). Raw source and filter modelling for dysarthric speech recognition. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, pp. 7377-7381. https://doi.org/10.1109/ICASSP43922.2022.9746553

[17] Yue, Z., Loweimi, E., Christensen, H., Barker, J., Cvetkovic, Z. (2022). Acoustic modelling from raw source and filter components for dysarthric speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30: 2968-2980. https://doi.org/10.1109/TASLP.2022.3205766

[18] Shahamiri, S.R., Lal, V., Shah, D. (2023). Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31: 3407-3416. https://doi.org/10.1109/TNSRE.2023.3307020

[19] Irshad, U., Mahum, R., Ganiyu, I., Butt, F.S., Hidri, L., Ali, T.G., El-Sherbeeny, A.M. (2024). UTran-dysarthric speech recognition: A novel transformer-based model using feature enhancement for dysarthric speech recognition. EURASIP Journal on Audio, Speech, and Music Processing, 2024(1): 1-18. https://doi.org/10.1186/s13636-024-00368-0

[20] Hsieh, I.T., Wu, C.H. (2024). Dysarthric speech recognition using curriculum learning and articulatory feature embedding. In Proceedings of the Interspeech, pp. 1300-1304. https://doi.org/10.21437/interspeech.2024-444

[21] Revathi, A., Sasikaladevi, N., Arunprasanth, D., Amirtharajan, R. (2024). A strategic approach for robust dysarthric speech recognition. Wireless Personal Communications, 134(4): 2315-2346. https://doi.org/10.1007/s11277-024-11029-y

[22] Hu, S., Xie, X., Geng, M., Jin, Z., et al. (2024). Self-supervised ASR models and features for dysarthric and elderly speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32: 3561-3575. https://doi.org/10.1109/TASLP.2024.3422839

[23] Wang, H., Jin, Z., Geng, M., Hu, S., et al. (2024). Enhancing pre-trained ASR system fine-tuning for dysarthric speech recognition using adversarial data augmentation. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, pp. 12311-12315. https://doi.org/10.1109/ICASSP48485.2024.10447702

[24] Kumar, R., Tripathy, M., Anand, R.S., Kumar, N. (2024). Residual convolutional neural network-based dysarthric speech recognition. Arabian Journal for Science and Engineering, 49(12): 16241-16251. https://doi.org/10.1007/s13369-024-08919-5

[25] Hsieh, I.T., Wu, C.H. (2025). Hierarchical curriculum learning for dysarthric speech recognition via multi-level knowledge distillation. IEEE Transactions on Audio, Speech and Language Processing, 33: 3553-3567. https://doi.org/10.1109/TASLPRO.2025.3597438

[26] Wang, H., Xie, X., Geng, M., Hu, S., et al. (2025). Phone-purity guided discrete tokens for dysarthric speech recognition. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, pp. 1-5. https://doi.org/10.1109/ICASSP49660.2025.10889032

[27] Singh, S., Wang, Q., Zhong, Z., Mendes, C., Hasegawa-Johnson, M., Abdulla, W., Shahamiri, S.R. (2025). Robust cross-etiology and speaker-independent dysarthric speech recognition. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, pp. 1-5. https://doi.org/10.1109/ICASSP49660.2025.10888041

[28] He, Y., Seng, K.P., Ang, L.M. (2025). Collaborative AI dysarthric speech recognition system with data augmentation using generative adversarial neural network. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 33: 2097-2111. https://doi.org/10.1109/TNSRE.2025.3570383

[29] Shahamiri, S.R., Mandal, K., Sarkar, S. (2025). Dysarthric speech recognition: An investigation on using depthwise separable convolutions and residual connections. Neural Computing and Applications, 37(12): 7991-8005. https://doi.org/10.1007/s00521-024-10870-3

[30] Wang, Q., Zhong, Z., Singh, S., Mendes, C., Hasegawa-Johnson, M., Abdulla, W., Shahamiri, S.R. (2025). 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, pp. 1-5. https://doi.org/10.1109/ICASSP49660.2025.10889046

[31] Yue, Z., Loweimi, E., Cvetkovic, Z., Barker, J., Christensen, H. (2026). Raw acoustic-articulatory multimodal dysarthric speech recognition. Computer Speech & Language, 95: 101839. https://doi.org/10.1016/j.csl.2025.101839

[32] Janbakshi, P., Kodrasi, I., Bourlard, H. (2021). Subspace-based learning for automatic dysartric speech detection. IEEE Signal Processing Letters, 28: 96-100. https://doi.org/10.1109/LSP.2020.3044503

[33] Farhadipour, A., Chapariniya, M., Vuković, T., Dellwo, V. (2024). Comparative analysis of modality fusion approaches for audio-visual person identification and verification. In Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024), pp. 168-177.

[34] Khan, I.U., Aslam, N., Aboulnour, M., Bashamakh, A., Alghool, F., Alsuwayan, N., Alturaif, R., Gull, H., Iqbal, S.Z., Hussain, T. (2024). Deep learning-based surface defect detection in steel products using convolutional neural networks. Mathematical Modelling of Engineering Problems, 11(11): 3006-3014. https://doi.org/10.18280/mmep.111113

[35] Kakde, A., Dale, M. (2025). Hybrid stock price forecasting with stacked LSTM and multi-source feature fusion. Mathematical Modelling of Engineering Problems, 12(12): 4191-4202. https://doi.org/10.18280/mmep.121208

[36] Wang, J., Xue, M., Culhane, R., Diao, E., Ding, J.,

Tarokh, V. (2020). Speech emotion recognition with dual-sequence LSTM architecture. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pp. 6474-6478. https://doi.org/10.1109/ICASSP40776.2020.9054629

[37] Hashim, F.A., Hussain, K., Houssein, E.H., Mabrouk, M. S., Al-Atabany, W. (2021). Archimedes optimization algorithm: A new metaheuristic algorithm for solving optimization problems. Applied Intelligence, 51(3): 1531-1551. https://doi.org/10.1007/s10489-020-01893-z

[38] Adagale, S.S., Gupta, P. (2025). Parallel deep convolution neural network for speech-based sentiment recognition. Multimedia Tools and Applications, 84(27): 32777-32796. https://doi.org/10.1007/s11042-024-20507-1

[39] Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J.R., Huang, T.S., Watkin, K.L., Frame, S. (2008). Dysarthric speech database for universal access research. In Interspeech, pp. 1741-1744. https://doi.org/10.21437/Interspeech.2008-480

[40] Rudzicz, F., Namasivayam, A.K., Wolff, T. (2012). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Language Resources and Evaluation, 46(4): 523-541. https://doi.org/10.1007/s10579-011-9145-0