Ingénierie des Systèmes d'Information
Vol. 31, No. 3, March, 2026, pp. 865-877
Journal homepage: http://iieta.org/journals/isi

Voice Pathology Detection Using Deep Learning and Feature Selection with Elite Spider Monkey Optimization Based on Periodic Chaotic Tent Map and Lévy Flight Function

Narendra Wagdarikar[1,2*], Sonal Jagtap[3]

[1] Department of E&TC Engineering, G H Raisoni College of Engineering and Management, Wagholi, SPPU, Pune 412207, India
[2] Department of E&TC Engineering, Smt. Kashibai Navale College of Engineering, SSPU, Vadgaon (Bk), Pune 411041, India
[3] Department of E&TC Engineering, NBN Sinhgad Technical Institutes Campus, Ambegaon bk, SPPU, Pune 411041, India

Corresponding Author Email: wagdarikarnarendra@gmail.com

https://doi.org/10.18280/isi.310318

Received: 23 November 2025
Revised: 25 January 2026
Accepted: 19 March 2026
Available online: 31 March 2026

Keywords:
voice pathology detection, deep learning, Long Short-Term Memory, feature selection, elite spider monkey optimization, Saarbrücken Voice Database

ABSTRACT

Voice pathology detection (VPD) is crucial due to the wide range of phonological and prosodic variations, abrupt voice changes, and overlapping symptoms across distinct voice pathologies. This paper presents the VPD using a one-dimensional Deep Convolutional Neural Network and Bidirectional Long Short-Term Memory (1D-DCNN-BiLSTM). The 1D-DCNN-BiLSTM accepts multiple acoustic features (MAF) that combine spectral-domain features (SDF), time-domain features (TDF), and phonatory features (PF) of pathological voice. It uses the improved elite spider monkey optimization (ESMOPL) algorithm with a periodic chaotic tent map (PCTM), a Levy flight function (LFF), and Elite learning (EL) to select essential features, thereby reducing the model's computational complexity. It achieved accuracies of 97%-100% for the two-class VPD and 97.90% for the four-class VPD (normal, cyste, dysphonia, and paralysis) on the Saarbrücken Voice Database (SVD). The feature selection using ESMOPL helps to minimize the computational complexity and provides superior accuracy than existing techniques.

1. INTRODUCTION

Voice pathology refers to disorders or abnormalities that affect the quality of a person's voice, often due to structural, neurological, or functional problems in the vocal folds and related systems. These conditions can range from mild, such as nodules or polyps, to severe, such as vocal fold paralysis and laryngeal cancer. A pathological voice usually sounds different from a healthy one, with changes in pitch, loudness, clarity, or overall tone, making communication more challenging. Because the voice is central to self-expression, social interaction, and professional activities, even minor impairments can impact confidence and overall quality of life [1-3].

Timely detection of voice disorders is essential to prevent complications and ensure effective treatment. Many pathologies, if identified early, can be managed with voice therapy or minor medical interventions, avoiding invasive surgeries later. For individuals whose professions rely heavily on their voice—such as teachers, singers, or broadcasters—early diagnosis helps protect their careers and vocal health. Automated detection systems also play an important role, as they reduce the need for costly and invasive procedures like laryngoscopy, making diagnosis faster, more comfortable, and more widely accessible [4, 5].

Traditionally, the diagnosis of voice pathology has relied on perceptual and clinical methods. Perceptual assessments are conducted by speech-language pathologists who listen carefully and rate qualities such as breathiness, roughness, and strain. Clinical examinations, including laryngoscopy and stroboscopy, provide direct visualization of the vocal folds, while acoustic measures such as jitter, shimmer, and harmonic-to-noise ratio capture voice irregularities. While these methods have been effective, they are often subjective, time-consuming, and sometimes invasive—creating a need for more objective and automated approaches. With advancements in artificial intelligence and signal processing, voice pathology detection has entered a new era. Modern approaches use machine learning and deep learning models to analyze acoustic, aerodynamic, and spectrographic features of voice signals. Techniques such as deep convolutional neural networks (DCNN), recurrent neural networks, and hybrid frameworks are showing impressive accuracy in identifying a wide range of disorders. In addition, mobile health technologies and telemedicine platforms enable remote monitoring of vocal health, making diagnosis easier. The development of large-scale voice pathology datasets and explainable AI models is also improving reliability, transparency, and clinical acceptance of these technologies [6, 7].

Even with rapid progress, several challenges remain in this field. A major hurdle is the lack of large, high-quality datasets, as collecting and labeling pathological voice samples often faces ethical, privacy, and practical barriers. Differences in language, age, gender, and accent further complicate the development of universal models. Detecting subtle signs of

pathology is also challenging, as normal variations in voice can resemble early-stage disorders. Moreover, while deep learning models achieve high performance, their "black-box" nature makes it hard for clinicians to trust their decisions fully. Ensuring interpretability, cross-dataset generalization, and real-time deployment are key areas that researchers continue to address [8-10].

This paper presents a DL-based model comprising a 1D-DCNN-BiLSTM to improve voice pathology detection (VPD) performance. The chief contributions of the system are provided as follows:

- Spectral, phonological, and temporal depiction of pathological voice using MFA that comprises SDF, PF, and TDF using 1D-DCNN and BiLSTM. The 1D-DCNN captures local feature correlations, and the BilSTM layer enhances long-term connectivity and temporal correlations within the features.
- Feature selection using improved ESMOPL to identify the most prominent and distinguishing features, reducing the complexity of the 1D-DCNN. The spider monkey optimization (SMO) uses the PCTM to create an initial, diverse population, Elite learning to increase the search space, and the LFF to increase the SMO's convergence and solution diversity.

The remaining article is summarized as follows: Section 2 reviews recent techniques for VPD. Section 3 provides details on the suggested VPD methodology. Section 4 delivers the augmentative discussions on the two-class and four-class VPD. Lastly, section 5 provides the conclusion and potential future directions of the work.

2. PAGE SETUP

In recent years, various ML and DL-based techniques have been utilized for VPD. Albadr et al. [11] proposed a Fast Learning Network (FLN)-based framework for voice pathology detection, which included normalization and pitch-based segmentation as preprocessing steps. Using MFCC features on the Saarbrücken Voice Database (SVD), the method achieved an accuracy of 84.64%, with precision, recall, and F1-score of 97.39%, 86.05%, and 86.80%, respectively. The key advantage of FLN lies in its hybrid nature, combining neural network and SVM principles to handle both binary and multi-class tasks while reducing overfitting—marking the first application of FLN in this domain. Around the same period, Latiff et al. [12] developed an approach using filtering-based preprocessing and MFCC features, tested on both SVD and the Malaysian Voice Pathology Database (MVPD). Multiple classifiers were examined, and the Online Sequential Extreme Learning Machine (OSELM) outperformed the SVM, Decision Tree, and Naïve Bayes classifiers. OSELM achieved 85.71% accuracy on the same database and 80.77% on a cross-database evaluation, highlighting its robustness and generalizability across datasets. Geng et al. [13] introduced a multimodal transmission network (MMTN) that integrates Mel-spectrograms with Electroglottograph (EGG) signals, combined using short-time Fourier transform mapping. On a dataset of 1,179 subjects from SVD, the model achieved 98.02% accuracy, with strong recall (98.23%) and F1-score (97.95%). This demonstrated the advantage of multimodal integration, though the authors suggested reducing model complexity in future work.

In another contribution, Abdulmajeed et al. [14] applied normalization along with MFCC, ZCR, and Mel-spectrogram features to compare LSTM and ANN. Using the SVD dataset, LSTM achieved superior results, with accuracies of 99.3% for /u/, 99.2% for /a/ and for sentence-level inputs, and 99% for /i/. These findings highlighted the strength of recurrent models over feedforward ANN, supported by a novel feature combination. A clinical study by Cala et al. [15] focused on incorporating vocal fold dynamics and articulatory positioning features within an explainable AI (XAI) framework. Based on a dataset of 287 patients, the system reached accuracies of 76% for females and 81% for males in distinguishing bilateral vocal fold (BLVF) from unilateral vocal fold paralysis (UVFP). However, subclass classification remained limited (60% for BLVF types), likely due to the relatively small dataset. Importantly, this study emphasized interpretability and age-related recovery analysis. Similarly, Gulsen et al. [16] used MFCC features with SVM, achieving 99.19% accuracy for male and 99.50% for female subjects on the SVD dataset through hyperparameter tuning and 10-fold cross-validation. This work confirmed the potential of classical machine learning when combined with appropriate optimization, while also highlighting gender-related differences in performance. Xiong et al. [17] proposed an adversarial feature disentanglement framework incorporating contrastive learning, aiming to suppress noise and disentangle pathological from non-pathological cues. Using both SVD and FEMH datasets, the model achieved accuracies of 87.94% and 92.06%, respectively, with visualizations validating feature separation.

In Arabic voice pathology, Bashir et al. [18] compared SVM, hybrid deep learning, and transfer learning approaches on the AVPD dataset. Transfer learning proved superior, reaching 96.88% accuracy with F1-score of 0.97, outperforming prior works by 1.53%. Brindha et al. [19] designed an ensemble of EfficientNetB0, ResNet50V2, DenseNet121, and CNN, combining MFCC, TQWT, and glottal features. On a clinical dataset, this ensemble achieved 94.02% accuracy, showing robustness across demographics. Importantly, deployment via TensorFlow Lite demonstrated feasibility on low-resource devices. Sasikala et al. [20] explored non-speech signals by analyzing EGG-based features such as spectral contrast, MFCC means, and open/closed quotient. With an SVM classifier, the system achieved 84% accuracy. Although lower than speech-based deep learning models, this approach demonstrated the potential of EGG signals for binary classification.

Another study by Farazi and Shekofteh [21] leveraged phonetic information via phone posterior probabilities (PPPs) alongside MFCCs. A CNN-based model trained on the AVFAD dataset achieved 87% accuracy on the test set and 93% on the validation set, highlighting the benefits of combining acoustic and phonetic cues, especially for spontaneous and read speech. Fu et al. [22] employed data augmentation and phase-space reconstruction (PSR) to construct trajectory graphs, which were then processed by a VGG-like CNN. This model achieved remarkably high accuracy across multiple datasets: 99.42% (MEEI), 97.30% (SVD), and 95.88% (clinical), confirming its strong generalization capacity.

Zhang et al. [23] developed a two-stage hybrid machine learning framework for stroke detection using speech. By enhancing speech with SEWUNet and combining handcrafted and deep features, accuracies exceeded 90% for spontaneous

voice (SV) and 95% for sustained speech (SS). Clinical trials achieved 100% recognition using a WeChat-based app, demonstrating real-time applicability.

El Omari et al. [24] explored pediatric pathology detection using cry signals, employing deep models such as YAMNet and VGG with a parallel 1D CNN. On a children's cry dataset, VGGish achieved 81% accuracy. This was the first attempt to use cry-based cues, opening the door to pediatric diagnosis. Dai et al. [25] presented a Multi-Scale Dynamic Feature Extraction Network (MSDFEN) that uses sinc filters and channel attention. Accuracies varied across datasets, with 98.83% on MEEI, 74.24% on SVD, and 84.09% on HUPA. These results suggested strong adaptability but dataset-dependent performance. In 2024, Farazi and Shekofteh [26] demonstrated the utility of spontaneous speech for pathology detection. Using MFCCs and Mel-spectrograms with CNN, RNN, and hybrid models, they achieved 85% test accuracy and 92% in evaluation. CNN performed best, reinforcing the value of real-life spontaneous data. Özbay et al. [27] introduced a metaheuristic feature selection approach by combining ZCR, RMS, and MFCC features with MELGWO. When integrated with SVM on SVD, this method achieved 99.5% accuracy with excellent F1 and recall, outperforming many prior studies.

Tirronen et al. [28] conducted a systematic analysis of MFCC frame segmentation lengths (20–500 ms) using SVM on SVD. Results showed that a 500 ms frame with 5 ms shift yielded the best accuracy, underlining the importance of temporal resolution in preprocessing. Won and Kim [29] explored few-shot learning (FSL) using transfer-learned embeddings on both voice and EGG signals. Results showed accuracies of 73.7% for voice and 82.6% for EGG, demonstrating the potential of FSL in limited-data scenarios, with EGG yielding better results. Arslan [30] applied mode decomposition techniques (EMD, EEMD, CEEMDAN) along with MFCC, LPC, and LPCC features. Using SVM (cubic kernel), the system reached 99.85% accuracy and F1-score on datasets like Voice ICar Federico II and SVD. Mode decomposition proved highly effective in enhancing cepstral features. Jegan and Jayagowri [31] combined handcrafted and deep features (MFCC, glottal cues, harmonic model, CNN descriptors) with feature selection via Slime Mould Optimizer. On AVPD and SVD, the method achieved 98.46% accuracy, showing the benefit of early fusion and dimensionality reduction. Er and İlhan [32] fused Hilbert-Huang Transform (HHT) with LSTM embeddings through Canonical Correlation Analysis (CCA). Their model combined classical ML classifiers and achieved 89.54% accuracy, with balanced precision, recall, and F1 values around 89%. Finally, Zhao et al. [33] proposed an interpretable CNN using multi-band AT-SincNet filters. Tested on MEEI, SVD, and HUPA, the system outperformed baselines with improvements of 0.17 in accuracy and 0.19 in F1. Blind testing showed 75.9% accuracy and 84.9% F1, proving its strength in generalization and interpretability.

Thus, from the extensive survey, we have identified the major research gaps as given below:

- Higher complexity of recent DL techniques leads to huge training time and trainable parameters.
- Higher feature dimensions and redundancy in features decrease the VPD system's accuracy.
- Existing feature selection lacks the ability to provide distinctive features because of poor convergence, inferior solution variability, lower search space, and poor exploration-exploitation balance.
- High overlapping similarity in pathological voice features leads to poor feature distinctiveness.

3. METHODOLOGY

The speech signal is first processed to obtain two complementary feature sets, as shown in Figure 1. First, it extracts multiple acoustic features directly from the raw speech signal, including SDF, TDF, and PF. These features are selected using an ESMOPL, which reduces dimensionality and retains only the most discriminative attributes. The selected features are then fed to a 1D-DCNN, which learns temporal patterns and variations across the feature vectors. After processing through both neural branches, the outputs are concatenated to form a unified feature representation combining learned spectral and temporal characteristics. Further, a Bi-LSTM layer is used to captures long-range temporal dependencies in the speech signal—particularly useful for modeling dynamically evolving pathological voice patterns. The sequence-aware representation from the Bi-LSTM is then fed through a fully connected layer and finally to a Softmax classifier, which categorizes the input speech as normal or pathological. Overall, the hybrid model leverages the strengths of both convolutional networks (for spatial and temporal feature learning) and recurrent networks (for sequence modeling). By combining spectrogram-based deep features with optimized acoustic features, the system provides a robust framework for early and accurate detection of voice disorders.

3.1 Proposed hybrid model

The proposed voice pathology detection framework comprises two parallel deep-learning branches that capture complementary information from the speech signal. The system processes 547 acoustic features using a 1D-DCNN architecture to learn localized correlation in SDF, TDF, and PF. The 1D-CNN, this branch also consists of three convolutional layers with 64, 128, and 256 filters at successive depths, each followed by ReLU activation and Max-Pooling to learn temporal variations within the feature vectors. These acoustic features represent key voice characteristics such as pitch variations, perturbation measures, energy distribution, and noise levels, and the 1D-CNN effectively captures their hidden temporal dependencies.

The outputs from both CNN branches are concatenated to form a unified high-level feature representation. This combined feature vector is passed through a Bi-LSTM network composed of two layers, each containing 50 hidden units. The bidirectional structure enables the model to learn forward and backward temporal dependencies, capturing dynamic voice patterns more accurately. After this sequence modeling stage, the framework uses a fully connected (FC) network with 10 layers that systematically refines the extracted deep features. Finally, the FC network's output is fed into a Softmax classifier for multiclass or binary classification. Together, these components form a robust and comprehensive framework capable of accurately detecting pathological voice conditions.

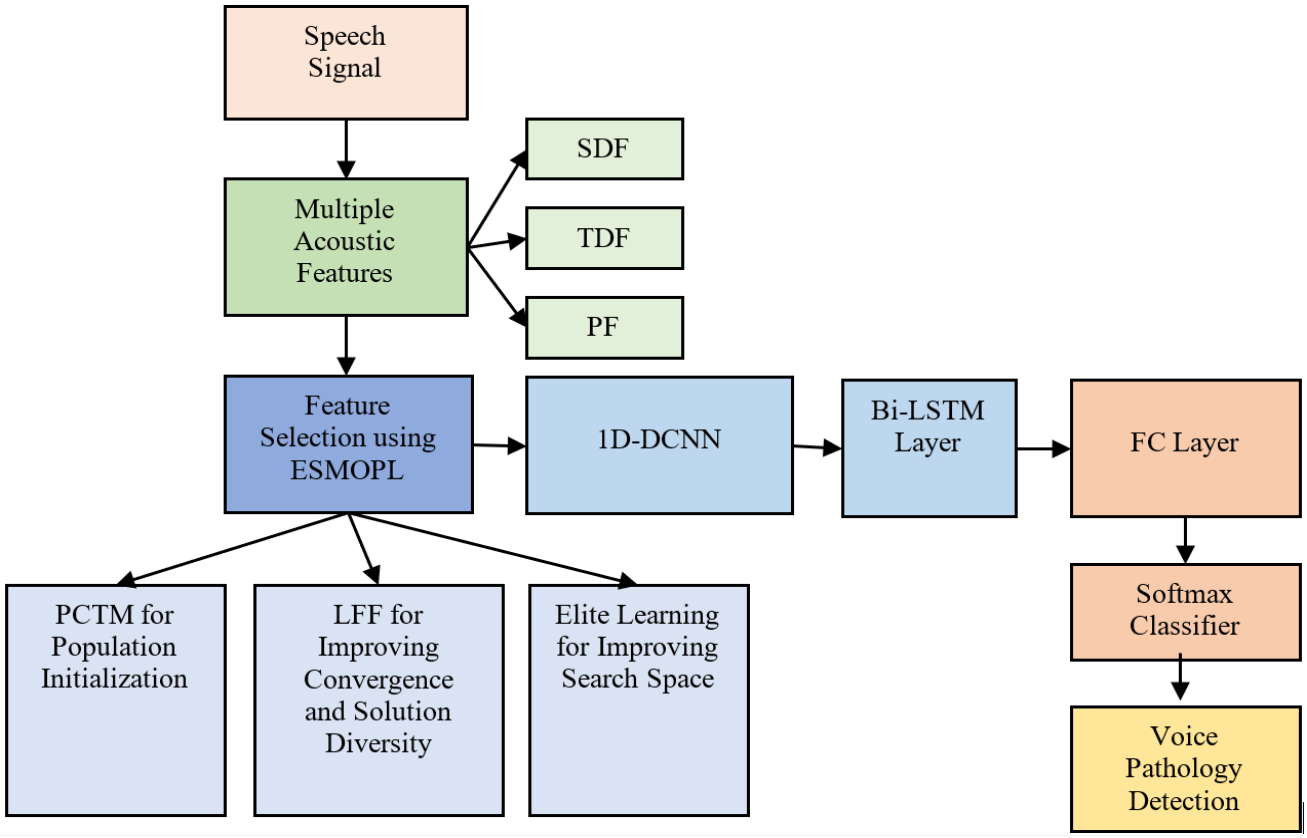

Figure 1. Flow diagram of the proposed voice pathology detection (VPD) system

Table 1. Details of multiple voice features for voice pathology detection characterization

Category	Feature	Long Form	No. of Features	Significance of Feature
SDF	MFCC	Mel-Frequency Cepstral Coefficients	13	Captures vocal tract characteristics to distinguish healthy vs. pathological voices.
	MFCC Δ	Delta Mel-Frequency Cepstral Coefficients	13	Represents how spectral features change over time to detect instability in speech.
	MFCC ΔΔ	Delta-Delta Mel-Frequency Cepstral Coefficients	13	Highlights dynamic acceleration patterns helpful in identifying fluctuating vocal disorders.
	LPCC	Linear Predictive Cepstral Coefficients	13	Models speech production behavior to detect abnormalities in vocal fold patterns.
	Spectral Kurtosis	Spectral Kurtosis Measure	257	Measures sharpness of the spectrum to detect irregular harmonics caused by vocal damage.
	Formants	Formant Frequencies (F1–F3)	3	Represents resonance frequencies that shift when vocal tract structure is affected.
	Formant Mean	Average Formant Frequency	1	Indicates consistent vocal tract behavior across time for disorder screening.
	Formant Variance	Variability in Formant Frequency	1	Tracks instability in articulation linked to pathological speech.
	Skewness	Spectral Skewness	1	Measures asymmetry in spectral distribution signaling abnormal energy patterns.
	Wavelet Features	Multi-Level Wavelet Coefficients	224	Captures multi-resolution changes in speech useful for detecting noisy pathological variations.
TDF	Pitch Frequency	Fundamental Frequency (F0)	1	Identifies pitch disruptions seen in diseases affecting vocal fold vibration.
	ZCR	Zero-Crossing Rate	1	Detects abrupt signal transitions linked to harshness or breathiness.
	Hjorth Parameters	Activity, Mobility & Complexity	3	Quantifies temporal dynamics reflecting instability in vocal signals.
PF	Jitter	Frequency Perturbation	1	Measures pitch variation, often elevated in vocal fold disorders.
	Shimmer	Amplitude Perturbation	1	Tracks loudness fluctuations associated with weak or damaged vocal cords.
	RMS	Root Mean Square Energy	1	Represents overall speech energy, reduced in cases of vocal weakness.
—	**Total Features**	—		**547**

3.2 Multiple acoustic features

The MAF details include the SDF, PF, and TDF for the pathological voice attributes, as shown in Table 1.

3.3 Feature selection using ISMO

The SMO algorithm is employed to select the most relevant MAF features and inspired by the social behavior of spider monkeys (SM), SMO models how these animals form large groups of 40–50 members, led by a dominant female, the global leader, who directs food-search activities. When food becomes scarce, the main group splits into smaller subgroups of 3–8 members, each guided by a local leader responsible for navigation and decision-making [1]. The algorithm mimics this behavior in four stages: initially locating potential food (solutions), sharing positional updates within subgroups, selecting the best subgroup-level solution, and finally choosing the best global solution using feedback from local leaders. If progress stalls, groups are further divided to explore new regions of the search space [34, 35]. The proposed ESMOPL enhances this strategy by incorporating PCTM, EL, and Levy flights to diversify exploration and reduce the risk of getting stuck in local minima, as shown in Figure 2.

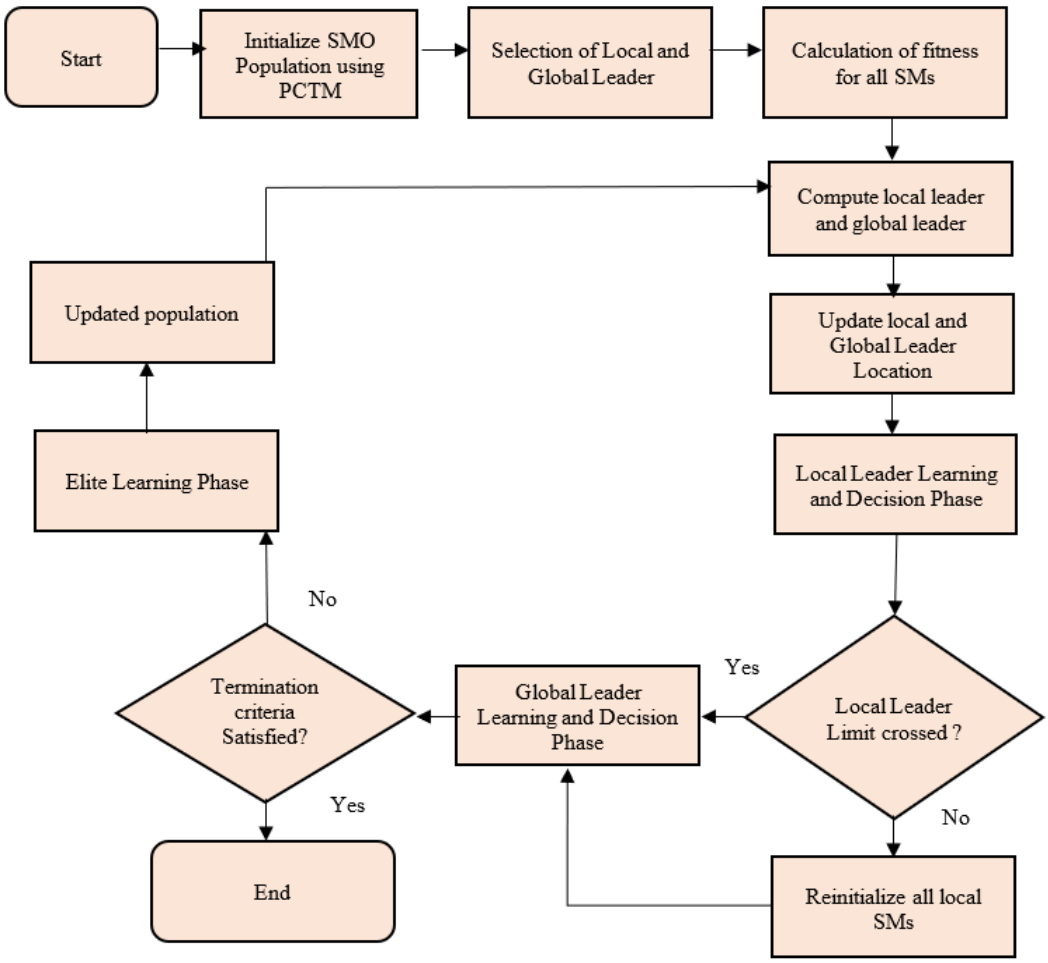

Figure 2. Flow chart of proposed ESMOPL for feature selection

Step 1: Initialization SM population

The first step involves initializing the population of SMs, which represents the possible sets of MAFs. The N sets of n features denotes the SM population which is initialized as given in Eq. (1) where sm denotes an individual monkey depicting a single feature.

$$SM = \begin{bmatrix} sm_{11} & sm_{12} & \cdots & C_{1n} \\ sm_{21} & sm_{22} & \vdots & C_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ sm_{N1} & sm_{N2} & \cdots & C_{Nn} \end{bmatrix} \quad (1)$$

$$sm = LBF + TM * (UBF - LBF) \quad (2)$$

The population is initialized using the periodic chaotic Tent Map (PCTM) to minimize non-uniformity and increase diversity. The TM is given in Eq. (3).

$$TM_{i+1} = \begin{cases} r * TM & 0 < TM < 0.5 \\ r * (1 - TM) & 0.5 \leq TM \leq 1 \end{cases} \quad (3)$$

$$r = 2 - 0.5 * sin\left(\frac{2\pi k}{K}\right) \quad (4)$$

Step 2: Compute the fitness of the population

The fitness of the feature selection depends upon the intraclass variability ($\sigma_{intra-class}$) and inter-class variability ($\sigma_{inter-class}$), entropy of features (EN) and covariance of the features (CV) as given in Eq. (5).

$$Fitness = \frac{\sigma_{intraclass}}{\sigma_{interclass}} + \frac{1}{EN} + \frac{1}{CV} \quad (5)$$

Step 3: Local leader Phase with LFF

The local group leader refines the locations of SMs based

on insights gained from both the regional leader and fellow group members. If a newly generated solution has lower fitness than the current local leader, the SM adjusts its location again according to Eq. (6).

$$sm_{newj} = sm_{ij} + R_n \times (LL_{kj} - sm_{ij}) + R_n \times (sm_{rj} - sm_{ij}) + levy \quad (6)$$

where, sm_{newj} denotes the modified location of SMs, sm_{ij} depict the previous location of SM, LL_{kj} symbolizes the j^{th} location of k^{th} local group, and sm_{rj} provides j^{th} location of r^{th} SM, whose location in the group is randomly modified and R_n depicts the random number distributed uniformly between 0 and 1.

The Lévy flight mechanism is employed to gradually perturb and expand the SMO population, thereby increasing solution diversity and improving convergence. The Lévy-based position update is defined by Eqs. (7) and (8), where β represents the Lévy exponent controlling the distribution shape, Γ denotes the gamma function, and θ acts as the step-size control parameter.

$$Levy = 0.01.\frac{rand1.\theta}{|rand2|^{\frac{1}{\beta}}} \quad (7)$$

$$\theta = \frac{\Gamma(1+\beta).sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right).\beta.2^{\left(\frac{\beta-1}{2}\right)}} \quad (8)$$

Step 4: Global Leader Phase with LFF

The global leader updates the positions of both individual SMs and their subgroups using the knowledge derived from international and local leaders, guided by probability-based fitness, as defined in Eqs. (9) and (10).

$$sm_{newj} = sm_{ij} + R_n \times (GL_{kj} - sm_{ij}) + R_n \times (sm_{rj} - sm_{ij}) + levy \quad (9)$$

where, GL_{kj} represents the jth dimension of $k\,th$ global leader location.

$$Prob_i = \frac{fitness_i}{\sum_{i=0}^{N} fitness_i} \quad (10)$$

Step 5: Global and Local Leader Learning Phase

The global leader adjusts the positions of the SMs by selecting the best-performing solution across all groups using a greedy selection strategy. Similarly, the local leader updates the SM positions within each subgroup based on their respective fitness values, also following a greedy selection approach. Ultimately, the SM achieving the highest fitness within a group is designated as the local leader.

Step 6: Local Leader Decision Phase

When stagnation occurs, all SMs within the local groups randomly adjust their positions based on decisions influenced by both the local and global leaders as defined by Eq. (11).

$$sm_{newj} = sm_{ij} + R_n \times (GL_j - sm_{ij}) + R_n \times (sm_{ij} - LL_{kj}) \quad (11)$$

If the global leader's position remains unchanged for several iterations, the population is split into smaller subgroups (fission). When these subgroups become large and can no longer influence the global leader, they are merged back into a single group (fusion). Thus, fission enhances exploration during stagnation, while fusion refocuses the search when regrouping is needed.

Step 7: Elite Learning Phase

The process of elite learning is shown in Figure 3, where two elite members with the highest fitness are used to share knowledge, thereby improving the search space and the effective use of the existing population. The algorithm selects two top-performing solutions from the elite 5% of the population, denoted as $E1$and $E2$. If the two selected solutions are not identical, a knowledge-sharing strategy is applied to generate improved versions, producing $E1_{new}$ and $E2_{new}$. Their fitness values are then evaluated, and the solution with the highest fitness is chosen as the global best—provided it also outperforms both original elite solutions. Otherwise, the better of the two new solutions is retained. The final selected solution is returned as the optimal feature set.

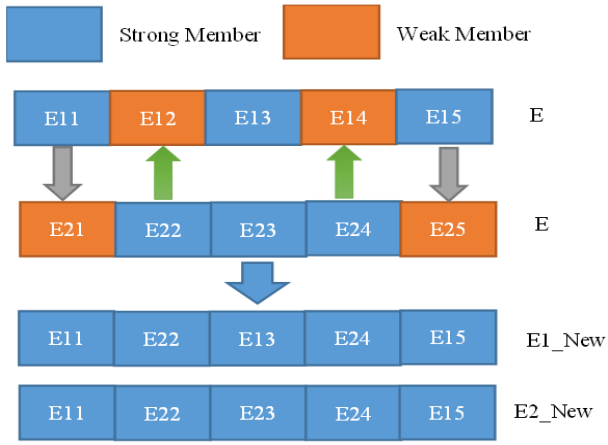

Figure 3. Elite learning phase

The algorithm is provided as given below:

Algorithm: Elite learning for Feature selection using ESMOPL
Input: Two elite solutions E1 and E2
Output: Global Best Solution

1. Select two elite members having best fitness
   E1: Random monkey from 5% elite group
   E2: Random monkey from 5% elite group
2. If E1~=E2
   Apply Knowledge Sharing Strategy
   Obtain New Solution as E1_new and E2_new
   Else
   Repeat step 1;
   end
3. Compute Fitness of E1_new and E2_new
4. If fitness (E1_new) > fitness (E2_new) and fitness (E1_new) > max (fitness (E1, E2))
   Global_Best=E1_new
   Else
   Global_Best=E2_new
   end
5. Provide the final global solution as Best Feature set

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

The suggested system is implemented using the SVD dataset [36], which includes three pathology classes (cyste, paralysis, dysphonia) and one regular class. A total of 200 samples of each class are considered for experimental evaluation. The hyperparameters of the proposed DL framework are provided in Table 2.

The performance evaluation of the proposed voice pathology detection framework demonstrates strong diagnostic capability across multiple classification scenarios. The model consistently achieves high recall, precision, F1-score, selectivity, NPV, and accuracy, reflecting its robustness in detecting various voice disorders, such as cysts, dysphonia, and paralysis, and in distinguishing pathological from normal voices. The confusion matrices for different results are given in Figure 4.

Table 2. Hyperparameters of the DL model used for VPD

Parameter	Specification
Train Test Split	70:30
Learning rate	0.001
Optimization method	Adam
Epoch	200
Batch size	32
Loss Fucntion	Cross entropy
Dropout	0.5
Momentum	0.8

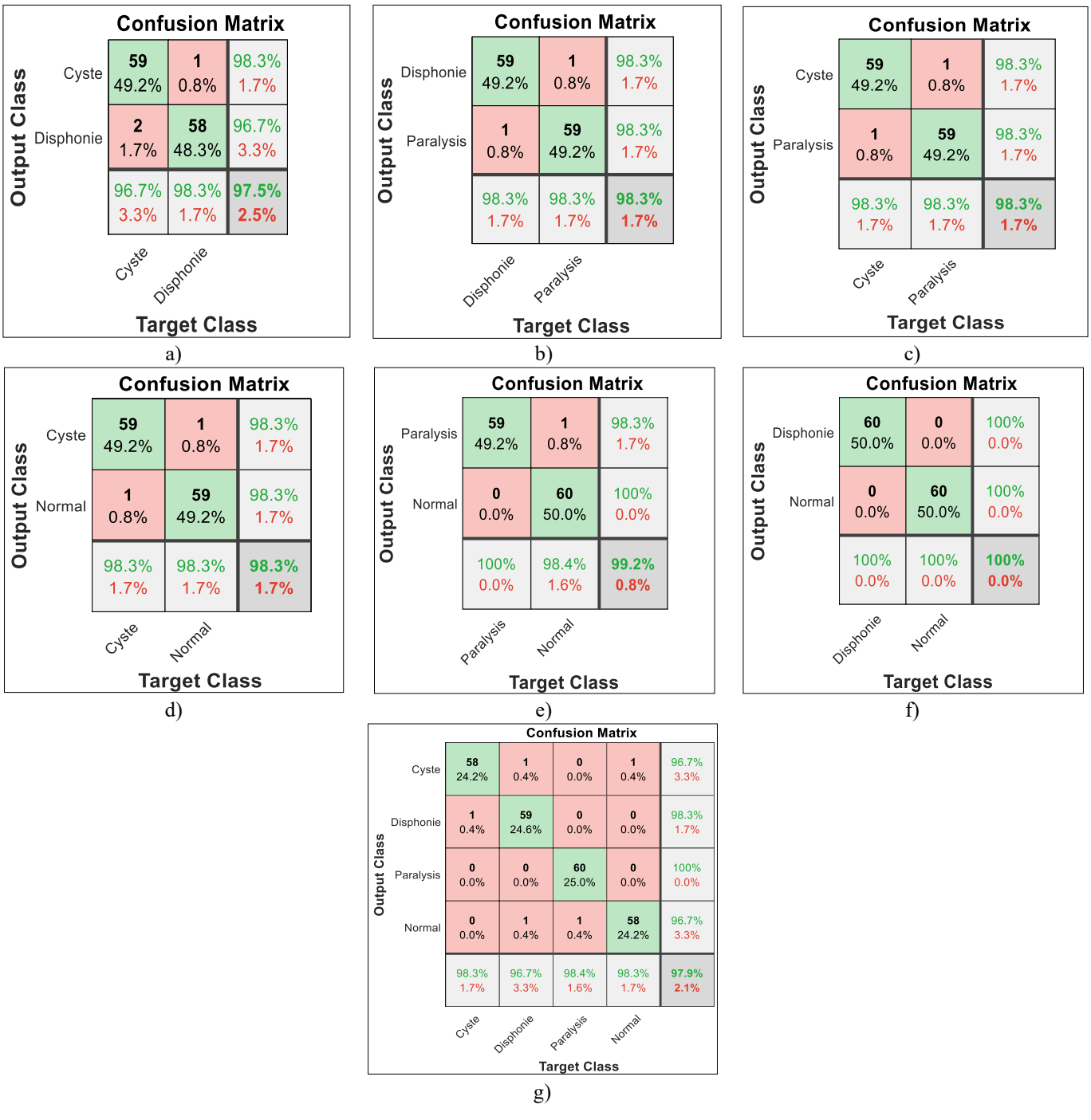

Figure 4. Confusion matrix for DL-based VPD a-f) 2-class VPD g) 4-class VPD for 300 features

In the binary classifications, the system shows exceptional reliability as given in Table 3. For Cyst vs Dysphonia, the model achieves a recall of 98.30% and a precision of 96.70%, leading to an F1-score of 97.49% and an overall accuracy of 97.50%. These high values indicate that the system effectively identifies both classes with minimal false positives and false negatives. Similarly, for Paralysis vs Dysphonia and Cyst vs Paralysis, all performance metrics reach 98.30%, indicating balanced detection and stable classification across disorders. When comparing Cyst vs Normal and Dysphonia vs Normal, the framework achieves near-perfect performance. In particular, Dysphonia vs Normal achieves 100% in all evaluation metrics, demonstrating the model's ability to differentiate pathological features from normal vocal patterns clearly.

The classification between Paralysis and Normal also shows strong performance, with accuracy of 99.20% and an F1-score of 99.14%. The precision value is 100%, meaning the model produces no false positives when identifying paralysis cases, while the recall of 98.30% confirms a very low rate of missed pathological cases. For the broader classification task of grouping normal voices against all three pathologies—Normal vs (Cyst, Dysphonia, and Paralysis)—the framework achieves 98.50% accuracy and an F1-score of 98.73%, indicating consistent generalization across multiple pathological categories. The multiclass classification involving all four categories—Normal vs Cyst vs Dysphonia vs Paralysis—achieves an accuracy of 97.90%, with balanced precision (97.92%) and recall (97.85%). This indicates that even when confronted with a more complex classification setup, the system retains outstanding discriminatory power. Overall, these statistical results confirm the reliability, stability, and effectiveness of the proposed model for robust voice pathology detection across multiple diagnostic scenarios. The results are visualized in Figures 5-10, respectively.

The accuracy comparison of VPD across different feature sizes shows a consistent upward trend as more features are selected using ESMOPL, as given in Table 4. For 2-class classification, accuracies rise significantly from around 86–90% at 50 features to peak values at 300 features, such as 97.5% for Cyste vs Dysphonia, 98.3% for Paralysis vs Dysphonia, and 100% for Dysphonia vs Normal. Beyond 300 features, performance begins to plateau or slightly decline, with values ranging between 96–98% up to 547 features. For the combined 4-class classification, accuracy improves from 86.7% at 50 features to 97.9% at 300 features, then marginally decreases to 96.0% at 547 features. Overall, the results indicate that selecting approximately 250–300 features yields optimal performance across both binary and multiclass VPD tasks, demonstrating the effectiveness of feature optimization in enhancing diagnostic accuracy while avoiding redundancy from excessive features.

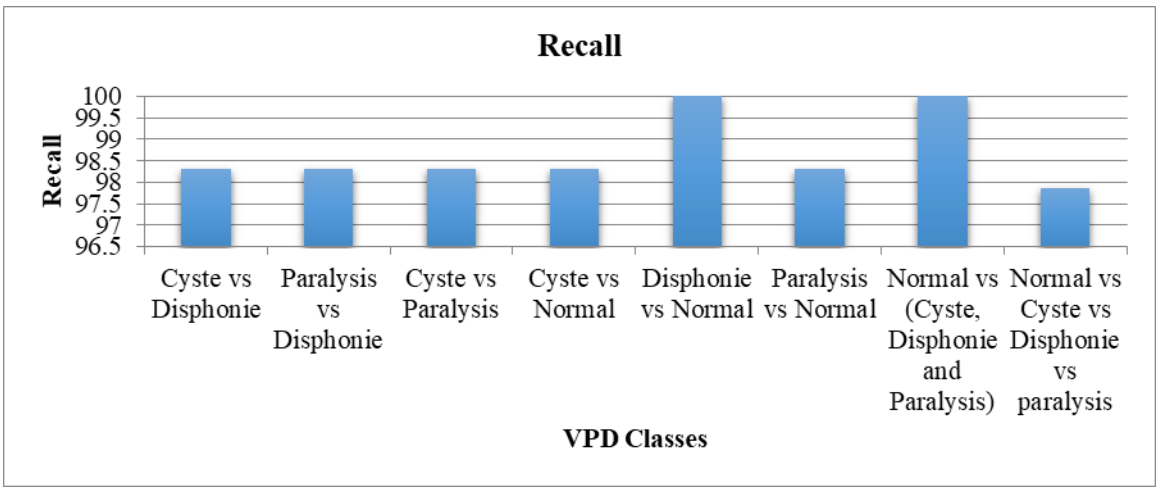

Figure 5. Recall comparison of the proposed DL model for different VPD classes

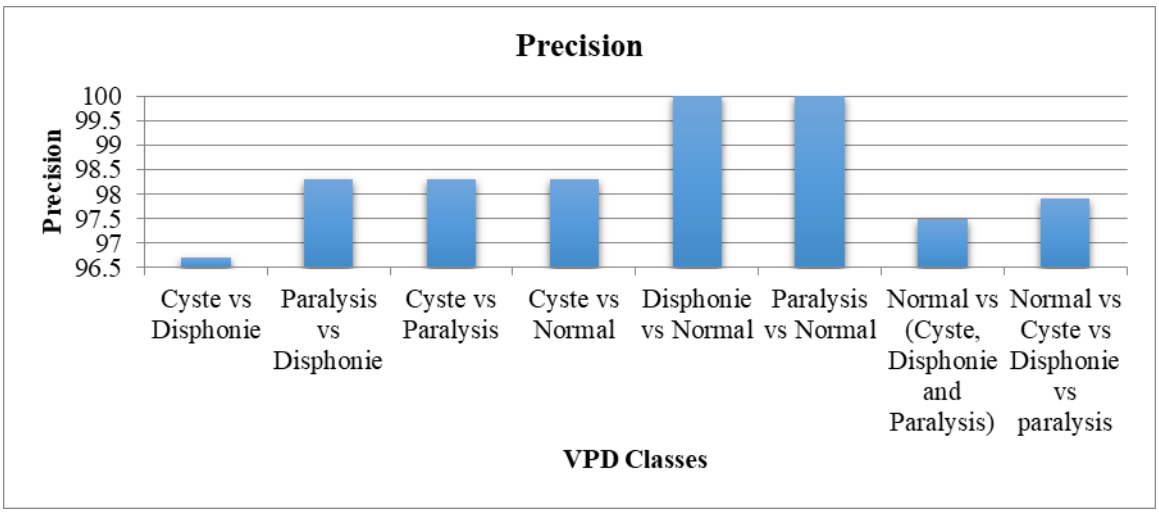

Figure 6. Precision comparison of the proposed DL model for different VPD classes

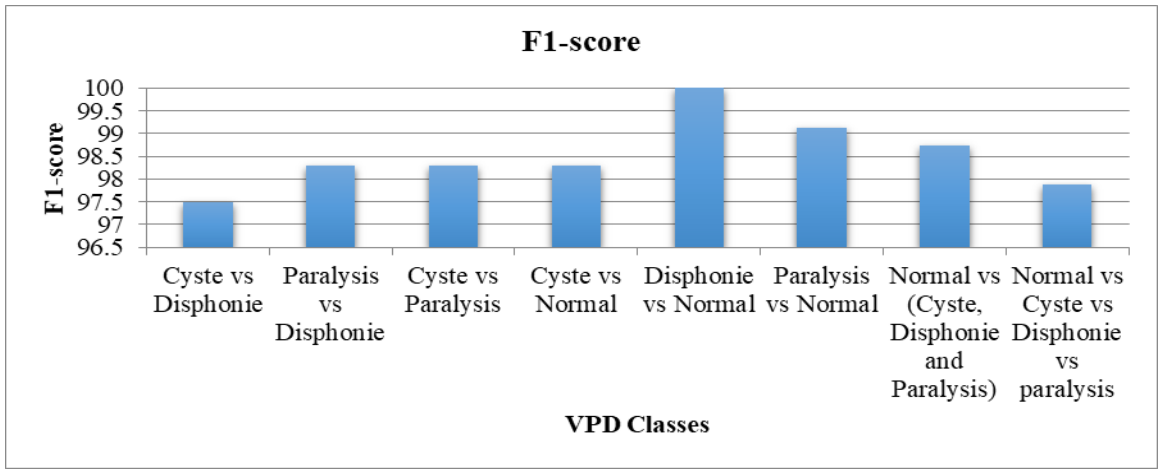

Figure 7. F1-score comparison of proposed model for different VPD classes

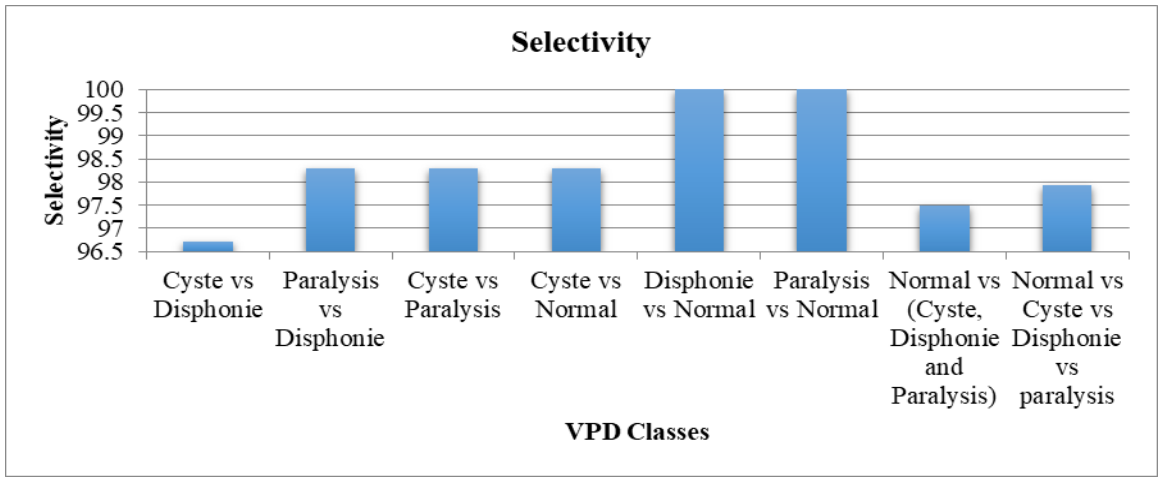

Figure 8. Selectivity comparison of proposed model for different VPD classes

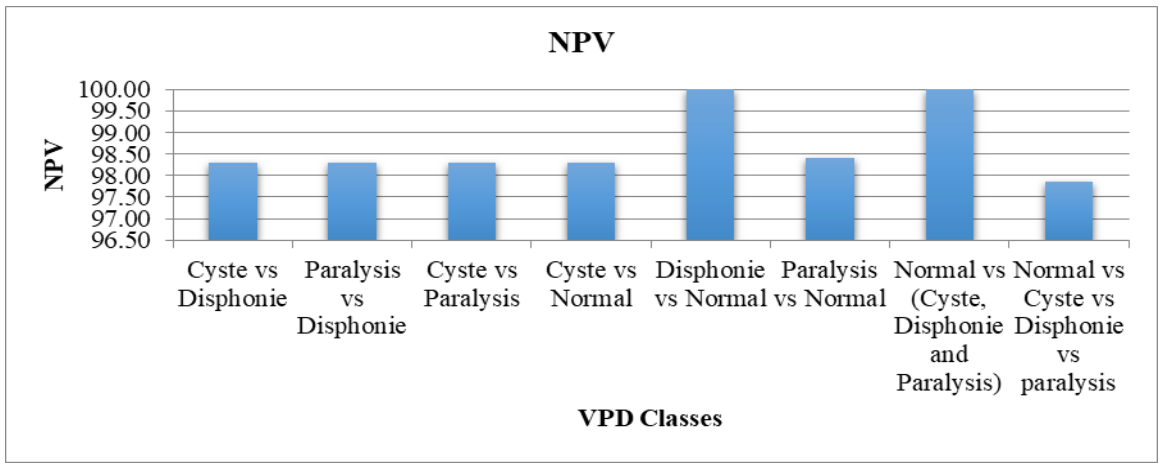

Figure 9. NPV comparison of proposed model for different VPD classes

Table 3. Results comparison for 2-class and 4-class VPD for 300 features

Classes	Recall	Precision	F1-Score	Selectivity	NPV	Accuracy
Cyste vs Disphonie	98.30	96.70	97.49	96.70	98.30	97.50
Paralysis vs Disphonie	98.30	98.30	98.30	98.30	98.30	98.30
Cyste vs Paralysis	98.30	98.30	98.30	98.30	98.30	98.30
Cyste vs Normal	98.30	98.30	98.30	98.30	98.30	98.30
Disphonie vs Normal	100.00	100.00	100.00	100.00	100.00	100.00
Paralysis vs Normal	98.30	100.00	99.14	100.00	98.40	99.20
Normal vs (Cyste, Disphonie and Paralysis)	100.00	97.50	98.73	97.50	100.00	98.50
Normal vs Cyste vs Disphonie vs paralysis	97.85	97.92	97.88	97.92	97.85	97.90

Table 4. Accuracy comparison for 2-class and 4-class VPD for different features

Classes	Number of Features Selected Using ESMOPL										
	50	100	150	200	250	300	350	400	450	500	547
Cyste vs Disphonie	86.9	89.0	91.0	92.7	95.2	97.5	97.1	96.6	96.2	95.9	95.5
Paralysis vs Disphonie	87.5	90.2	91.9	94.0	96.5	98.3	98.0	97.5	97.2	96.8	96.3
Cyste vs Paralysis	86.4	88.9	91.3	93.8	96.3	98.3	97.9	97.5	97.2	96.8	96.4
Cyste vs Normal	86.9	88.9	91.2	93.6	95.7	98.3	97.9	97.5	97.0	96.7	96.3
Disphonie vs Normal	89.9	91.5	93.5	95.6	98.0	100.0	99.6	99.2	98.8	98.3	97.8
Paralysis vs Normal	86.4	88.9	91.5	93.9	96.6	99.2	98.8	98.3	97.9	97.4	97.1
Normal vs (Cyste, Disphonie and Paralysis)	87.3	89.9	92.1	94.0	96.2	98.5	98.1	97.8	97.3	96.9	96.5
Normal vs Cyste vs Disphonie vs paralysis	86.7	88.5	90.7	93.3	95.4	97.9	97.6	97.1	96.7	96.4	96.0

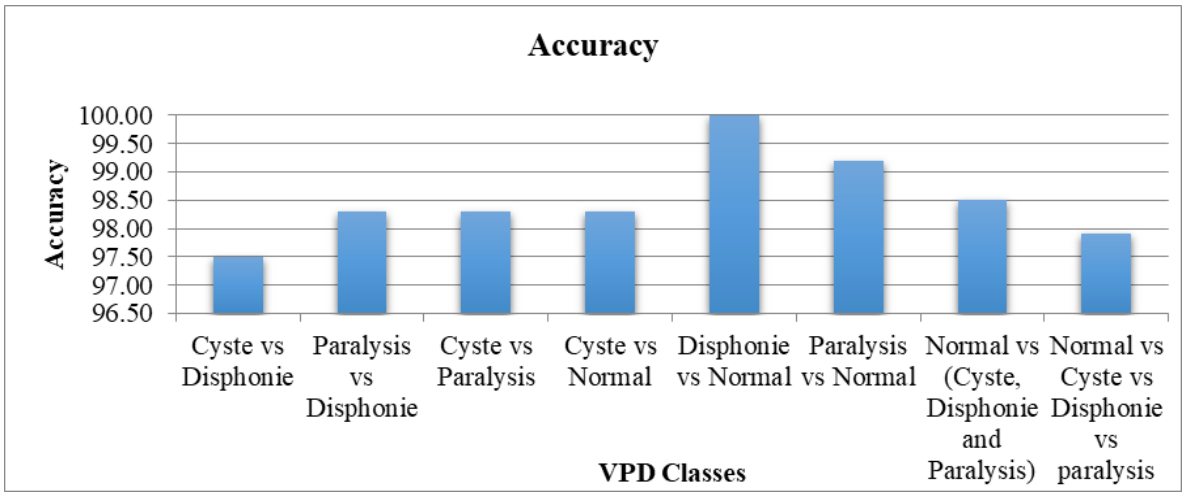

Figure 10. Accuracy comparison of the proposed model for different VPD classes

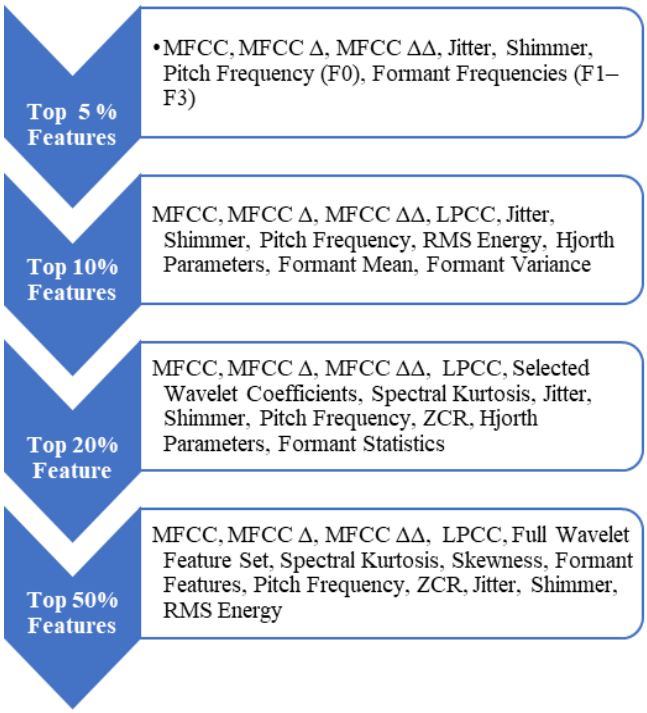

Figure 11. Importance ranking of features selected using ESMOPL

The feature ranking results shown in Figure 11 show that MFCC and its dynamic variants (Δ and ΔΔ) consistently appear in the top 5–10%, highlighting their strong ability to capture vocal tract and spectral variations caused by pathology. Perturbation features such as jitter and shimmer, along with pitch frequency and formant features, are also highly ranked due to their sensitivity to vocal fold irregularities. As the feature subset expands to the top 20–50%, wavelet coefficients and spectral kurtosis become prominent, providing multi-resolution and harmonic irregularity information.

Table 5 compares the performance of different feature selection algorithms such as particle swarm optimization (PSO), Gray Wolf Optimization (GWO), and SMO across three model configurations—DCNN, BiLSTM, and the hybrid DCNN-BiLSTM—for four-class voice pathology detection. For the standalone DCNN model, accuracy gradually improves from 90.7% with PSO to 92.7% with GWO and 92.5% with SMO, peaking at 93.48% with the proposed ESMOPL method. In the BiLSTM model, a similar trend is observed where accuracy increases from 93.56% with PSO and 93.33% with GWO to 94.87% using SMO, and further to 95.87% with ESMOPL. The hybrid DCNN-BiLSTM model achieves the highest overall performance, starting at 95.2% with PSO, rising to 96% with GWO, 96.3% with SMO, and reaching a maximum of 97.9% with ESMOPL. These results clearly indicate that ESMOPL consistently outperforms traditional meta-heuristic optimizers across all models, with the hybrid architecture achieving the highest classification accuracy.

The training time comparison in Table 5 shows that the proposed ESMOPL-based feature selection consistently reduces training time across all models. For DCNN, training time decreases from 1234 s (PSO) to 1065 s (ESMOPL), while BiLSTM decreases from 1323 s to 1140 s. Similarly, the hybrid DCNN-BiLSTM model achieves a notable reduction from 1679 s (PSO) to 1431 s (ESMOPL). This improvement is attributed to the selection of a compact, discriminative feature subset, which reduces computational complexity and

accelerates convergence. The DCNN-BiLSTM needs ~392K trainable parameters for the original 547 features and ~360K for 300 features selected using ESMOPL.

Table 5. Results comparison with existing algorithms

Algorithm	Feature Selection	Overall Accuracy for 4 Class	Training Time (sec)
DCNN	PSO	90.7	1234
	GWO	92.702	1231
	SMO	92.545	1176
	ESMOPL	93.484	1065
BiLSTM	PSO	93.56	1323
	GWO	93.33	1259
	SMO	94.871	1220
	ESMOPL	95.868	1140
DCNN-BiLSTM	PSO	95.2	1679
	GWO	96	1676
	SMO	96.3	1548
	ESMOPL	97.9	1431

Table 6. Ablation study for DCNN-BiLSTM for MAF for 300 features chosen using ESMOPL for 4-class VPD

Epoch	Optimization Algorithm	Learning Rate Vaues	Accuracy
50	SGDM	0.1	78.25%
		0.01	86.65%
		0.001	89.10%
		0.0001	88.75%
	RMSProp	0.1	80.75%
		0.01	86.35%
		0.001	89.20%
		0.0001	88.45%
	Adam	0.1	82.23%
		0.01	87.35%
		0.001	92.90%
		0.0001	91.20%
100	SGDM	0.1	82.25%
		0.01	90.65%
		0.001	93.10%
		0.0001	92.75%
	RMSProp	0.1	84.75%
		0.01	90.35%
		0.001	93.20%
		0.0001	92.45%
	Adam	0.1	86.23%
		0.01	91.35%
		0.001	95.90%
		0.0001	94.20%
200	SGDM	0.1	85.25%
		0.01	93.65%
		0.001	96.10%
		0.0001	95.75%
	RMSProp	0.1	87.75%
		0.01	93.35%
		0.001	96.20%
		0.0001	95.45%
	Adam	0.1	89.23%
		0.01	94.35%
		0.001	97.90%
		0.0001	96.20%

Table 6 presents an ablation study of the proposed DCNN-BiLSTM model for 4-class voice pathology detection (VPD) using 300 MAF features selected by ESMOPL, highlighting the influence of optimization algorithms, learning rates, and training epochs. At 50 epochs, Adam consistently outperforms SGDM and RMSProp, achieving a maximum accuracy of 92.90% at a learning rate of 0.001, whereas SGDM and RMSProp reach 89.10% and 89.20%, respectively. Increasing the training to 100 epochs improves overall performance, with Adam attaining 95.90% accuracy at 0.001, compared to 93.10% (SGDM) and 93.20% (RMSProp). The best performance is observed at 200 epochs, where Adam with a learning rate of 0.001 achieves the highest accuracy of 97.90%, demonstrating stable convergence and superior optimization capability. These results confirm that Adam, with a moderate learning rate and sufficient training epochs, is most effective for the proposed VPD framework.

To assess the stability of the results, the model is analyzed for performance based on mean accuracy, standard deviation, and confidence interval (CI) of the proposed methods using 10-fold cross-validation with 300 features selected using ESMOPL for 4-class VPD, as given in Table 7. The DCNN model achieves a mean accuracy of 93.25% with a standard deviation of ± 1.85%, resulting in a 95% confidence interval (CI) of [91.60%, 94.90%], indicating moderate variability across training settings. The BiLSTM model shows slightly improved stability with a mean accuracy of 93.80%, a standard deviation of ± 1.62%, and a 95% CI of [92.35%, 95.25%]. The proposed DCNN-BiLSTM framework demonstrates the most consistent and superior performance, achieving a higher mean accuracy of 96.55% with a lower standard deviation of ± 1.10%, yielding a narrow 95% CI of [95.55%, 97.55%], confirming its robustness and reliability for voice pathology detection.

Table 7. Statistical result analysis for DCNN-BiLSTM for VPD for 10-fold cross validation

Method	Mean Accuracy	Standard Deviation	CI for 95%
DCNN	93.25	± 1.85%	[91.60%, 94.90%]
BiLSTM	93.80%	± 1.62%	[92.35%, 95.25%]
DCNN-BiLSTM	96.55%	± 1.10%,	[95.55%, 97.55%]

Table 8. Results comparison with existing algorithms

Author & Year	Feature Extraction	Classifier/Model	Accuracy
Albadr et al., 2025	MFCC	FLN	84.64%
Xiong et al., 2025	Task-oriented features	AFE + CCL module	87.94%
Er & İlhan (2024)	HHT + LSTM embeddings + CCA fusion	SVM, RF, KNN	89.54%
Proposed Method	MAF + ESMOPL	DCNN-BiLSTM-	97.9 %

The performance comparison with recent voice pathology detection studies highlights the superiority of the proposed method as given in Table 8. Albadr et al. (2025), using MFCC features and an FLN classifier, achieved 84.64% accuracy, while Xiong et al. (2025) improved performance to 87.94% using task-oriented features combined with AFE and CCL modules. Er and İlhan (2024) further enhanced accuracy to 89.54% by integrating HHT-based features, LSTM

embeddings, and CCA fusion with traditional classifiers such as SVM, RF, and KNN. In contrast, the proposed approach, which utilizes MAF features optimized through ESMOPL and is classified using a DCNN-BiLSTM architecture, significantly outperforms prior methods, achieving 97.9% accuracy and demonstrating its effectiveness in capturing both spectral characteristics and temporal dynamics of pathological speech.

4. CONCLUSIONS

This work highlights the importance of accurate voice pathology detection, as vocal disorders often present with overlapping symptoms and irregular speech variations, making diagnosis challenging. The proposed approach combines a 1D-DCNN and a BiLSTM to learn local correlation and temporal patterns in MAFs of voice signals, while the improved Spider Monkey Optimization algorithm selects the most relevant features, reducing computational complexity without compromising performance. Experimental results on the Saarbrücken Voice Database demonstrate high accuracy, achieving 97%–100% for binary classification and 97.90% for four-class classification involving normal, cyst, dysphonia, and paralysis cases.

In the future, this work can be extended by using larger, multilingual datasets to improve generalization, developing lightweight, real-time models for mobile or edge deployment, and using generative models to synthesize additional pathological data for underrepresented classes. In future work, we plan to integrate explainable AI techniques, such as attention visualizations, saliency maps, and SHAP/LIME-based feature attribution, to highlight influential features and time segments, thereby improving model transparency and clinical trust without compromising performance. Moreover, future clinical validation and integration with telemedicine systems may help transition the proposed method from research to practical healthcare applications.

REFERENCES

[1] Wagdarikar, N., Jagtap, S. (2025). Two-way voice feature representation for disease detection based on voice using 1D and 2D deep convolution neural network. Applied Acoustics, 233: 110615. https://doi.org/10.1016/j.apacoust.2025.110615

[2] Bogdan, F., Lascu, M.R. (2025). Advances and challenges in deep learning for acoustic pathology detection: A review. Technologies, 13(8): 329. https://doi.org/10.3390/technologies13080329

[3] Bhangale, K.B., Kothandaraman, M. (2022). Survey of deep learning paradigms for speech processing. Wireless Personal Communications, 125: 1913-1949. https://doi.org/10.1007/s11277-022-09640-y

[4] Bhangale, K., Mohanaprasad, K. (2022). Speech emotion recognition using mel frequency log spectrogram and deep convolutional neural network. In Futuristic Communication and Network Technologies. VICFCNT 2020. Lecture Notes in Electrical Engineering, pp. 241-250. https://doi.org/10.1007/978-981-16-4625-6_24

[5] Sankaran, A., Kumar, L.S. (2024). Advances in automated voice pathology detection: A comprehensive review of speech signal analysis techniques. IEEE Access, 12: 181127-181148. https://doi.org/10.1109/ACCESS.2024.3508884

[6] Islam, R., Tarique, M., Abdel-Raheem, E. (2020). A survey on signal processing based pathological voice detection techniques. IEEE Access, 8: 66749-66776. https://doi.org/10.1109/ACCESS.2020.2985280

[7] Bhangale, K., Kothandaraman, M. (2023). Speech emotion recognition based on multiple acoustic features and deep convolutional neural network. Electronics, 12(4): 839. https://doi.org/10.3390/electronics12040839

[8] Liu, G.S., Jovanovic, N., Sung, C.K., Doyle, P.C. (2024). A scoping review of artificial intelligence detection of voice pathology: Challenges and opportunities. Otolaryngology–Head and Neck Surgery, 171(3): 658-666. https://doi.org/10.1002/ohn.809

[9] Mohammed, M.A., Abdulkareem, K.H., Mostafa, S.A., Abd Ghani, M.K., Maashi, M.S., Garcia-Zapirain, B., Oleagordia, I., Alhakami, H., AL-Dhief, F.T. (2020). Voice pathology detection and classification using convolutional neural network model. Applied Sciences, 10(11): 3723. https://doi.org/10.3390/app10113723

[10] Bhangale, K.B., Mohanaprasad, K. (2021). A review on speech processing using machine learning paradigm. International Journal of Speech Technology, 24: 367-388. https://doi.org/10.1007/s10772-021-09808-0

[11] Albadr, M.A.A., Ayob, M., Tiun, S., AL-Dhief, F.T., Al-Daweri, M.S., Homod, R.Z., Abbas, A.H. (2025). Fast learning network algorithm for voice pathology detection and classification. Multimedia Tools and Applications, 84: 18567-18598. https://doi.org/10.1007/s11042-024-19788-3

[12] Latiff, N.M.A., Al-Dhief, F.T., Md Sazihan, N.F.S., Baki, M.M., Malik, N.N.N.A., Albadr, M.A.A., Abbas, A.H. (2025). Voice pathology detection using machine learning algorithms based on different voice databases. Results in Engineering, 25: 103937. https://doi.org/10.1016/j.rineng.2025.103937

[13] Geng, L., Liang, Y., Shan, H.F., Xiao, Z.T., Wang, W., Wei, M. (2025). Pathological voice detection and classification based on multimodal transmission network. Journal of Voice, 39(3): 591-601. https://doi.org/10.1016/j.jvoice.2022.11.018

[14] Abdulmajeed, N.Q., Al-Khateeb, B., Mohammed, M.A. (2025). Voice pathology identification system using a deep learning approach based on unique feature selection sets. Expert Systems, 42(1): e13327. https://doi.org/10.1111/exsy.13327

[15] Cala, F., Frassineti, L., Cantarella, G., Buccichini, G., Battilocchi, L., Manfredi, C., Lanatà, A. (2025). Towards an explainable artificial intelligence system for voice pathology identification and post-treatment characterisation. Biomedical Signal Processing and Control, 104: 107530. https://doi.org/10.1016/j.bspc.2025.107530

[16] Gulsen, P., Gulsen, A., Alci, M. (2025). Machine learning models with hyperparameter optimization for voice pathology classification on Saarbrücken Voice Database. Journal of Voice. https://doi.org/10.1016/j.jvoice.2024.12.009

[17] Xiong, Y.F., Guo, D.Y., Shen, L.P., Mo, W., Yang, H., Lin, Y. (2025). Adversarial feature disentanglement framework for voice pathology detection. In ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

Hyderabad, India, pp. 1-5. https://doi.org/10.1109/ICASSP49660.2025.10888569
[18] Bashir, R.N., Shahid, M.A., Rashid, T., Faheem, M., Saidani, T., Saidani, O., Khan, A.R. (2025). Voice pathology identification using mel spectrogram features and deep learning. Signal, Image and Video Processing, 19: 909. https://doi.org/10.1007/s11760-025-04527-4
[19] Brindha, G., S, P., N, S.N., G, V.K., K, V.K. (2025). Detection of voice pathologies and classification using deep learning in healthcare. In 2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, pp. 1945-1950. https://doi.org/10.1109/ICIRCA65293.2025.11089865
[20] S, S., S, A.K., V, S.K., M, G.P., T, I., S, D. (2025). Voice pathology detection using machine learning and electroglottography signals. In 2025 3rd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, pp. 1-4. https://doi.org/10.1109/ICAECA63854.2025.11012399
[21] Farazi, S., Shekofteh, Y. (2025). Evaluation of phone posterior probabilities for pathology detection in speech data using deep learning models. International Journal of Speech Technology, 28: 99-116. https://doi.org/10.1007/s10772-024-10166-w
[22] Fu, D.L., Zhang, X.H., Chen, D.D., Hu, W.P. (2025). Pathological voice detection based on phase reconstitution and convolutional neural network. Journal of Voice, 39(2): 353-364. https://doi.org/10.1016/j.jvoice.2022.08.028
[23] Zhang, J., Qiu, Y.Y., Liu, Y.C., Xiao, Y., Yang, J.Y., Yang, X., Ma, M., Song, A.G. (2025). Early stroke diagnosis and evaluation based on pathological voice classification using speech enhancement. Computers in Biology and Medicine, 196: 110940. https://doi.org/10.1016/j.compbiomed.2025.110940
[24] El Omari, M., El Belghiti, Y., Belmajdoub, H., Minaoui, K., Saoudi, S. (2025). Early detection of voice pathology from cry analysis using non-interpretable features and parallel 1D CNN. In Engineering Applications of Neural Networks. EANN 2025. Communications in Computer and Information Science, pp. 100-111. https://doi.org/10.1007/978-3-031-96196-0_8
[25] Dai, Z.Y., Jiang, Y.Y., Cao, L.Y., Zhang, X.J., Tao, Z. (2025). MSDFEN: Multi-scale dynamic feature extraction network for pathological voice detection. Applied Acoustics, 230: 110438. https://doi.org/10.1016/j.apacoust.2024.110438
[26] Farazi, S., Shekofteh, Y. (2024). Voice pathology detection on spontaneous speech data using deep learning models. International Journal of Speech Technology, 27: 739-751. https://doi.org/10.1007/s10772-024-10134-4
[27] Özbay, E., Özbay, F.A., Khodadadi, N., Gharehchopogh, F.S., Mirjalili, S. (2024). Multifeature fusion method with metaheuristic optimization for automated voice pathology detection. Journal of Voice. https://doi.org/10.1016/j.jvoice.2024.08.018
[28] Tirronen, S., Kadiri, S.R., Alku, P. (2024). The effect of the MFCC frame length in automatic voice pathology detection. Journal of Voice, 38(5): 975-982. https://doi.org/10.1016/j.jvoice.2022.03.021
[29] Won, J.H., Kim, D.H. (2024). Metric-based few-shot transfer learning approach for voice pathology detection. IEEE Access, 12: 159226-159238. https://doi.org/10.1109/ACCESS.2024.3480523.
[30] Arslan, Ö. (2024). A machine learning approach for voice pathology detection using mode decomposition-based acoustic cepstral features. Mathematical Modelling and Numerical Simulation with Applications, 4(4): 469-494. https://doi.org/10.53391/mmnsa.1473574
[31] Jegan, R., Jayagowri, R. (2024). Optimized early fusion of handcrafted and deep learning descriptors for voice pathology detection and classification. Healthcare Analytics, 6: 100369. https://doi.org/10.1016/j.health.2024.100369
[32] Er, M.B., İlhan, N. (2025). Voice pathology detection based on canonical correlation analysis using Hilbert–Huang transform and LSTM features. Arabian Journal for Science and Engineering, 50: 11693-11711. https://doi.org/10.1007/s13369-024-09599-x
[33] Zhao, D.H., Qiu, Z.X., Jiang, Y.J., Zhu, X.C., Zhang, X.J., Tao, Z. (2024). A depthwise separable CNN-based interpretable feature extraction network for automatic pathological voice detection. Biomedical Signal Processing and Control, 88: 105624. https://doi.org/10.1016/j.bspc.2023.105624
[34] Sharma, H., Hazrati, G., Bansal, J.C. (2018). Spider monkey optimization algorithm. In Evolutionary and Swarm Intelligence Algorithms. Studies in Computational Intelligence, pp. 43-59. https://doi.org/10.1007/978-3-319-91341-4_4
[35] Agrawal, V., Rastogi, R., Tiwari, D.C. (2018). Spider monkey optimization: A survey. International Journal of System Assurance Engineering and Management, 9: 929-941. https://doi.org/10.1007/s13198-017-0685-6
[36] Pützer, M., Barry, W.J. (2008). Saarbruecken voice database. Zenodo, v2. https://doi.org/10.5281/zenodo.16874898