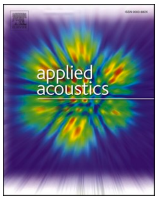
Applied Acoustics 233 (2025) 110615

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust



Two-way voice feature representation for disease detection based on voice using 1D and 2D deep convolution neural network

Narendra Wagdarikar [a,b,*], Sonal Jagtap [c]

[a] Department of E&TC Engineering, G H Raisoni College of Engineering and Management, Wagholi, Pune, India
[b] Department of Electronics and Telecommunication, Smt. Kashibai Navale College Of Engineering, Vadgaon(Bk), Pune, India
[c] Smt. Kashibai Navale College Of Engineering, Department of E&TC Engineering, Vadgaon(Bk), Pune, India

ARTICLE INFO

Keywords:
Voice pathology
Deep learning
Deep convolution neural network
Multiple voice features

ABSTRACT

Voice pathology deals with detecting diseases with the help of the voice, as diseases significantly impact the voice. Machine learning (ML) and deep learning (DL) schemes have been presented for disease detection using voice. However, the outcomes of the system are limited due to poor spectro-temporal representation, less feature distinctiveness, low-frequency resolution problems, lower detection rates, etc. This article presents voice-based pathology using two-way voice feature representation (TWVFR), which consists of two parallel arms of a Deep Convolution Neural Network (DCNN) for feature representation. The first parallel arm considers the Mel Frequency Cepstral Coefficient Spectrogram (MFCCS), fed to 2-D DCNN to characterize the spectral domain characteristics of the voice signal. The second approach consists of multiple voice features (MVF), such as spectral domain (SD), time domain (TD), and voice quality (VQ) features. The essential features are selected using the Spider Monkey optimization algorithm and given to 1D-DCNN. The last layer features are combined and given to a fully connected layer followed by the Softmax classifier. The Softmax classifier classifies the speech signal into normal and diseased voices. The system outcomes are validated on the Saarbruecken Voice Dataset (SVD) for four class disease classifications: Bulb Paralysis, Cyste, Polyp, and Normal. The suggested TWVFR scheme provides improved overall accuracy of 98.33%, recall of 0.98, precision of 0.98, F1-score of 0.98, selectivity of 0.98, and negative predictive rate of 0.98 compared to existing methods. The TWVFR helps to enhance the feature depiction and provides an overall accuracy of 98.33% than MVF-1D DCNN (95.45%). The suggested TWVFR scheme provides improved overall accuracy of 98.33%, recall of 0.98, precision of 0.98, F1-score of 0.98, selectivity of 0.98, and negative predictive rate of 0.98 compared to existing methods.

1. Introduction

Voice is an effective, reliable, and simple way of human communication. Human voice depicts the verbal as well as emotional content. Stress, affective states, diseases, age, gender, and disorder generally influence the human voice. A variety of circumstances may cause voice disorders. Among them include facial discomfort, fatigue, environmental changes, infections of the vocal tissue, and muscular dystrophy [1]. Vocal noise surges due to the detrimental effect of voice pathology on voice functioning and vibration consistency. The typical voice became weak, strained, and hoarse [2], impacting the voice's quality [3]. Currently, subjective factors are used to skew the assessment of voice pathology detection systems [4]. The auditory-perceptual assessment, often used with visual laryngostroboscopy assessment in hospitals, is an example of a subjective evaluation [5]. Several clinical procedures are used to measure auditory-perceptual characteristics to quantify the rate of severity diagnosis [6].

Nevertheless, such assessment techniques are time-consuming, labor-intensive, and highly sensitive to parameter variations [7]. Furthermore, they require a comprehensive physical evaluation of the patient, which may be challenging for those with serious illnesses. Identifying and analyzing voice sounds with a computer-assisted instrument without surgery is an example of objective assessment. Inaudible noises may also be determined by automated detection [1]. Since these evaluation techniques are objective, they don't rely on human judgment.

Additionally, they are simple to use since several online recording programs allow speech recordings to be accessed remotely. Studies like

[*] Corresponding author at: Department of E&TC Engineering, G H Raisoni College of Engineering and Management, Wagholi, Pune, India.
E-mail address: wagdarikarnarendra@gmail.com (N. Wagdarikar).

https://doi.org/10.1016/j.apacoust.2025.110615
Received 7 June 2024; Received in revised form 28 November 2024; Accepted 17 February 2025
Available online 25 February 2025

[8] have developed vocal processing methods that can effectively combine with a machine learning method to detect voice pathology automatically within a single framework. We do this to distinguish healthy individuals from those with voice pathologies accurately. The research literature extensively uses several vocal pathology databases to evaluate voice pathology objectively. The most frequently used databases for the study of voice pathology are the Saarbruecken Voice Database (SVD) [9], the Arabic Voice Pathology Database (AVPD) [10], and the Massachusetts Eye and Ear Infirmary Database (MEEI) [11]. Academics often analyze the vocalization of the vowel /a/ [1,11] due to its easy availability in various language databases [2]. Researchers also investigate additional vowel pairings [1,12]. Interestingly, most researchers involved in vocal diseases have limited their datasets to certain kinds of abnormalities [12].

The traditional methods of voice pathology include the manual inspection and observation of the patient, where the experts interact with the patient or ask them to read specific paragraphs. The manual voice pathological diagnosis is inaccurate and less precise because of fatigue, tiredness, inadequate knowledge of the expert, and the unresponsive nature of the patient [6,7]. The automatic ML and DL-based voice pathology systems have shown noteworthy contributions and helped to increase the preciseness and reliability of the system. The DL-based voice pathological systems offer superior feature representation, the capability to handle larger datasets, higher classification rates, and better distinctive features [8,9]. Most DL-based voice pathology systems use a spectrogram representation of voice as input for the DL framework. However, the performance of such systems is limited because of poor correlation in the spectral and time-domain characteristics of the pathological voice, poor generalization capability, complex DL framework, fewer feature distinctiveness, and inability to capture correlation in long-term and short-term attributes of the voice. Therefore, this paper presents the TWVFR that utilizes the voice spectrogram to characterize the spatial and spectral properties of the voice. It also uses traditional SD, TD, and VQ features to offer the distinctive spectro-temporal properties of pathological voices [10–12].

This paper presents the TWVFR scheme for voice disease detection. The main contributions of the paper are summarized as follows:

- Voice representation using multiple voice features that provide spectral, temporal, and voice quality attributes of the pathological voice and 1D DCNN. The 1D DCNN provides better correlation in the different SD, TD, and VQ features and offers superior correlation and dependency in the local and global voice features.
- The Improved Spider Monkey Optimization-based feature selection scheme selects prominent features from multiple voice features. The ISMO, with Levy's flight function, gradually increases the population's solution diversity and search space.
- Spectral and spatial feature representation of pathological voice using MFCCS and 2D DCNN. The 2D DCNN offers the spatial and spectral characteristics of the voice spectrogram to depict the minor changes in the voice spectrum due to abnormality.

The remaining article is organized as follows: Section 2 delivers the information regarding material and methodology. Section 3 offers the experimental results and discusses them. Section 4 depicts the conclusion and future scopes of the proposed ASSR for a potential boost in performance.

2. Related work

Voice pathology has attracted the wide attention of researchers who have deployed distinct ML and DL-based schemes for automatic analysis of voice pathology. In most cases, the process of illness diagnosis takes place following the clinical interpretation of the voice's features [13]. Some of the things that make up vocal characteristics are the glottal-to-noise excitation ratio (GNE) [14], the Mel frequency cepstral coefficients (MFCC) [15], the multidimensional voice program parameters (MDVP) [16], and many more. For further information about the circumstances behind speech difficulties, please refer to [17]. After retrieving the vocal features, various conventional classification methods detect voice pathology. The vast majority of studies have used classifiers [19,20], Gaussian mixture models (GMM), random forests (RF), artificial neural networks (ANN) [18], and support vector machines (SVM) to identify targets. It has been observed that the results of the study tend to vary significantly from one another. A classification method, a vocal feature, and a sample of voice pathology are employed in the investigations due to the various set choices that are applied. Based on these observations, we can draw the following conclusions: The majority of research focuses on one speech task, namely the sustained phonation of the vowel /a/ (also known as a language-independent speech task), is the focus of the majority of the research. Most research primarily focuses on examining a single voice segment. The vowel /a/ is particularly important within the autonomous language-speaking task context. Most analyses on restricted acoustic diseases utilize the SVD, AVPD, and MEEI databases.

Most frequently, researchers retrieve the conventional dysphonic feature to define the voice aspect of a specific individual vocal disorder. Artificial neural networks (ANN), random forests (RF), and support vector machines (SVM) are the most commonly used machine learning techniques for diagnosing vocal disorders [22]. There are a variety of applications that might benefit from the employment of machine learning (ML), including medical diagnostics [23], cancer detection [24], intelligent building applications [25], and other applications [1,11,26]. Machine learning techniques benefit jobs involving selective detection and classification [27,28]. These technologies are used in the analysis of diseased voices [29]. These methods have been used in numerous speech identification applications. Among the many challenging aspects of speech detection research, one of the most challenging areas is still identifying and categorizing voice pathology procedures [30]. Karaman et al. [31] presented transfer learning for Parkinson's disease detection. It provides an overall accuracy of 89.75 % for DenseNet161. The models show the feasibility of the implementation on standalone devices.

Further, Tuncer et al. [32] proposed a 1D Local binary pattern (LBP) for texture feature representation of the pathological voice. It provided an accuracy of 98.73 % for the SVD dataset. The LBP delivers the impact of the disease on the local prosody of the voice. Further, Tuncer and Dogan [33] explored dynamic LBP to enhance the effectiveness of the conventional LBP that gave 89.17 % accuracy for the daily voice dataset. Tuncer et al. [34] presented multilevel texture features for representing the voice characteristics of the diseased voice. It presented the multi-centered and multi-threshold ternary pattern for describing the attributes of the diseased voice. It provides enhanced results for the SVM classifier (98.09 %) for classifying the healthy and cordectomy from the SVD dataset. It is observed that the feature selection using iterative neighborhood component analysis (INCA) assists in selecting the prominent features to minimize the computational intricacy of the system. Fujimura et al. [35] provided 1D DCNN for the hierarchical representation of the voice signal for disease detection. It provided 88.3 % accuracy for voice disease detection. The voice pathology system's chief goal is to determine the voice's grade, roughness, breathiness, asthenia, and strain (GRBAS) to characterize the impact of disease on the voice. Hammami et al. [36] utilized the empirical mode decomposition and discrete wavelet transform (EMD-DWT) to analyze voice disease detection. It resulted in 99.29 % accuracy for the SVD dataset (Normal, Laryngitis, Rekurrensparses, Hyperfunktionelle Dysphonia, Funktionelle Dysphonia, Dysphonia) Higher Order Statistics (HOS) features and SVM classifier. It provided 93.1 % accuracy for DWT-SVM. Al-Dhip et al. [37] presented MFCC for feature extraction and online sequential extreme learning machine (OSELM) for classification. It provided 85 % accuracy for two classes (600 samples/class) of pathology voice detection for the SVD dataset. MFCC provides the human perceptual capability to characterize the pathological voice.

Further, Syed et al. [38] proposed CNN and LSTM for pathological voice detection. It provides 87.1 % and 86.525 accuracy for CNN and LSTM, respectively, for 12-class disease detection for the SVD dataset. CNN provides a hierarchical representation of voice signals. The LSTM depicts the long-term dependencies and temporal characteristics of the pathological voice. It is observed that the synthetic minority oversampling technique (SMOTE) helps to enhance the effectiveness of the DL architectures. Lee and Ji [39] suggested that CNN along linear prediction cepstral coefficients (LPC), oversampled with SMOTE, provides 98.89 % accuracy for the classification of standard (687 samples) and pathological voice (1354 samples) of the SVD dataset. It results in a class imbalance problem for multiclass disease detection. This approach poorly represents the GRBAS scales of diseased voice. Abdulmajeed et al. [40] presented features such as MFCC, zero crossing rate (ZCR), and mel spectrograms for voice representation. The LSTM-based classifier provides an accuracy of 99.3 % for neutral pitch /u/ vowel samples and 99.2 % for sentence samples.

The extensive literature review shows that various DL-based schemes have shown noteworthy contributions to the overall effectiveness of voice pathology [45–47]. However, the effectiveness of the DL-based system is challenging because of the incapability to capture the minor variations over pathological voice, intricate DL framework, higher trainable parameters, imbalance in spectral and temporal properties, lower long-term dependency, and inferior voice representation. The spectrogram-based voice representation fails to provide the temporal characteristics and long-term correlation in the pathological voices [48–50]. Therefore, this research aims to provide superior spectral-temporal representation and better long-term correlation using the TWVFR system to enhance the effectiveness of voice pathology.

3. Material and methodology

This section details the material and methodology utilized to simulate the proposed DL-based voice pathology system.

3.1. Dataset

The Saarbruecken Speech Database [21], a collection of speech recordings and EGG signals from more than two thousand people, is used for disease detection. There are records of 687 persons in good health, including 428 males and females. The study comprised a total of 1356 people, 727 of whom were female and 629 of whom were male, who

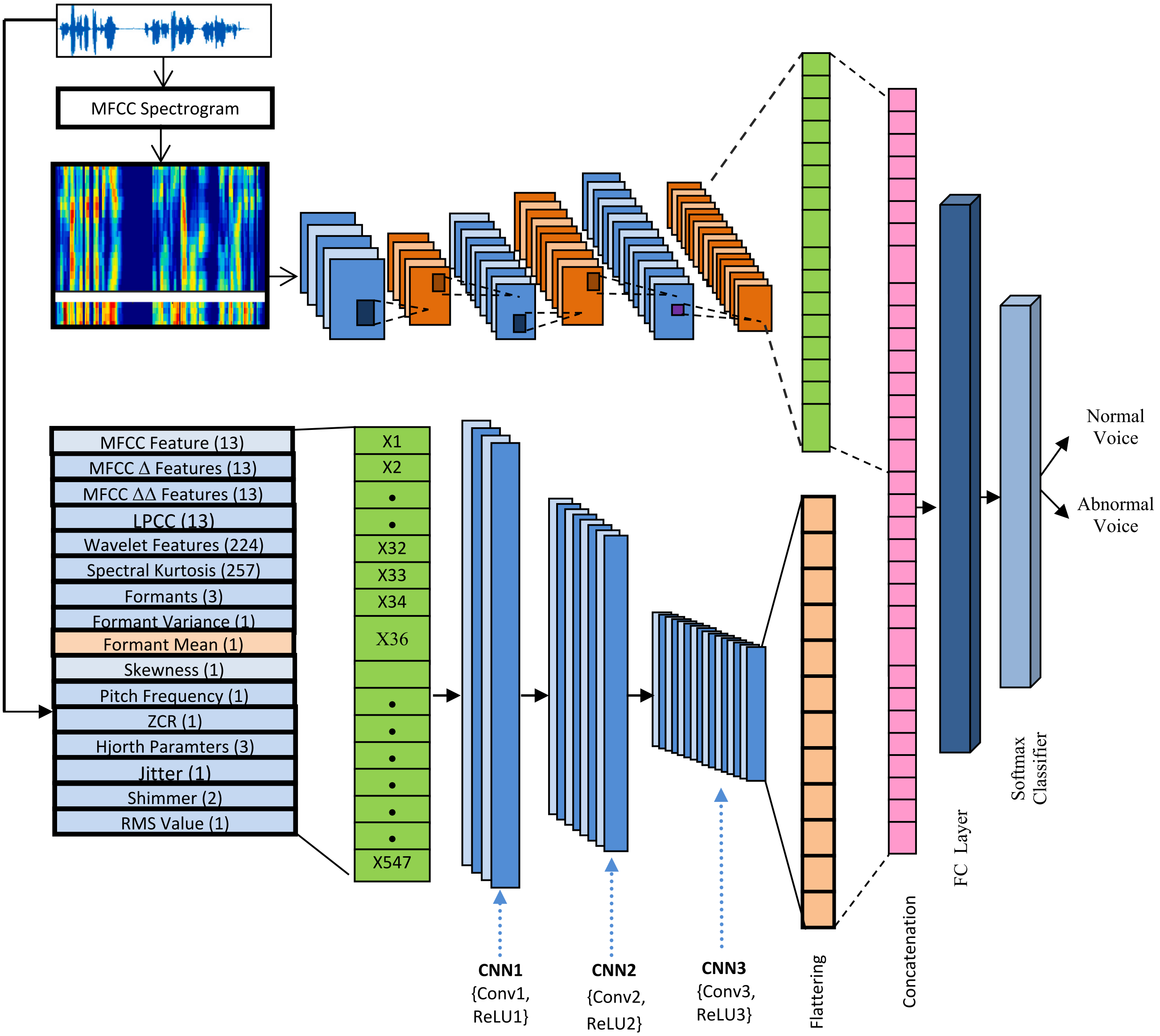


Fig. 1. Process of proposed pathological voice detection.

were diagnosed with one or more of the 71 different disorders. A total of 259 males and 1356 females were present. We have considered 200 samples for each of the three pathologies, such as paralysis, cysts, polyps, and normal voice. Within the confines of a single recording session, there are recordings of the following components that are included: In addition to the standard pitch, the vowels /i/, /a/, and /u/ may also be pronounced at high and low pitch. It is possible to generate the vowels /i/, /a/, and /u/ using a rising-falling pitch. It also consists of a phrase in German: “Guten Morgen, wie geht es Ihnen?” (“Good morning, how are you?”). Every single one of the samples of the sustained vowels has a duration that ranges from one to three seconds, and they were all recorded at a frequency of fifty kilohertz with a resolution of sixteen bits.

3.2. Methodology

The proposed DL-based voice flow diagram is illustrated in Fig. 1. It consists of TWVFR of the speech signal for pathological voice-based disease detection.

The first approach consists of a 2-D MFCC spectrogram given to 2-D DCNN to capture the spectro-temporal characteristics of the signal. The second approach comprises acoustic features such as SD, TD, and VQ. The essential features are selected using the Spider Monkey optimization algorithm and given to 1D-DCNN. The last layer features are combined and given to a fully connected layer followed by the Softmax classifier. The Softmax classifier classifies the speech signal into normal and diseased voices.

3.3. Multiple acoustic acoustic features

Acoustics aspects of the speech signal are the physical qualities of the speech signal, which include frequency, amplitude, and loudness. The speech signal represents these features. The acoustic feature set presented comprises discrete SD, TD, and VQ components responsible for characterizing speech illness. Acoustic features that have been extracted include the Mel frequency cepstral coefficients (MFCC), the linear prediction cepstral coefficients (LPCC), the Wavelet Packet Transform (WPT), the zero crossing rate (ZCR), the spectrum centroid, the spectral rolloff, the root mean square (RMS), the spectral kurtosis (SK), the jitter, the shimmer, the pitch frequency, the formants, and the mean and standard deviation of the formants. A moving average filter is applied to the transmission of the speech signal before the different characteristics are computed to reduce the amount of noise and disruptions present in the speech signal. Information on the various MVFs can be found in Table 1.

A. MFCC.

MFCC provides the spectral information of the speech and characterizes the human hearing perception. Fig. 2 shows the process flow of computation of MFCC coefficients [47].

Normalization of the raw signal voice signal is accomplished by the use of pre-emphasis during the process of MFCC coefficient extraction. The pre-emphasis is effective by reducing the amount of noise and disruptions present in the raw speech (x(n)). With a frameshift of fifty percent, or twenty milliseconds, the filtered signal is then divided into frames that are forty milliseconds long. When assuming a frame width of 40 ms and a 50 % overlap, a total of 199 frames are created for voice signals that are four seconds long. In addition, a single hamming window with a value of α equal to 0.46 and several samples per frame length (N) of 30 ms can collect the frequency components closest to one another, as shown by equation (1).

$$\mathrm{H(n)} = (1-\alpha) - \alpha.\cos\left(\frac{2\pi n}{(N-1)}\right), \quad 0 \le n \le N-1 \tag{1}$$

As shown in (2), the following step involves the use of the Discrete Fourier Transform (DFT) to transform the time-domain pathology voice signal into the frequency-domain (X(k)). One example of the features of the vocal tract is provided by Equation 3, which is the power spectrum of the DFT. To get the speech-hearing perceptual information, the signal is next processed by passing it through M(24) number of Mel Frequency triangle filter banks ($\nabla_m(k)$) as described in equation (4). The transfer of linear frequency to Mel frequency and vice versa may be accomplished using equations 5 and 6.

Table 1
Details of MVFs.

Types of Features	Feature	Number of Features	Total Features	Total MVF
Spectral Features	MFCC	13	539	**547**
	MFCCΔ	13		
	MFCCΔΔ	13		
	LPCC	13		
	Spectral Kurtosis	257		
	Formants	3		
	Formant Mean	1		
	Formant Variance	1		
	Skewness	1		
	Wavelet Features	224		
Time Domain Features	Pitch Frequency	1	5	
	ZCR	1		
	Hjorth Parameter (Activity, Mobility, Complexity)	3		
Voice Quality Features	Jitter	1	3	
	Shimmer	1		
	RMS	1		

$$\mathrm{X(k)} = \sum_{n=0}^{N-1} \mathrm{x(n).H(n).e^{-j2\pi nk/N}}, \; 0 \le n, k \le N-1 \tag{2}$$

$$X_k = \frac{1}{N}|X(k)|^2 \tag{3}$$

$$\mathrm{ET_m} = \sum_{k=0}^{k=1} \nabla_m(k).X_k; \;\; m = 1, 2, \cdots.M \tag{4}$$

$$\mathrm{Mel} = 2595\log\left(1+\frac{f}{700}\right) \tag{5}$$

$$f = 700\left(10^{\frac{Mel}{2595}} - 1\right) \tag{6}$$

Afterward, the Discrete Cosine Transform (DCT) of the log-filter bank energy signal provides L number of cepstral coefficients as given by (7).

$$\mathrm{MFCC_i} = \sum_{m=1}^{M} \log_{10}(\mathrm{ET_m}).cosj\left((m+0.5)\frac{\pi}{m}\right) \quad forj = 1, 2, \cdots L \tag{7}$$

The MFCC provides a total of 39 features. These characteristics include one feature: the energy of the speech signal, 12 MFCC coefficients, and 26 first- and second-order derivatives of the MFCC features. The derivative qualities are essential to accurately characterize the transition in aberrant speech [26,27].

B. RMS.

Based on the root mean squares of the amplitudes of the speech samples (x_i), the root mean squares (RMS) (x_{RMS}) Offers the loudness of the pathology signal. An estimate of the root mean square (RMS) of the speech signal with N samples is provided by Equation 8.

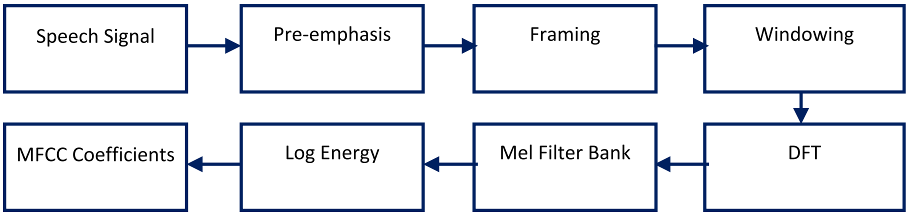


Fig. 2. Process flow of MFCC.

$$x_{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{8}$$

C. ZCR.

The ZCR offers the transition of the signal over through the zero line, which reflects the amount of noise in the voice signal. The ZCR may be computed in the time domain using Equation 9, which is provided. A value of 1 is returned by the sign function for positive sample amplitude, while a value of 0 for negative sample amplitude over a given time frame (t).

$$ZCR_t = \frac{1}{2}\left(\sum_{n=1}^{N}(\mathrm{sign}(x[n]) - \mathrm{sign}(x[n-1]\right) \tag{9}$$

D. LPCC.

The LPCC is the SD feature produced from the linear predictive analysis to capture the speech quality-specific phonetic representation of the speech signal. The LPCC is effective at giving features of the human vocal tract, which allow for the unique characterization of the expressive content that is present in speech [46,47], and [48]. In linear predictive analysis, it is possible to estimate the *n* th samples using the information of the preceding samples, as shown in equation (12).

$$x(n) = a_1x(n-1) + a_2x(n-2) + a_3x(n-3) + \cdots\cdots\cdots + a_px(n-p), \tag{12}$$

where $a_1, a_2, \cdots a_p$ are the constants over the speech frame. These linear predictor coefficients predict the speech sample. Equation 13 is used to analyze the error between predicted $\widehat{x}(n)$ and actual sample x(n).

$$x(n) = x(n) - \widehat{x}(n) = x(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{13}$$

To obtain the unique predictive coefficients, the sum of the squared difference of error (e_n) between predicted $\widehat{x}(n)$ and actual sample x(n) is computed using (14). Here, m represents the number of samples in the frame.

$$e_n = \sum_m \left| x(m) - \sum_{k=1}^{p} a_k x(m-k) \right|^2 \tag{14}$$

The LP coefficients are computed by solving equation 15. The LPCC coefficients are calculated using (15—18).

$$\frac{dE_n}{da_k} = 0 \quad \text{for } k = 1, 2, 3, \cdots\cdots p \tag{15}$$

$$C_0 = \log_e(p) \tag{16}$$

$$LPCC_m = a_m + \sum_{k=1}^{m-1}\frac{k}{m}C_k a_{m-k}; \qquad \text{for } 1 < m < p \tag{17}$$

$$LPCC_m = \sum_{k=m-p}^{m-1}\frac{k}{m}C_k a_{m-k}; \qquad \text{for } m > p \tag{18}$$

The proposed approach considers 13 LPCC coefficients as features [4647].

E. Spectral Kurtosis.

Spectral kurtosis (SK) is a statistical measure that presents a sequence of transients and their positions in standard deviation. The non-Gaussianity or flatness of the speech spectrum at its centroid is characterized by this phenomenon, which demonstrates the influence of differences in arousal and valence in the speech spectrum [5,7,9]. When attempting to estimate the spectral kurtosis of the voice stream, Equation 19 is used accordingly.

$$SK = \frac{\sum_{k=b1}^{b2}(f_k - \mu_1)^4 s_k}{(\mu_2)^4 \sum_{k=b1}^{b2} s_k} \tag{19}$$

Here, μ_1 and μ_2 represents the spectral centroid and spectral spread, respectively, s_k is spectral value over k bins, b_1 and b_2 are the lower and upper bound of the bins where spectral skewness of speech is estimated.

F. Jitter and Shimmer.

Variations in frequency and amplitude of the pathology signal are called jitter and shimmer, respectively. Irregular vibrations of the vocal folds generate these variations. The sounds of jitter and shimmer are a representation of the hoarseness, roughness, and breathiness that are present in the sound. The jitter's average absolute value may be found using Equation 20 [13].

$$\text{Jitter} = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}| \tag{20}$$

Where T_i stands for the period in sec, and N represents a number of periods. Equation 21 represents the average value of shimmer.

$$\text{Shimmer} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N}A_i}, \tag{21}$$

where A_i is the peak-to-peak amplitude of pathology speech, and N depicts a number of periods.

G. Pitch Frequency.

Pitch (f_0) is significant in exemplifying the voiced part of speech. The pitch of the speech is estimated by computing the difference between the peaks derived from the autocorrelation of the speech signal [57].

H. Formants.

The peak frequencies in the speech spectrum with a more significant energy concentration are called formants. Specifically, it describes the resonance phenomena in the vocal tract, which is very useful for describing speech disorders' impact on the resonance phenomenon. The formants are obtained from the MFCC spectrogram, and for assessment, three formats, f_1, f_2, and f_3, are taken into consideration. Furthermore, to give the variances in the formants, the mean and standard deviation of formats are estimated using three different formats [5,7]. Equations 22–24 provide formants (fm), mean of formants (fm_u), and the standard deviation of formants (fm_σ) respectively.

$$fm = \{f_1, f_2, f_3\} \tag{22}$$

$$fm_u = \frac{f_1 + f_2 + f_3}{3} \tag{23}$$

$$\mathrm{fm}_{\sigma} = \sqrt{\frac{\sum_{i=1}^{3} (\mathrm{f_i} - \mathrm{fm_u})^2}{3}} \tag{24}$$

I Wavelet Packet Decomposition Features.

WPT depicts the intricate information of complex patterns of voice, images, and patterns. The WPT features the disparities in the voice at different scales and levels due to diseases. The Daubechies (db2) filter decomposes the voice into different frequency components. It enables analysis of the disease's effect at various frequency bands. The wavelet basis function $\Psi_j^i(n)$ utilized for decomposing the voice into L-level are described in Fig. 25 and 26.

$$\Psi_j^{2i}(n) = \sum_k h(k)\Psi_{j-1}^i\left(n - 2^{j-1}k\right) \tag{25}$$

$$\Psi_j^{2i+1}(n) = \sum_k g(k)\Psi_{j-1}^i\left(n - 2^{j-1}k\right) \tag{26}$$

The high pass Quadrature mirror filter (QMF) and low pass QMF utilized for the WPT are represented by $g(k)$ and $h(k)$ as shown in Equations 27 and 28.

$$h(k) = \left\langle \Psi_j^{2i}(u), \Psi_{j-1}^i\left(u - 2^{j-1}k\right) \right\rangle \tag{27}$$

$$g(k) = \left\langle \Psi_j^{2i+1}(u), \Psi_{j-1}^i\left(u - 2^{j-1}k\right) \right\rangle \tag{28}$$

Equation 29 splits the pathology voice into subcomponents at level j, where $X_j^i(k)$ denotes k^{th} WPT at ith packet at j level.

$$x(n) = \sum_{i,k} X_j^i(k)\Psi_j^i\left(n - 2^j k\right) \tag{29}$$

Equation 30 describes the energy computation of the local wavelet.

$$X_j^i(k) = \left\langle x(n), \Psi_j^i\left(n - 2^j k\right) \right\rangle \tag{30}$$

The wavelet coefficient $X_j^i(k)$ denotes the confined WPT weights depicted by $\Psi_j^i\left(n - 2^j k\right)$ as given in Equation 31.

$$X_j^i(k) = \left\langle x(n), \Psi_j^i\left(n - 2^j k\right) \right\rangle \tag{31}$$

Equation 32 gives a distinct WPT set for the L level.

$$X_L(k) = \begin{bmatrix} X_L^0(k) \\ X_L^1(k) \\ . \\ . \\ X_L^{2^{L-1}}(k) \end{bmatrix} \tag{32}$$

The voice is decomposed into five levels to analyze the impact of disease on the human voice. The fifth level of WPT decomposition generates 32 coefficients. Seven statistical and spectral features are computed for every packet, including kurtosis, skewness, median, mean, energy, variance, and standard deviation. The seven features of 32 packets produce a total of 224 WPT features. These features characterize the effect of disease on the intonation, prosody, and timbre of the voice.

The overall feature consists of a total of 547 features. The SD, TD, and VC features are concatenated to form the final feature vector. The SD, TD, VC, and total features are represented by equations 33–36, respectively. The MVFs are provided to the 1D DCNN to enhance the connectivity and representation power of the global and local voice features.

$$Feat_{SD} = \{MFCC_{1-39}, LPCC_{1-13}, SC_{1-257}, Fm_{1-3}, Fmm_1, Fmv_1, SK_1, WPT_{1-224}\} \tag{33}$$

$$Feat_{TD} = \{PF_1, ZCR_1, HP_{1-3}\} \tag{34}$$

$$Feat_{SD} = \{Jitter_1, Shimmer_1, RMS_1\} \tag{35}$$

$$Feat = \{Feat_{SD}, Feat_{TD}, Feat_{VC}\} \tag{36}$$

3.4. Feature selection using SMO

The SMO algorithm is utilized to select the prominent features of the MVF. Bansal et al. [41] investigated the spider monkey algorithm (SMO) for multi-objective numerical optimization and metaheuristic problem-solving. The spider monkey survives in a group of 40–50 entities as a part of a fission–fusion society. Most animals live in a group of 40 to 50 individuals, and a mature female can control them, called a global leader. The global leader is responsible for food search. When it fails to discover enough food, it splits the chief group into several subgroups of 3–8 monkeys to find food sources independently. Every subgroup has a local group leader accountable for route discovery, planning, and decision-making. Spider Monkey's food discovery strategy includes four stages. The first stage deals with beginning food discovery and evaluating the distance of the home place from a food source. The second stage consists of communicating and updating position and distance information with the local leader, other members, and others. In the third stage, the regional leader decides the best position of the subgroup. If the position remains unchanged for the stipulated time, then all subgroup members decide to forage the food diversely independently. In the final stage, the global leader decides on the finest position of food source with the help of information gathered from local leaders. The global leader divides the groups into small groups if stagnation occurs for a particular group [42–44]. The flow chart of the proposed ISMO is shown in Fig. 3. The proposed ISMO uses the levy flight function to gradually increase the SMO position to increase the diversity of the solution and avoid the local minima problem.

A. Initialization Phase.

The initialization phase consists of the generation of *M* spider monkey. The initial population includes *N* subgroups for food foraging (finding optimal channel set),

where each subgroup encompasses *n* members to indicate optimal channel value. The spider monkey population is given by equation 34.

$$SM = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \vdots & C_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ C_{N1} & C_{N2} & \cdots & C_{Nn} \end{bmatrix} \tag{34}$$

B. Fitness Function

The fitness function of the proposed channel selection scheme depends upon interclass variability ($\sigma_{interclass}$) and intra-class variability ($\sigma_{intra-class}$) of the channels as given in equation 35.

$$Fitness = \frac{\sigma_{interclass}}{\sigma_{intraclass}} \tag{35}$$

The intra-class variability represents the closeness in same class channel information, whereas interclass variance depicts the distinctiveness of two classes. The intra-class and interclass variability is computed using equations 36 and 37.

$$\sigma_{interclass} = \frac{1}{Num \cdot CN \cdot CN} \sum_{i=1}^{Num} \sum_{j=i+1}^{Num} \sum_{k=1}^{CN} \sum_{l=1}^{CN} f(i_k, j_l) \tag{36}$$

$$\sigma_{interclass} = \frac{1}{Num \cdot CN \cdot CN} \sum_{i=1}^{NumClass} \sum_{k=1}^{CN} \sum_{l=1}^{CN} f(i_k, j_l) \tag{37}$$

Where $Num, CN, i, j, kandl$ represents the number of classes (2), number of total voice channels, samples of 1st class (standard), samples of 2nd class (disease), channels of fits class and channels of second class

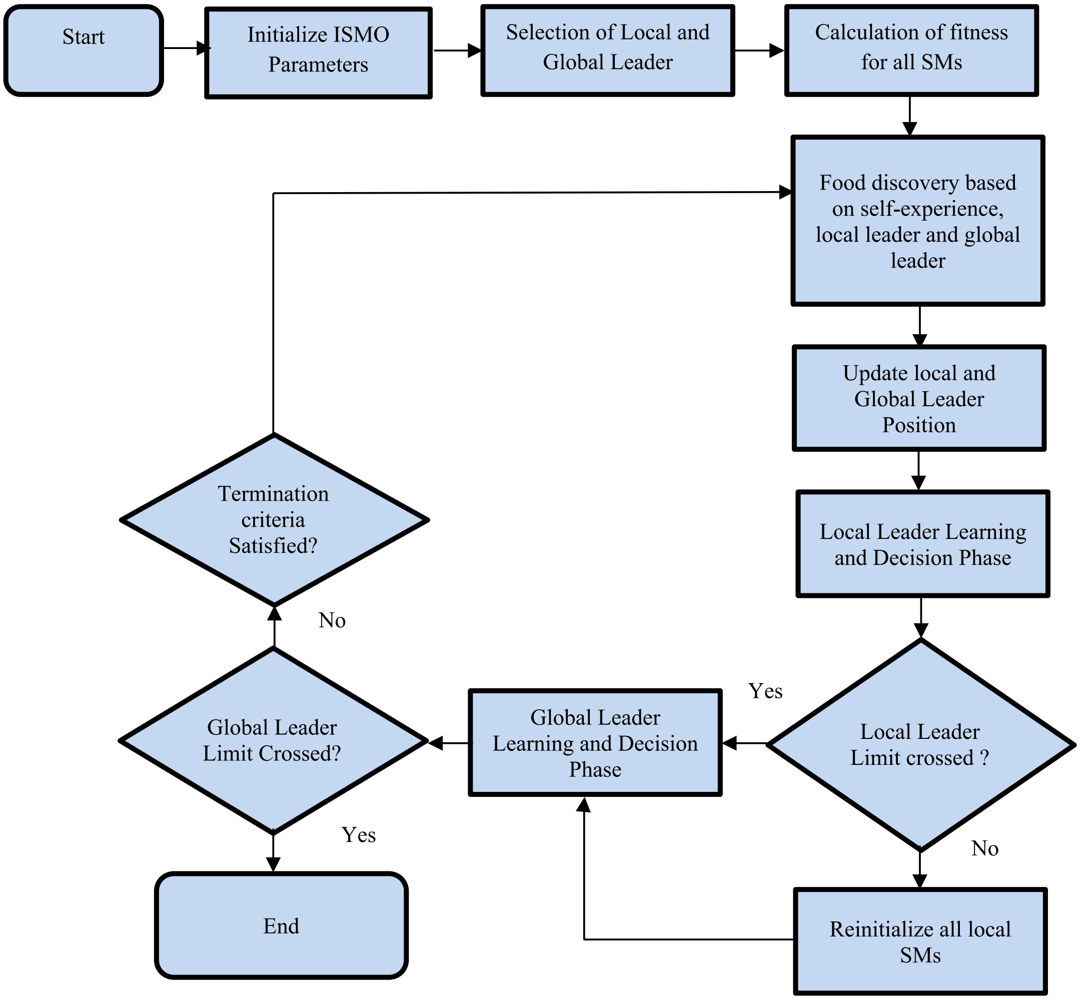

Start
Initialize ISMO Parameters
Selection of Local and Global Leader
Calculation of fitness for all SMs
Food discovery based on self-experience, local leader and global leader
Update local and Global Leader Position
Local Leader Learning and Decision Phase
Local Leader Limit crossed ?
Yes
No
Reinitialize all local SMs
Global Leader Learning and Decision Phase
Global Leader Limit Crossed?
Yes
No
Termination criteria Satisfied?
End

Fig. 3. Flow chart of proposed SMO for channel selection.

respectively. Also, the Euclidean distance metrics $f(x,y)$ measure the similarity between two channels, x, and y, having m samples in each channel as given by equation 38.

$$f(x,y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \tag{38}$$

C. Local Leader Phase.

The local group leader updates the positions of SMs using the experience of regional leaders and members. If the new SM's fitness is inferior to the previous local leader solution, then SM updates its position using equation 39.

$$sm_{newj} = sm_{ij} + R_n \times \left(LL_{kj} - sm_{ij}\right) + R_n \times \left(sm_{rj} - sm_{ij}\right) + levy \tag{39}$$

Where, sm_{newj}, indicates the updated position of SMs, sm_{ij} shows the old position of SM, LL_{kj} depicts the j^{th} position of k^{th} local group, and sm_{rj} provides j^{th} position of r^{th} SM, whose position in the group is randomly updated and R_n depicts the random number distributed uniformly between 0 and 1. The levy flight function is utilized to stepwise increase the SMO population, which boosts the solution diversity. The Levy function is given by equations 40 and 41. Here, β is levy's exponent which controls the shape and distribution, Γ is gamma function, and θ controls the step size.

$$Levy = 0.01\frac{rand1.0}{|rand|^{1/\beta}} \tag{40}$$

$$\theta = \frac{\Gamma(1+\beta).\sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right).\beta.2^{\left(\frac{\beta-1}{2}\right)}} \tag{41}$$

D. Global Leader Phase.

The global leader modifies the position of SMs and subgroups based on the experience of international and local leaders and probability fitness using equations 42 and 43.

$$sm_{newj} = sm_{ij} + R_n \times \left(GL_{kj} - sm_{ij}\right) + R_n \times \left(sm_{rj} - sm_{ij}\right) + levy \tag{42}$$

where GL_{kj} represents the *jth* dimension of *kth* global leader position.

$$Prob_i = \frac{fitness_i}{\sum_{i=0}^{N} fitness_i} \tag{43}$$

E. Global Leader Learning Phase.

The global leader updates the position of SM based on the best fitness value among all groups using a greedy selection scheme.

F. Local Leader Learning Phase.

The local leader updates the position of SM based on fitness value in individual local groups using a greedy selection scheme. However, the *SM* with the highest fitness value is chosen as the local leader.

G. Local Leader Decision Phase.

In case of stagnation, all individual SMs of local groups modify their position arbitrarily using decisions made by local and global leaders using equation 42.

$$sm_{newj} = sm_{ij} + R_n \times \left(GL_j - sm_{ij}\right) + R_n \times \left(sm_{ij} - LL_{kj}\right) \tag{44}$$

H. Global Leader Decision Phase.

The global leader divides the group into smaller groups when the global leader's position remains unchanged for a given time and maximum iterations. Suppose large groups are formed and deprived of the ability to modify the global leader's position. In that case, the global leader combines the entire group into a single group, known as the fission and fusion strategy. In cases of stagnation, the global leader splits the groups into smaller groups.

3.5. DCNN

The 1-D DCNN is utilized to learn the feature depiction of the 1-D features chosen using SMO. The 1-D DCNN framework consists of three convolution (Conv) layers, three batch normalization (BN), three rectified linear layers (ReLU), and a flattening layer (FL). However, 2-D DCNN consists of three Conv layers, three BN, three ReLU layers, three maximum pooling layers, and an FL layer. The output neurons of the FL layers of the 1-D DCNN and 2-D DCNN are concatenated together and given to a fully connected layer that improves the connectivity in the features depicted by 1-D DCNN and 2-D DCNN. The DCNN in both parallel arms uses 32, 64, and 128 Conv filters at the first, second, and third layers. The Conv layer provides local correlation and connectivity in the pathological voice features. The Conv layer offers hierarchical abstract-level features of voice that describe distinctive features that depict variation in voice due to disease or disorder. In this layer, the input signal is convolved with multiple convolution kernels to provide multi-level hierarchical features as given in equations 45 and 46.

$$V_{conv}(x) = Feat^{*}K \tag{45}$$

$$V_{conv}(x) = \sum_{j=1}^{Nf} Feat(i).K(x-i) \tag{46}$$

Here, Feat denotes voice features, K stands for convolution filter and V_{conv} denotes convolution output.

Batch normalization converts the deep features to the normalized format to minimize the outliners. It assists to fasten the training performance of DCNN. The BN operation for a batch size of b is given by equation 47. Here, μ_b and σ_b denote mean and variance over batch b, $\propto$ and β indicates scale and offset. The ReLU layer boosts the deep feature's non-linear characteristics to improve classification accuracy. It replaces the negative neurons with 0 values to minimize linearity. It helps to speed up the training process. The ReLU activation is provided in equation 48.

$$BN(x) = \propto . \frac{V_{conv}(x) - \mu_b}{\sigma_b} + \beta \tag{47}$$

$$ReLU(x) = max(BN(x), 0) \tag{48}$$

The maximum pooling layer in 2-D DCNN helps to select the crucial features from the 2 × 2 window. Selecting the maximum value over the window increases feature distinctiveness and minimizes the feature dimensions. The FL layer converts the multi-dimensional deep features to 1-D features vector. The FCL layer enhances the correction and connectivity of hierarchical deep features. The final layer consists of a probabilistic softmax classifier where a label with higher probability is classified as an output class. The suggested TWVFR combines the advantages of 1-D DCNN as well as 2-D DCNN to enhance the feature depiction of pathological voice.

4. Experimental results and Discussions

The suggested system uses MATLAB2020b on a personal computer with 20 GB RAM and a Windows operating environment. The system's training performance (training accuracy and loss) is shown in Fig. 4. The TWVFR system is trained using the ADAM optimization algorithm with an initial learning rate of 0.01, batch size of 32, 200 epochs, and cross-entropy loss function. It uses 70 % of data for training and 30 % for testing. It offers 100 % training accuracy for 200 iterations and shows stability in training. The proposed model needs 103.2 K trainable parameters and needs 1895 sec for the training.

The MFCCS provides the local changes in the pitch variation over the speech signal frame. It depicts the variation in the speech signal's time and frequency domain characteristics due to the voice signal. Fig. 5 illustrates the visualizations of the MFCCS for normal and diseased (cyst) voice.

The confusion matrix for the TWVFR and MVF-1D DCNN are shown in Fig. 6 for four class voice pathology. The TWVFR provides an overall accuracy of 98.3 % for cysts, 100 % for dysphonia, 98.3 % for paralysis, 96.7 % for normal class, and 98.33 % overall accuracy for four classes. Meanwhile, MVF-1D DCNN delivered 95 % for cysts, 95 % for dysphonia, 95 % for paralysis, 96.75 for normal, and 95.45 for all four classes.

The system outcomes are estimated using accuracy, recall, precision F1-score, selectivity, and NPV, as displayed in Figs. 7-12. The TWVFR-based ASSR's performance is analyzed for the different numbers of MVFs selected using the SMO algorithm. It provides better results for the 300 features than the total features and the MFCCS.

It offers precisions of 0.97, 0.98, 0.98, and 0.99 for the MFCCS-DCNN, MVF-DCNN, proposed method without SMO, and proposed method with SMO, respectively. The suggested method provides the improved recall of 0.97, 0.98, 0.99, and 0.99 for the MFCCS-DCNN, MVF-DCNN, proposed method without SMO, and proposed method with SMO, respectively. It shows the proper balancing between the F1-score of the system. The MVF helps characterize the different diseases' effects on the voice. The outcomes of the suggested schemes are observed to be superior to the MFCCS-DCNN and MVF-DCNN. Combining the 1D DCNN and 2D DCNN assists in capturing the short and long-term dependencies of the voice to represent the disease. Table 1 accurately compares the suggested scheme with various system implementations.

The convergence of the ISMO-based feature selection is compared with traditional techniques like particle swarm optimization (PSO), genetic algorithm (GA), and traditional SMO (SMO) as shown in Fig. 13. The proposed ISMO provides a superior solution compared with the GA, PSO, and SMO. The suggested TWVFR scheme offers an overall accuracy of 99.20 % for ISMO, 98.20 % for SMO, 97.50 % for PSO, and 96.80 % for GA. The optimal population split in the local leader and local group in ISMO ensures superior solution diversity, better convergence, and elite solutions than PSO, GA, and SMO.

The outcome of the suggested TWVFR scheme is compared with the traditional techniques, as given in Table 2. It provides 95.20 % accuracy for the MFCC spectrogram features and 2D DCNN for the four-class disease detection. The system provides an overall voice disease detection accuracy of 91.50 %, 83.30 %, and 82.80 % for the 1D DCNN for the SD, TD, and VQ features, respectively. Thus, it is observed that the SD, temporal, and VQ features are essential for the representation of speech. Combining two types of vice features significantly boosts the system's overall accuracy.

It delivers an overall accuracy of 93.60 % for the combination of SD and TD features, where 1D DCNN is used to improve these features' connectivity and correlation. The collaboration of TD and VQ features offers 88.90 % accuracy. The amalgamation of SD and VQ features gives an overall accuracy of 93.20 % for voice disease detection. The MVFs combine all three features to improve the voice's distinctiveness and characterize the impact of disease on the voice. This results in 95.42 %

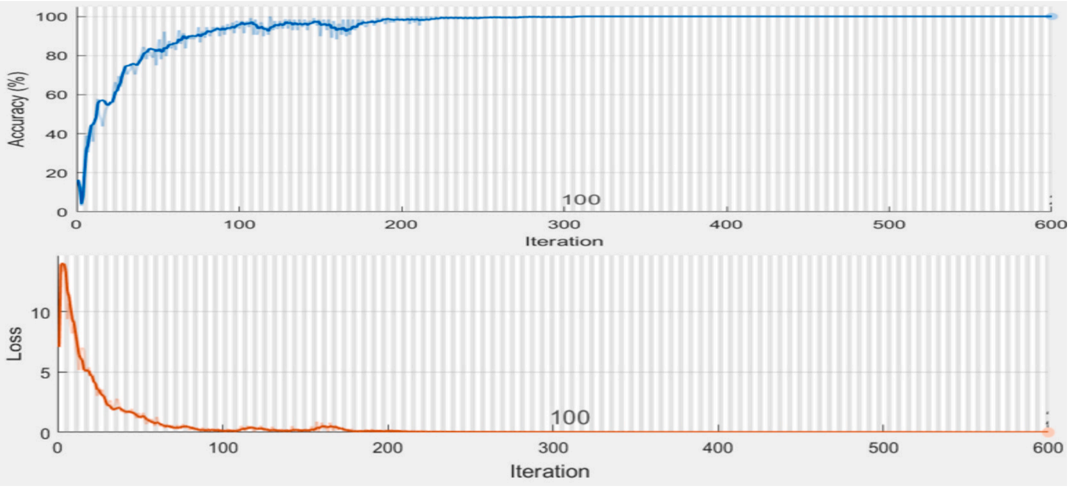
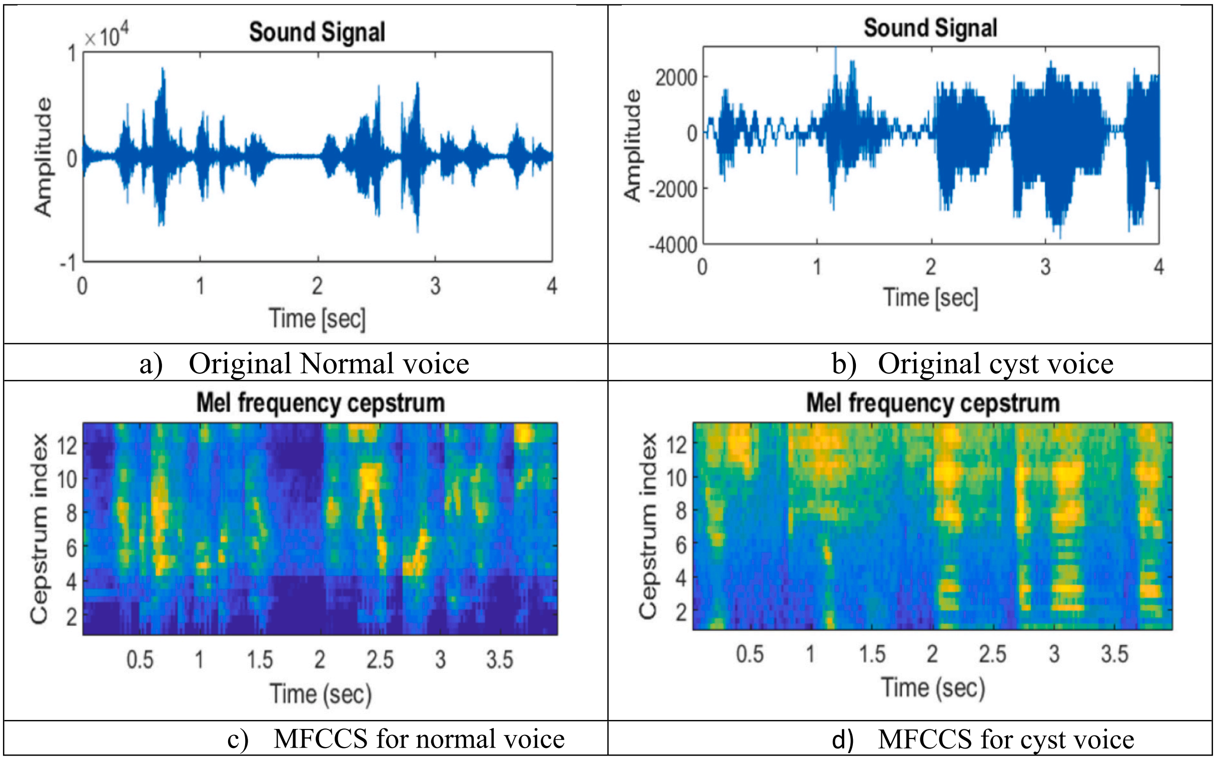


Fig. 4. Training performance of the proposed system.



a) Original Normal voice

b) Original cyst voice

c) MFCCS for normal voice

d) MFCCS for cyst voice

Fig. 5. MFCCS representation for sample normal and diseased voice.

accuracy for the 1D DCNN and shows a noteworthy boost in the accuracy provided by a single combination of two types of features. The 1D DCNN provides the correlation in the SD, temporal, and VQ characteristics of the voice and provides detailed information regarding the impact of disease on the local and global characteristics of the features. Further, the proposed TWVFR helps improve the voice's spatial and SD characteristics by adding MFFCS + 2D DCNN. The TWVFR provides an improved overall accuracy of 98.33 % for voice disease detection.

The proposed scheme's results are assessed for implementing existing algorithms such as LSTM, DCNN, Gated Recurrent Unit (GRU), Deep Belief Network (DBN), and Deep Neural Network (DNN) for two-class and four-class disease detection using voice as given in Table 3. The TWVFR provides an overall accuracy of 98.33 %, which has shown improvement over DCNN (94.60 %), LSTM (95.20 %), DBN (95.20 %), and GRU (96.56 %). The TWVFR offers % overall accuracy of 95.45 % for two-class disease detection, which has shown superiority over traditional techniques.

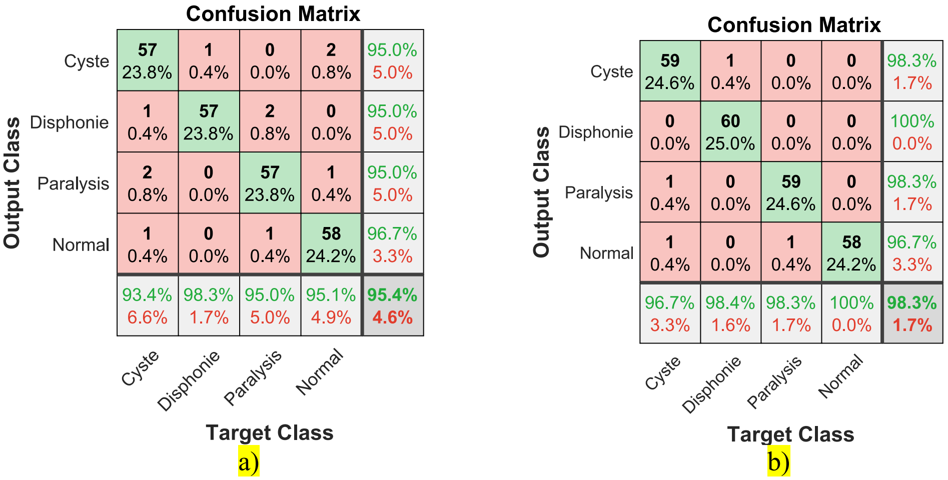


Fig. 6. Confusion matrix for 4-class disease detection a) MVF-1D DCNN b) TWVFR.

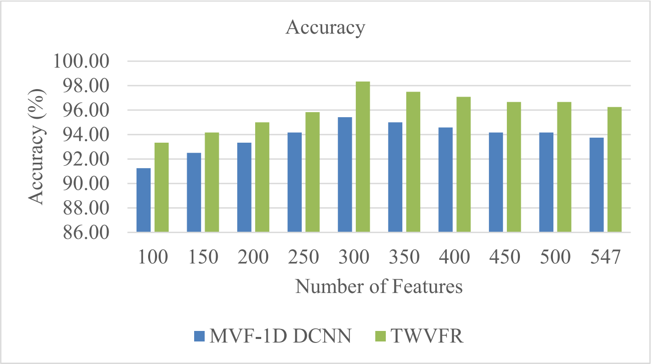


Fig. 7. Accuracy of MVF-1D DCNN and TWVFR for different features selected using SMO.

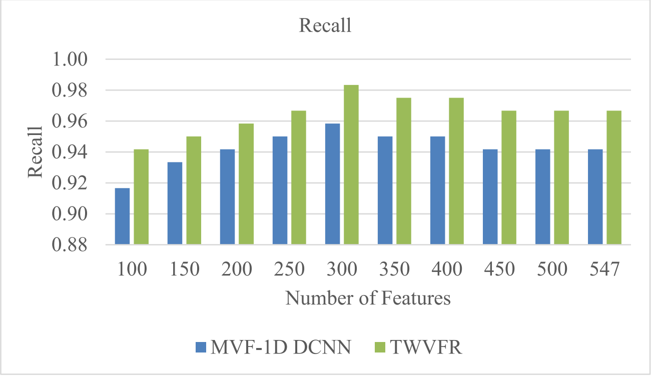


Fig. 8. Recall of MVF-1D DCNN and TWVFR for different features selected using SMO.

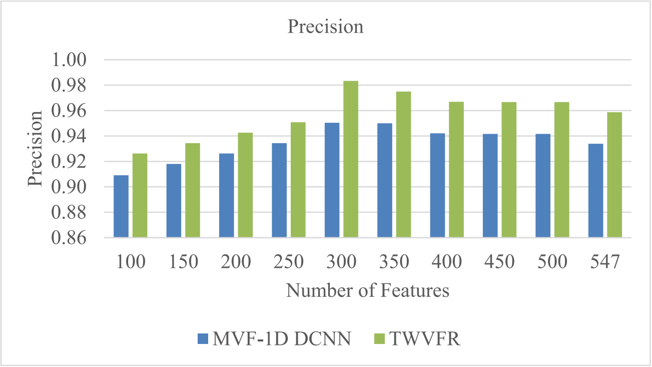


Fig. 9. Precision of MVF-1D DCNN and TWVFR for different features selected using SMO.

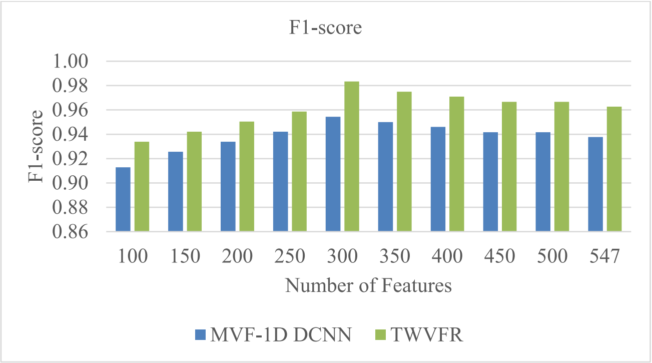


Fig. 10. F1 score of MVF-1D DCNN and TWVFR for different features selected using SMO.

5. Conclusions and Discussions

Thus, this article provides the DL-based voice pathology using TWVFR. The MFCCS-2D DCNN provides the SD and TD characteristics of the pathology voice. The MVF, along with 1D DCNN, provides the effect of disease on the voice regarding spectral changes, time domain changes, and VQ changes. The TWVFR scheme provides superior feature depiction compared with conventional ASSR schemes. The SMO-based feature selection algorithm selects the vital feature with maximum entropy, variance, and mean. It helps to minimize the time and total trainable parameters. The DCNN improves the distinctiveness of traditional features and provides a good correlation between local (frame level) and global-level features of the pathological signal. It captures the change over the speech's SD, TD, and VQ due to pathology. This results in an accuracy of 98.33 % for TWVFR and DCNN. In the future, the ASSR scheme's outcomes can be enhanced by optimizing the hyperparameters of the DCNN. The scheme can be extended for larger disease classes of disease detection for larger datasets. The effectiveness of the TWVFR system v = can be further enhanced by optimizing the hyper-parameters of the DL algorithm. The DL-based models are highly abstracted and act

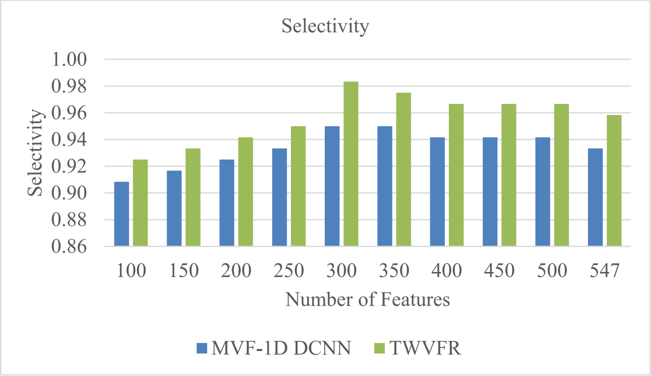


Fig. 11. Selectivity of MVF-1D DCNN and TWVFR for different features selected using SMO.

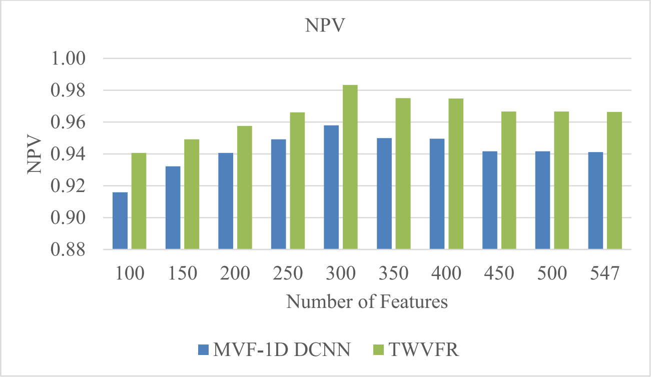


Fig. 12. NPV of MVF-1D DCNN and TWVFR for different features selected using SMO.

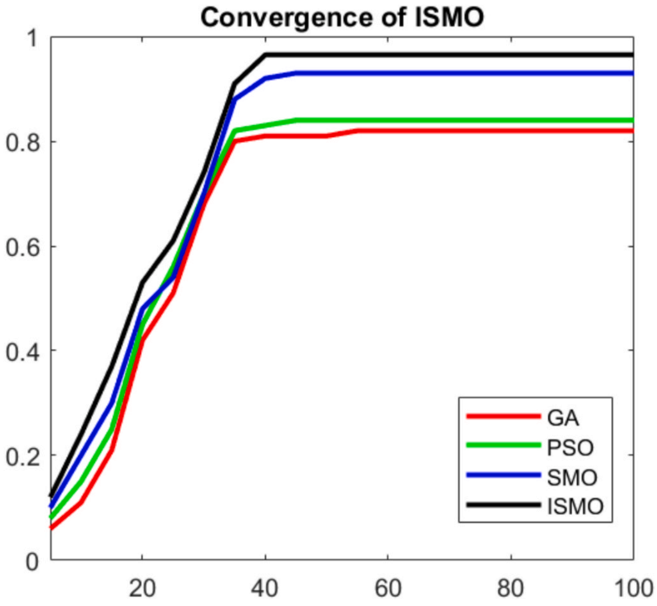


Fig. 13. Convergence comparison of ISMO with PSO, GA and SMO.

as a “black-box” system. Thus, in the future, the interpretability and explainability of the system can be enhanced by collaborating with explainable artificial intelligence (XAI) to boost the trust and reliability of the DL-based voice pathology system.

Table 2
Performance comparison with traditional methods.

Features	Classifier	Accuracy
MFCCS	2D DCNN	95.20 %
SD Features	1D DCNN	91.50 %
TD Features	1D DCNN	85.30 %
VQ Features	1D DCNN	82.80 %
SD + TD Features	1D DCNN	93.60 %
TD + VQ Features	1D DCNN	88.90 %
SD + VQ Features	1D DCNN	93.20 %
MVF (SD + TD + VQ) (300 features)	1D DCNN	95.42 %
TWVFR	1D DCNN and 2D DCNN	**98.33 %**

Table 3
Comparison of TWVFR with implementation of traditional algorithms.

Method	**Accuracy**	
	2-class	**4-class**
DCNN	96.20 %	94.60 %
LSTM	96.40 %	95.20 %
DBN	96.20 %	95.20 %
GRU	96.40 %	96.56 %
TWVFR	99.66 %	98.33 %

CRediT authorship contribution statement

Narendra Wagdarikar: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Sonal Jagtap:** Writing – review & editing, Validation, Supervision, Resources.

Funding

This research received no external funding.

Declaration of competing interest

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

Data availability

Data will be made available on request.

References

[1] FAl-Dhief, F.T.; Latiff, N.M.A.A.; Malik, N.N.N.A.; Salim, N.S.; Baki, M.M.; Albadr, M.A.A.; Mohammed, M.A. A Survey of Voice Pathology Surveillance Systems Based on Internet of Things and Machine Learning Algorithms. IEEE Access 2020, 8, 64514–64533.
[2] Titze IR, Martin DW. Principles of Voice Production; the Journal of the Acoustical Society of America. Acoust Soc Am 1998;104:1148.
[3] Teager H. Some observations on oral air flow during phonation. IEEE Trans Acoust Speech Signal Process 1980;28:599–601.
[4] Hillenbrand J, Houde RA. Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. J Speech Lang Hear Res 1996;39:311–21.
[5] Saenz-Lechon N, Godino-Llorente JI, Osma-Ruiz V, Gómez-Vilda P. Methodological issues in the development of automatic systems for voice pathology detection. Biomed Signal Process Control 2006;1:120–8.
[6] Markaki M, Stylianou Y. Using modulation spectra for voice pathology detection and classification September 2009;3–6:2514–7.
[7] Hossain MS, Muhammad G, Alamri A. Smart healthcare monitoring: A voice pathology detection paradigm for smart cities. Multimed Syst 2019;25:565–75.
[8] Mehta DD, Hillman RE. Voice assessment: Updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods. Curr Opin Otolaryngol Head Neck Surg 2008;16:211.
[9] Al-Nasheri A, Ali Z, Muhammad G, Alsulaiman M. Voice pathology detection using auto-correlation of different filters bank November 2014;10–13:50–5.

[10] Muhammad G, Alsulaiman M, Ali Z, Mesallam TA, Farahat M, Malki KH, et al. Voice pathology detection using interlaced derivative pattern on glottal source excitation. Biomed Signal Process Control 2017;31:156–64.

[11] Al-Nasheri A, Muhammad G, Alsulaiman M, Ali Z, Malki KH, Mesallam TA, et al. Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions. IEEE Access 2017;6:6961–74.

[12] Amami R, Smiti A. An incremental method combining density clustering and support vector machines for voice pathology detection. Comput Electr Eng 2017; 57:257–65.

[13] Muhammad G, Alhamid MF, Hossain MS, Almogren AS, Vasilakos AV. Enhanced living by assessing voice pathology using a co-occurrence matrix. Sensors 2017;17: 267.

[14] Michaelis D, Gramss T, Strube HW. Glottal-to-noise excitation ratio–a new measure for describing pathological voices. Acta Acust United Acust 1997;83:700–6.

[15] Saldanha JC, Ananthakrishna T, Pinto R. Vocal fold pathology assessment using mel-frequency cepstral coefficients and linear predictive cepstral coefficients features. J Med Imaging Health Inform 2014;4:168–73.

[16] Al-Nasheri A, Muhammad G, Alsulaiman M, Ali Z, Mesallam TA, Farahat M, et al. An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification. J Voice 2017;31: 113–e9.

[17] Mekyska J, Janousova E, Gomez-Vilda P, Smekal Z, Rektorova I, Eliasova I, et al. Robust and complex approach of pathological speech signal analysis. Neurocomputing 2015;167:94–111.

[18] Abd Ghani MK, Mohammed MA, Arunkumar N, Mostafa SA, Ibrahim DA, Abdullah MK, et al. Decision-level fusion scheme for nasopharyngeal carcinoma identification using machine learning techniques. Neural Comput Appl 2020;32: 625–38.

[19] Abdulkareem KH, Mohammed MA, Gunasekaran SS, Al-Mhiqani MN, Mutlag AA, Mostafa SA, et al. A Review of Fog Computing and Machine Learning: Concepts, Applications, Challenges, and Open Issues. IEEE Access 2019;7:153123–40.

[20] Mohammed, M.A.; Al-Khateeb, B.; Rashid, A.N.; Ibrahim, D.A.; Ghani, M.K.A.; Mostafa, S.A. Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images. Comput. Electr. Eng. 2018, 70, 871–882. [CrossRef].

[21] Barry, J.; Püutzer, M. Saarbrucken Voice Database, Institute of Phonetics, Univ. of Saarland. Available online: http://www.stimmdatenbank.coli.uni-saarland.de/ (accessed on 30 Sept 2023).

[22] Harar, P.; Alonso-Hernandezy, J.B.; Mekyska, J.; Galaz, Z.; Burget, R.; Smekal, Z. Voice pathology detection using deep learning: A preliminary study. In Proceedings of the 2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI), Funchal, Portugal, 10–12 July 2017; pp. 1–4.

[23] Mohammed MA, Ghani MKA, Hamed RI, Ibrahim DA, Abdullah MK. Artificial neural networks for automatic segmentation and identification of nasopharyngeal carcinoma. J Comput Sci 2017;21:263–74.

[24] Mohammed MA, Ghani MKA, Arunkumar NA, Hamed RI, Abdullah MK, Burhanuddin MA. A real time computer aided object detection of nasopharyngeal carcinoma using genetic algorithm and artificial neural network based on Haar feature fear. Future Gener Comput Syst 2018;89:539–47.

[25] Djenouri D, Laidi R, Djenouri Y, Balasingham I. Machine learning for smart building applications: Review and taxonomy. ACM Comput Surv (CSUR) 2019;52: 1–36.

[26] Alhussein M, Muhammad G. Voice pathology detection using deep learning on mobile healthcare framework. IEEE Access 2018;6:41034–41.

[27] Bhangale KB, Kothandaraman M. Survey of deep learning paradigms for speech processing. *Wirel Pers Commun* 2022;125(2):1913–49.

[28] Bhangale, Kishor, and K. Mohanaprasad. "Speech emotion recognition using mel frequency log spectrogram and deep convolutional neural network." In *Futuristic Communication and Network Technologies: Select Proceedings of VICFCNT 2020*, pp. 241-250. Springer Singapore, 2022.

[29] Bhangale, Kishor, Piyush Ingle, Rajani Kanase, and Divyashri Desale. "Multi-view multi-pose robust face recognition based on VGGNet." In *Second International Conference on Image Processing and Capsule Networks: ICIPCN 2021 2*, pp. 414-421. Springer International Publishing, 2022.

[30] Bhangale K, Kothandaraman M. Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network. *Electronics* 2023;12(4): 839.

[31] Karaman O, Çakın H, Alhudhaif A, Polat K. Robust automated Parkinson disease detection based on voice signals with transfer learning. *Expert Syst Appl* 2021;178: 115013.

[32] Tuncer T, Dogan S, Ertam F. 'Automatic voice based disease detection method using one dimensional local binary pattern feature extraction network,'. Appl Acoust, Dec. 2019;155:500–6.

[33] Tuncer T, Dogan S. 'Novel dynamic center based binary and ternary pattern network using m4 pooling for real world voice recognition,'. Appl Acoust, Dec. 2019;156:176–85.

[34] Tuncer T, Dogan S, Özyurt F, Belhaouari SB, Bensmail H. Novel multi center and threshold ternary pattern based method for disease detection method using voice. *IEEE Access* 2020;8:84532–40.

[35] Fujimura S, Kojima T, Okanoue Y, Shoji K, Inoue M, Omori K, et al. 'Classification of voice disorders using a one-dimensional convolutional neural network,'. J Voice, Mar 2020.

[36] I. Hammami, L. Salhi, and S. Labidi, ''Voice pathologies classification and detection using EMD-DWT analysis based on higher order statistic features,'' IRBM, Jan. 2020.

[37] AL-Dhief, Fahad Taha, Nurul Mu'azzah Abdul Latiff, Nik Noordini Nik Abd Malik, Naseer Sabri, Marina Mat Baki, Musatafa Abbas Abbood Albadr, Aymen Fadhil Abbas, Yaqdhan Mahmood Hussein, and Mazin Abed Mohammed. "Voice pathology detection using machine learning technique." In *2020 IEEE 5th international symposium on telecommunication technologies (ISTT)*, pp. 99-104. IEEE, 2020.

[38] Syed SA, Rashid M, Hussain S, Zahid H. Comparative analysis of CNN and RNN for voice pathology detection. *Biomed Res Int* 2021, 2021,:1–8.

[39] Lee J-N, Lee J-Y. An Efficient SMOTE-Based Deep Learning Model for Voice Pathology Detection. *Appl Sci* 2023;13(6):3571.

[40] Abdulmajeed NQ, Al-Khateeb B, Mohammed MA. Voice pathology identification system using a deep learning approach based on unique feature selection sets. *Expert Syst* 2023:e13327.

[41] Bansal, Jagdish Chand, Harish Sharma, Shimpi Singh Jadon, and Maurice Clerc. "Spider monkey optimization algorithm for numerical optimization." *Memetic computing* 6 (2014): 31-47.

[42] Patel VP, Rawat MK, Patel AS. Local neighbour spider monkey optimization algorithm for data clustering. *Evol Intel* 2023;16(1):133–51.

[43] Sharma, Apoorva, Nirmala Sharma, and Kavita Sharma. "Enhancing the social learning ability of spider monkey optimization algorithm." In *Proceedings of the International Conference on Intelligent Vision and Computing (ICIVC 2021)*, pp. 413-435. Cham: Springer International Publishing, 2022.

[44] Belaiche L, Kahloul L, Houimli M, Bousnane S, Benharzallah S. Multi-Swarm-based Parallel Spider Monkey Optimization Algorithm. In: In *2022 International Conference on Advanced Aspects of Software Engineering (ICAASE)*; 2022. p. 1–6.

[45] Tirronen S, Kadiri SR, Alku P. The effect of the MFCC frame length in automatic voice pathology detection. *J Voice* 2022.

[46] Javanmardi F, Kadiri SR, Alku P. A comparison of data augmentation methods in voice pathology detection. *Comput Speech Lang* 2024;83:101552.

[47] Kadiri SR, Alku P. Analysis and detection of pathological voice using glottal source features. *IEEE J Sel Top Signal Process* 2019;14(2):367–79.

[48] Kadiri, Sudarsana Reddy, Manila Kodali, and Paavo Alku. "Severity classification of Parkinson's disease from speech using single frequency filtering-based features." *arXiv preprint arXiv:2308.09042* (2023).

[49] Kadiri S, Kethireddy R, Alku P. Parkinson's disease detection from speech using single frequency filtering cepstral coefficients. In: *Interspeech*. International Speech Communication Association (ISCA); 2020. p. 4971–5.

[50] Kadiri SR, Alku P, Yegnanarayana B. Extraction and utilization of excitation information of speech: A review. *Proc IEEE* 2021;109(12):1920–41.