Mathematical Modelling and Numerical Simulation with Applications, 2024, 4(4), 469–494

https://dergipark.org.tr/en/pub/mmnsa

ISSN Online: 2791-8564 / Open Access

https://doi.org/10.53391/mmnsa.1473574

RESEARCH PAPER

# A machine learning approach for voice pathology detection using mode decomposition-based acoustic cepstral features

**Özkan Arslan** [1,*,‡]

[1]Department of Electronics and Communication Engineering, Tekirdağ Namık Kemal University, Çorlu, 59860 Tekirdağ, Türkiye

[*]Corresponding Author

[‡]oarslan@nku.edu.tr (Özkan Arslan)

## Abstract

In this paper, a mode decomposition analysis-based adaptive approach is proposed to provide high diagnostic performance for automated voice pathology detection systems. The aim of the study is to develop a reliable and effective system using adaptive cepstral domain features derived from the empirical mode decomposition (EMD), ensemble empirical mode decomposition (EEMD), and complete empirical mode decomposition with adaptive noise (CEEMDAN) methods. The descriptive feature sets are obtained by applying mel-frequency cepstral coefficients (MFCCs) and their derivatives, linear predictive coefficients (LPCs) and linear predictive cepstral coefficients (LPCCs) techniques to each decomposition level. The class-balanced data are generated on the VOice ICar fEDerico II database samples using the synthetic minority oversampling technique (SMOTE). The ReliefF algorithm is used to select the most effective and distinctive features. A combination of selected features and a support vector machine (SVM) classifier is used to identify pathological voices. In the pathology detection approach, the results show that the cepstral features based on EMD and SVM-cubic achieves the highest performance with 99.85% accuracy, 99.85% F1-score and 0.997 Matthews correlation coefficient (MCC). In pathology-type classification, the cepstral features based on EEMD and SVM-quadratic approach provided the highest performance with 96.49% accuracy, 96.46% F1 and 0.949 MCC values. The comprehensive results of this study reveal that mode decomposition-based approaches are more successful and effective than traditional methods for detection and classification of pathological voices.

**Keywords**: Voice pathology; SMOTE algorithm; mode decomposition; cepstral-domain coefficients; ReliefF algorithm; support vector machine

**AMS 2020 Classification**: 68T01; 68T07; 68T10

## 1 Introduction

Voice is considered as a subcomponent of speech, which is one of the most important daily communication tools of humans. In voice-related fields, the health problems have always been

➤ **Received:** 25.04.2024 ➤ **Revised:** 14.12.2024 ➤ **Accepted:** 18.12.2024 ➤ **Published:** 30.12.2024

seriously addressed. The social, professional, and interpersonal communication components can all be profoundly impacted by pathological voice issues [1]. Voice pathology, a global public health problem, has a high incidence and currently, it often includes vocal cysts, vocal fold nodules, keratosis, vocal folds paralysis, laryngitis and dysphonia [2]. The most prevalent voice issue, dysphonia, which affects 10% of the population, is frequently associated with changes in voice quality, pitch, and intensity in the upper respiratory tract [3, 4]. Phonetic symptoms or physiological irregularities serve as the basis for the pathological voice diagnostic in medicine. The following are some of the standard medical diagnosis techniques: Laryngoscopy, stroboscopic, and endoscopic procedures [5]. These diagnostic procedures are intrusive, time-consuming, and expensive, which means that they call for specialized tools and qualified medical professionals. Therefore, a non-invasive and efficient computer-aided automatic pathological voice identification system that does not require a clinical setting or a specialist would greatly improve human voice health. In addition, another important point is that it is possible to remotely evaluate the voice health and treatment process with a computer-aided diagnosis system. The advanced development of signal processing techniques and artificial intelligence (AI) has greatly contributed to automated and smart healthcare applications such as voice pathology detection (VPD).

The feature extraction and classification procedures have been widely used in the majority of VPD systems. In previous studies, feature extraction methods based on acoustic parameters include pitch [6], jitter and shimmer [7], harmonics to noise ratio [8], and cepstrum-based features have been introduced to assess the health status of voice. In VPD applications, mel-frequency cepstral coefficients (MFCC), linear prediction coefficients (LPC) and linear prediction cepstral coefficients (LPCC) techniques are frequently used as cepstrum-based acoustic parameters [9–12]. The exact type of vocal pathology can frequently be detected by using a classifier algorithm to the obtained acoustic distinctive features. The VPD framework has utilized a variety of classifiers. Gaussian mixture model (GMM) [13], Artificial neural network (ANN) [14], k-nearest neighbor (k-NN) [15], random forest (RF) [16], and support vector machine (SVM) [17] are effective and heavily used classifiers. In [13], the author used MFCC and its derivatives (delta and acceleration) features with hidden Markov model (HMM) and GMM classifiers to detect pathological voice. It has been reported that 94.44% and 95.74% accuracy were obtained for HMM and GMM, respectively. Chen et al. [15] proposed to use the LPCC features extracted from the voice signals separated by Hilbert-Huang transform with the KNN classifier. Thus, they achieved 93.3% accuracy with the LPCC-based HHT approach using optimal levels. A vocal disorders classification method was developed by Akbari and Arjmandi [18] utilizing energy and entropy features obtained by discrete wavelet packet transform (DWPT) combined with a multilayer neural network (ML-NN) and multiclass linear discriminant analysis (MC-LDA). They achieved classification accuracies of 96.67% and 97.33% for MC-LDA and ML-NN, respectively. Similarly, in [19], the authors presented a classification model using a combination of EMD and DWT decomposition based high order statistics features and SVM algorithm. It was reported that the VPD system had a 94.82% accuracy rate in the proposed study.

Recent studies have proposed deep learning algorithms as an alternative to traditional acoustic parameter and machine learning techniques in the automated identification of abnormal voices. In deep learning-based approaches, two-dimensional spectrogram images are used with convolutional neural networks (CNN) [20] and one-dimensional acoustic features are used with long-short term memory (LSTM) [21], which is one of the recurrent neural networks (RNN). In addition, CNN-LSTM architecture can be used to learn complex features obtained from spectrogram images [22]. The accuracy rate of these approaches in detecting pathological voice varies between 77% and 95.41%. In order for deep learning-based methods to achieve good performance, large amounts of data must be used. However, this may not always be possible as limited data can be collected

in the clinical conditions. These limitations encourage researchers to design more dynamic and high-performance systems by feature extraction engineering in pathological voice detection. Based on the above discussion, the aim of this paper is to obtain an effective feature set that enables automatic detection and classification of pathological voices. The main contributions of this study are summarized as follows:

- The study provided the development of a reliable and effective model for the detection of pathological voice and the classification of three types of pathology: hyperkinetic dysphonia, hypokinetic dysphonia, and reflux laryngitis.
- The proposed system extracts cepstral-domain acoustic features directly from the raw voice signals and from each level of the decomposed signals by the EMD, EEMD and CEEMDAN methods.
- The proposed system has been implemented with synthetic dataset obtained by the SMOTE algorithm, and the performance gain of the synthetic dataset has been revealed.
- The proposed system has been compared with state-of-the-art studies by many performance metrics and its contributions and achievements to the literature have been emphasized.

The rest of the paper is organized as follows. In Section 2, the proposed VPD system based on EMD, EEMD and CEEMDAN techniques is introduced. In Section 3, numerical results are presented to evaluate the detection and classification performance of the proposed VPD. In Section 4, comparisons with state-of-the-art studies are made and the performance gain of the proposed VPD is revealed. Finally, the paper is concluded in Section 5.

## 2 Materials and methods

In this study, a method for classifying healthy and diseased voices is given that is based on the cepstral-domain acoustic features extracted by EMD, EEMD, and CEEMDAN which are Hilbert-Huang transform-based algorithms. The block diagram of the proposed approach is illustrated in Figure 1 and is organized the following steps: (1) The healthy and pathological voice signals are obtained by the publicly available VOICED dataset. (2) The signals are normalized with the z-score method and decomposed by the EMD, EEMD and CEEMDAN methods. (3) The intrinsic mode functions (IMFs) of the decomposed signals as well as the raw signals are used to extract the MFCCs, its derivatives (delta and acceleration), LPCs, and LPCCs cepstral-domain features. (4) Then, a synthetic dataset is generated by increasing the amount of data with the SMOTE technique and the ReliefF feature selection algorithm is utilized to get the more limited and effective feature set. (5) The reduced selected feature set is supported as input to SVM classifier models. (6) The results of classification models are evaluated with a large number of performance metrics and verification is performed.

### Database

The acoustic voice records of healthy and diseased individuals were acquired for this research derived from the publicly available VOICED (VOice ICar fEDerico II) dataset [23] provided by the PhysioNet organization. The VOICED dataset contains a total of 208 (150 pathological and 58 healthy) voice data recorded from 73 male and 135 female subjects. The pathological voices include commonly encountered disorders of hyperkinetic dysphonia, hypokinetic dysphonia and reflux laryngitis. Validation of healthy and pathological voices including vocal fold disorders was performed by medical professionals. Table 1 gives the distribution of healthy and diseased voices in the dataset by age and gender. In a room with a moderate amount of humidity and background noise (less than 30 dB), the vowel /a/ was continuously recorded for five seconds by subjects ranging in age from 18 to 70.

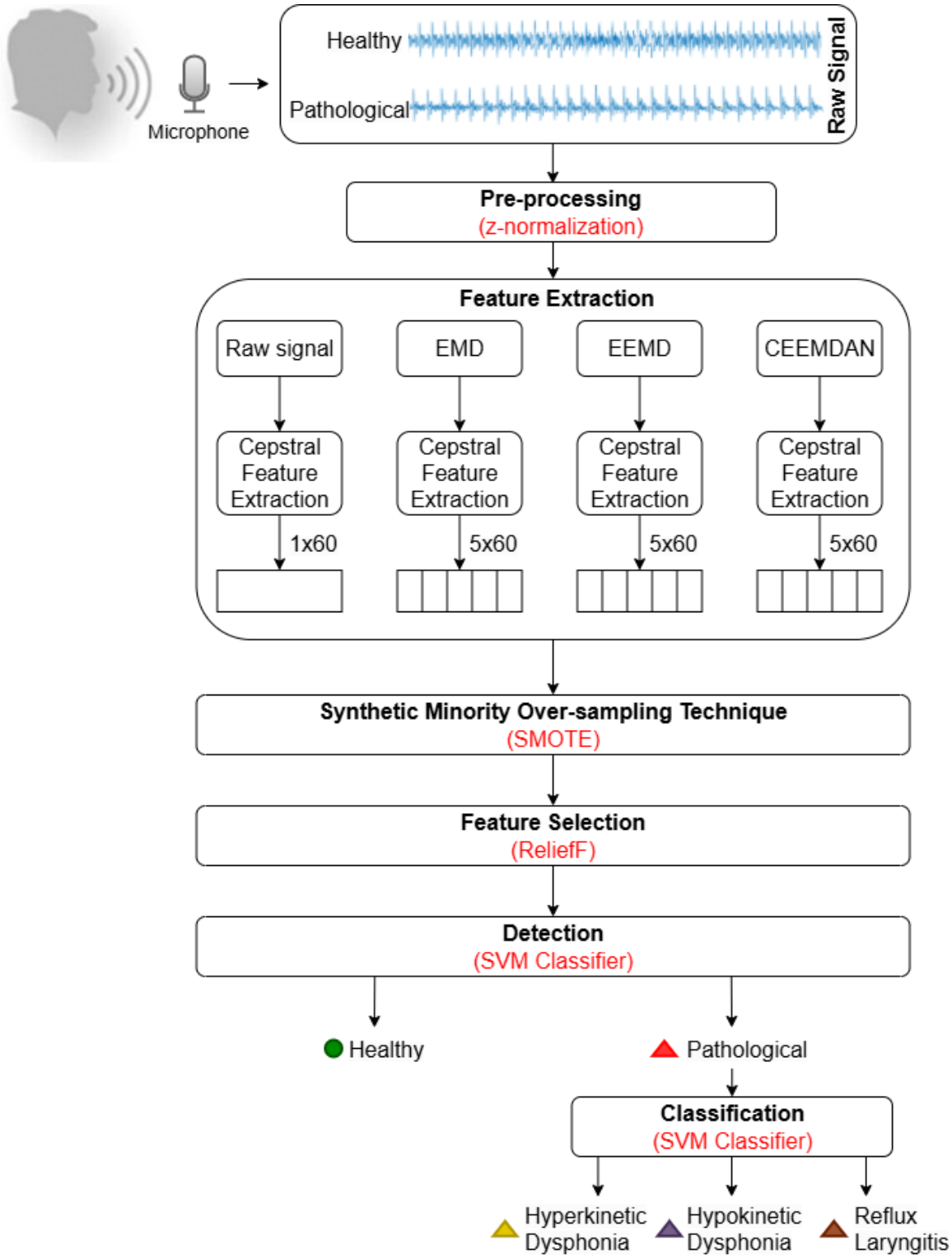


**Figure 1.** Block diagram of the proposed voice pathology detection and classification framework

The voice data have been collected by 8000 Hz sampling rate and 32-bit resolution using a microphone held at a 45-degree angle at a distance of about 20 cm.

**Table 1.** Distribution of pathology and healthy voices in VOICED database

| | | Age (years) | Female | Male | Total |
|---|---|---|---|---|---|
| Pathologies | # of hyperkinetic voices | 18-34 | 12 | 7 | 19 |
| | | 35-49 | 16 | 7 | 23 |
| | | ≥50 | 21 | 9 | 30 |
| | # of hypokinetic voices | 18-34 | 9 | 2 | 11 |
| | | 35-49 | 10 | 2 | 12 |
| | | ≥50 | 13 | 5 | 18 |
| | # of reflux laryngitis voices | 18-34 | 1 | 2 | 3 |
| | | 35-49 | 9 | 8 | 17 |
| | | ≥50 | 9 | 9 | 18 |
| | Total | | 100 | 51 | 151 |
| Healthy | # of healthy voices | 18-34 | 21 | 7 | 28 |
| | | 35-49 | 9 | 8 | 17 |
| | | ≥50 | 6 | 6 | 12 |
| | Total | | 36 | 21 | 57 |

## Data pre-processing

The voice recordings were made by using the microphone of a mobile device It was done in a quiet and not too dry room using a system that can obtain the voice signal in real time using its microphone. However, each recording was filtered with a Hanning windowed low-pass FIR filter to remove any noise accidentally added during the acquisition. The participants were instructed to pronounce the audio sample with a constant sound intensity as they would during normal conversation. Specific training tests were performed for each subject approximately two/three times before enrollment, and then the collected data were stored anonymously.

Although the same environment and microphone were used in the acquisition of all sound recordings, the sound amplitudes may differ between subjects. The amplitude of the features extracted from the voice signals is greatly affected by these differences. Therefore, all voice signals are normalized to ensure that the extracted features are not affected by the amplitude change and are more meaningful. In this study, the amplitudes of the voice signals were normalized using the z-normalization method. If the voice data is $X$, the z-normalized data $X_{norm}$ is expressed as: $X_{norm} = \frac{X-\mu}{\sigma}$, where $\mu$ and $\sigma$ denote the mean and standard deviation of all voice data, respectively.

## Mode decomposition methods

Empirical mode decomposition (EMD), a method used for analyzing of non-linear and non-stationary signals, was proposed by Huang et al. [24] in the late 1990s. In contrast to signal analysis methods such as wavelet and Fourier transform, EMD is an intuitive, direct, and adaptive method that uses a data-driven and data-derived basis function. Due to its adaptive nature, the EMD method is extensively used in sound/acoustic signal processing applications such as the classification of heart valve disorder [25], and the detection of disease voice signals [26]. The signal is divided into intrinsic mode functions (IMFs) by the EMD technique, each of which has a different frequency component. The following two conditions must be ensured for obtaining IMFs: (1) there can only be one difference or an equal number of extrema and zero crossings in the dataset, and (2) the average value of the envelope generated by the local maximums and minimums in the whole dataset is zero. Thus, the EMD process, also known as the sifting process, consists of obtaining all functions classified as IMFs. The steps involving the basic operations of EMD are given in Algorithm 1.

**Algorithm 1** Process steps for the EMD method

1: **Input:** Original signal $x(t)$;
2: Identify all local maxima and minima extrema for $x(t)$;
3: Calculate upper and lower envelope of $x(t)$;
4: Calculate the mean of both envelopes, $m_1(t)$;
5: Subtracting the mean from the original signal, difference signal $h_1(t) = x(t) - m_1(t)$;
6: It is evaluated whether $h_1(t)$ satisfies two IMF condition or SD < 0.3;
7: If $h_1(t)$ is not provide IMF conditions, update the signal and continue the steps 2 to 5;
8: The residue signal is obtained, $r_1(t) = x(t) - IMF_1$;
9: Iterate steps 2 to 8 on residue signal becomes a monotonic function;
10: **Output:** Find all the IMFs of the signal and residue signal sequence, $IMF_i$ and $r_i$ for $i = 1, 2, \ldots, k$;

The original signal $x(t)$ with IMFs and the residual signal can be defined as:

$$x(t) = \sum_{i=1}^{k} IMF_i(t) + r_k(t). \tag{1}$$

The stopping criterion (SD) is calculated to complete the sifting process in $T$ steps and can be defined as:

$$SD_i = \sum_{t=0}^{T} \frac{|IMF_{i+1}(t) - IMF_i(t)|^2}{IMF_i(t)^2}. \tag{2}$$

A mode mixing problem can arise when many IMFs include signals of the same scale or multiple IMFs contain signals of very different scales. In order to address the scale decomposition problem, ensemble empirical mode decomposition (EEMD), a method of noise-assisted data analysis, has been proposed [27]. The IMF components are described by the EEMD as the average of an ensemble of white noise additive signals of limited amplitude. Thus, the $i^{th}$ trial version of the signal $x(t)$ with added white noise can be expressed as:

$$x^i(t) = x(t) + a_0 w^i(t), \tag{3}$$

where $w^i(t)$ denotes the white noise in $i^{th}$ trial and $a_0$ represents the amplitude. The $IMF_k^i$ is calculated with different realizations of white noise, and the average $k$-th $\overline{IMF}_k$ is expressed as:

$$\overline{IMF}_k = \frac{1}{L} \sum_{i=1}^{L} IMF_k^i. \tag{4}$$

The EEMD method basically includes the following concepts: (1) The collection of white noise added to the signal cancels each other out with the help of the ensemble average. Thus, there can only be one signal component in the mixing of signal and white noise. (2) To search for all possible solutions, finite-amplitude white noise needs to be summed with the signal. (3) It is necessary to add noise to the signal to obtain real and physically meaningful IMFs compared to the EMD method.

The EEMD adds white noise to the signal in order to address the mode mixing issue of the EMD

algorithm. In this case, the noise cannot be completely separated from the signal, causing the IMFs obtained by the EEMD to contain both noise and signal. Therefore, the complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) algorithm, which provides the spectral separation of modes containing noisy signals, is proposed to solve this problem [28]. The CEEMDAN algorithm provides a low-cost and efficient computation to reconstruct the original signal. As a result of using the CEEMDAN approach, the first residual signal can be calculated as:

$$r_1(t) = x(t) - \overline{IMF}_1, \tag{5}$$

where $IMF_1$ is the first average function calculated by EEMD, and the second average $\overline{IMF}$ is obtained as:

$$\overline{IMF} = \frac{1}{L}\sum_{i=1}^{L} E_1(r_1(t) + a_0 E_1(w^i(t))). \tag{6}$$

Finally, $r_k$ denotes the final residual signal, and the $k+1$ average IMF can be calculated as:

$$\overline{IMF}_{k+1} = \frac{1}{L}\sum_{i=1}^{L} E_1(r_k(t) + a_k E_k(w^i(t))), \text{ for } k = 2, 3, \ldots, K, \tag{7}$$

where $E_k(\cdot)$ represents the operator that enables the $k$-th IMF to be obtained by the EMD, and $a_k$ denotes the amplitude that allows the selection of the SNR.
Figure 2 illustrates an instance of subtracting 5-level IMFs from the pathological and healthy voice signal and the frequency components of these signals relative to the IMFs. As can be seen from Figure 2, the frequency components for pathological and healthy voice signals differ according to the modes. Mode #1 contains the dominant frequency for both pathological and healthy voice signals, while the frequency bandwidth of pathological signals is greater than for healthy signals. Thus, it is seen that the frequency difference in the modes can be used effectively in the classification of voice signals.

## Feature extraction

In traditional voice analysis and processing, two techniques are widely used for acoustic information extraction. The first technique uses a parametric modeling approach that has been developed to closely resemble the resonance structure of the human vocal tract and is based on linear predictive coding (LPC) and linear predictive cepstral coefficients (LPCC). The second technique is approaches based on Mel-frequency cepstral coefficients (MFCC) and their derivatives parameterized by the windows of the voice signal. The voice signal has to be preprocessed,framed, and windowed in order to extract the cepstral domain-based parameters. Therefore, the following steps are performed before the acoustic parameters are extracted.
*1) Pre-emphasis:* The digitized voice signal s(n) is flattened spectrally using a digital technique called pre-emphasis, which also makes it less susceptible to finite precision effects. The pre-emphasis system output $\hat{s}(n)$ is calculated as:

$$\hat{s}(n) = s(n) - \alpha s(n-1), \tag{8}$$

where $\alpha$ is the pre-emphasis coefficient, and the value of 0.95 is commonly used.
*2) Frame-blocking:* The voice signals are analyzed and examined in short-term frames due to their

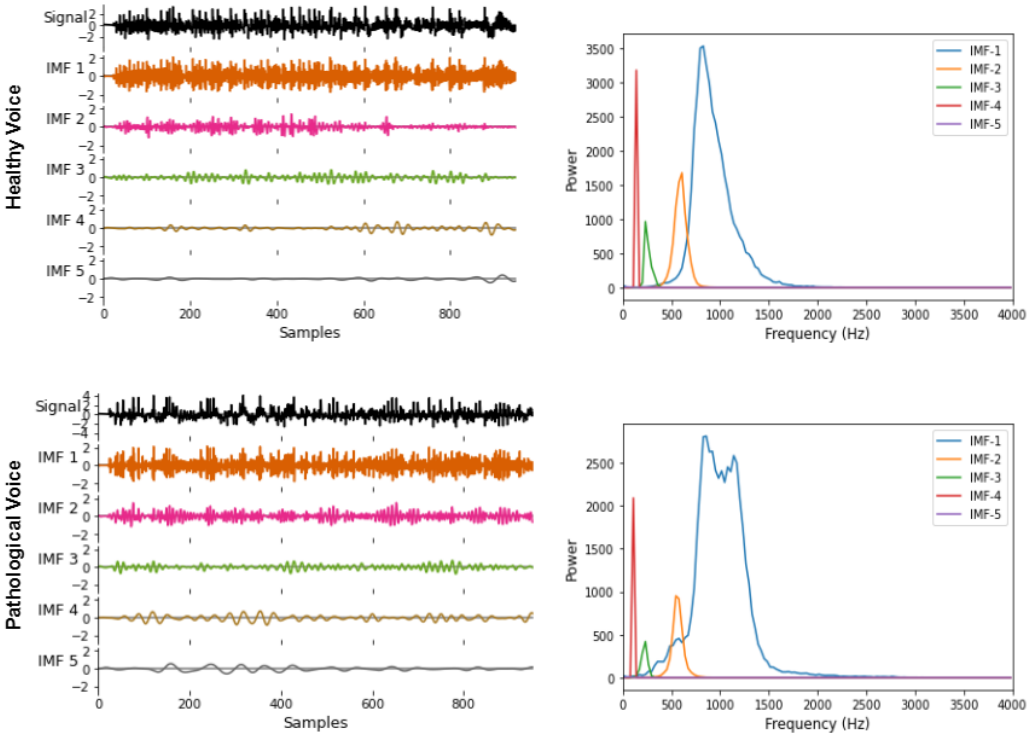


**Figure 2.** Intrinsic mode functions (IMFs) and Hilbert spectrum of healthy and pathological voice signals

time-varying nature. Thus, analysis frame-blocks are obtained that allow the parameters to be modeled dynamically. The pre-emphasized voice signal framed is expressed as:

$$x_l(n) = \hat{s}(Ml + n), \quad n = 0, 1, \ldots, N-1, \quad l = 0, 1, \ldots, L-1, \tag{9}$$

where, $M$ and $N$ is the number of adjacent frame samples and frame blocks, respectively. Also, $L$ represents the total number of frames.

*3) Windowing:* A windowing operation is applied to each frame to minimize signal discontinuities at the start and finish of the frames. The windowed signal $\hat{x}_l(n)$ is defined as:

$$\hat{x}_l(n) = x_l(n)w(n), \quad 0 \leq n \leq N-1, \tag{10}$$

where $w(n)$ denotes the window function. The Hamming window function is used for this study and is expressed as:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1. \tag{11}$$

The block diagram representation of the calculation steps of the LPC and LPCC parameters is illustrated in Figure 3. The LPC is obtained by autocorrelation of a frame with window function applied. Then, LPCC is obtained by LPC parameter conversion.

As a linear combination of prior voice samples, a given voice signal $s(n)$ can be estimated as follows:

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p), \tag{12}$$

where $a_1, a_2, \ldots, a_p$ are constants used for the analysis frame of voice signals.

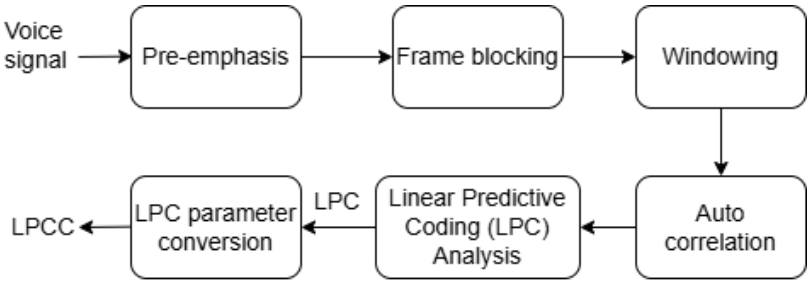


**Figure 3.** Block diagram of the LPC and LPCC computation process

The mel-frequency cepstral coefficient (MFCC) parameters are frequently utilized in voice and speech processing applications [29–31]. Figure 4 illustrates the process for MFCC feature extraction from a voice signal. The MFCC technique is developed on the basis that the human auditory system is more sensitive to low than high frequencies. In order to analyze non-stationary signals such as voice, it is crucial to consider the energy of the frequency bands. Acoustic signals can be easily and effectively analyzed in different frequency bands by using filter banks. In filter bank analysis, the logarithm process is applied to the energy coefficients to increase the dynamic range. After extracting the log-filterbank features, the MFCC coefficients of the signals are obtained using the Discrete Cosine Transform (DCT).

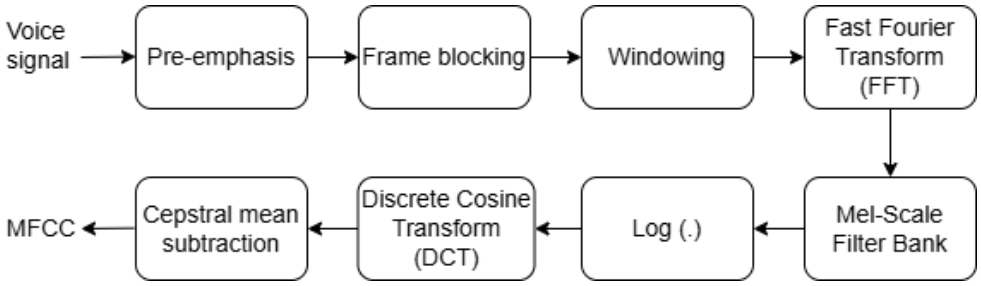


**Figure 4.** Block diagram of MFCC feature extraction from voice signal

The mel is expressed as a unit of pitch period and perceptually equidistant sounds are divided into equal numbers of mel's. A set of filters that extract energy from each frequency band is used to calculate the MFCC, and certain frequencies in Hz are calculated as follows:

$$Mel(f) = 2595 * \log_{10}(1 + \frac{f}{700}). \tag{13}$$

Spectral coefficients are calculated using the fast Fourier transform (FFT) after applying pre-emphasis, frame-blocking, and Hamming windowing. The calculation and derivation of the cepstral coefficients are formulated as follows:

$$c(n) = \sum_{m=0}^{M-1} \log_{10}(s(m)) \cos(\frac{\pi n(m-0.5)}{M}), \quad n = 0, 1, \ldots, C-1, \tag{14}$$

where $s(m)$, $c(n)$ and $M$ denote the mel spectrum, cepstral, and total number of filters, respectively. $C$ is the number of MFCCs, and the first 13 coefficients are generally used in voice/speech processing applications.

In this study, raw voice signals are decomposed into IMF modes by the EMD, EEMD, and

CEEMDAN methods. Considering the Hilbert spectrum given in Figure 2, in which the spectral components are obtained according to the modes, the first 5 IMF modes are used for feature extraction from voice signals, since they contain distinctive frequency components. The raw signals are processed to extract the features that will be employed in the classification of the voice signals, and the decomposed IMFs of each mode are also achieved. A feature vector of the MFCCs, velocity (first-order derivative, Δ-MFCC), acceleration (second-order derivative, ΔΔ-MFCC), LPC, and LPCC is obtained from the raw signal and each IMFs. A total of 60 coefficients (13 MFCCs, 12 Δ-MFCCs, 11 ΔΔ-MFCCs, 12-LPCs, and 12-LPCCs) are obtained from signal and each mode. Thus, the classification of voice signal is performed using both coefficients and derivatives, and mode frequency component differences. The Hilbert spectrum shows that the pathological signal has a wider frequency band than the healthy signal. Therefore, the cepstral coefficients obtained from each filter bank are of great importance in distinguishing the signals. Table 2 shows a description and number of the features that were derived by the raw signal, EMD, EEMD, and CEEMDAN.

**Table 2.** Description and number of obtained features.

| Features | Description | Raw | EMD | EEMD | CEEMDAN |
|---|---|---|---|---|---|
| MFCCs | Cepstral coeffs of raw signal and all IMFs | 13 | 65 | 65 | 65 |
| Δ-MFCCs | First-order derivative (velocity) of MFCCs | 12 | 60 | 60 | 60 |
| ΔΔ-MFCCs | Second-order derivative (acceleration) of MFCCs | 11 | 55 | 55 | 55 |
| LPCs | Predictive coding coeffs for raw signal and all IMFs | 12 | 60 | 60 | 60 |
| LPCCs | Cepstral coeffs of LPC for raw signal and all IMFs | 12 | 60 | 60 | 60 |
| Total | | 60 | 300 | 300 | 300 |

## Synthetic data augmentation using SMOTE algorithm

In recent years, with the progress and developments in the field of artificial intelligence, big data classification approaches have provided a great advantage for diagnostic research in medicine. Medical data is often inconsistent due to different conditions and difficulties in collecting samples in clinical conditions. The class imbalance in the dataset is a serious obstacle that negatively affects the classification performance. Therefore, the Synthetic Minority Over-sampling Technique (SMOTE) algorithm based on the principle of generating random sample points can be used to improve the imbalance rate [32]. The SMOTE algorithm provides new data by randomly interpolating between a number of samples and their neighbors' samples. The performance of the classification is improved by increasing the rate of data imbalance by producing a given number of artificial samples. The SMOTE algorithm realization procedure can be summarized as follows: (1) For each minority sample $x_i (i = 1, 2, ..., n)$, the distance of the minority sample to the other samples is calculated and its $k$ nearest neighbors are obtained. (2) As a subset, the $m$ nearest neighbors are randomly selected from the set of $k$ nearest neighbors of each sample $x_i$ and denoted as $x_{ij} (j = 1, 2, ..., m)$. Thus, artificially obtained minority samples $p_{ij}$ are expressed as:

$$p_{ij} = x_i + rand(0, 1) \times (x_{ij} - x_i), \tag{15}$$

where $rand(0, 1)$ is a random number generator that is uniformly distributed in the range of [0,1]. The SMOTE method is one of the most popular over-sampling methods and has been widely adopted in many applications. Moreover, many SMOTE extensions have been developed for the generation of synthetic data: k-NN SMOTE, Borderline SMOTE, and Adaptive Synthetic Sampling (ADASYN). Figure 5 shows the distributions of the original dataset samples and the synthetically resampled datasets created with SMOTE methods. Data samples expanded with the SMOTE

technique have a higher degree of intra-class clustering, and the degree of intra-class clustering is significantly improved compared to other over-sampling techniques.

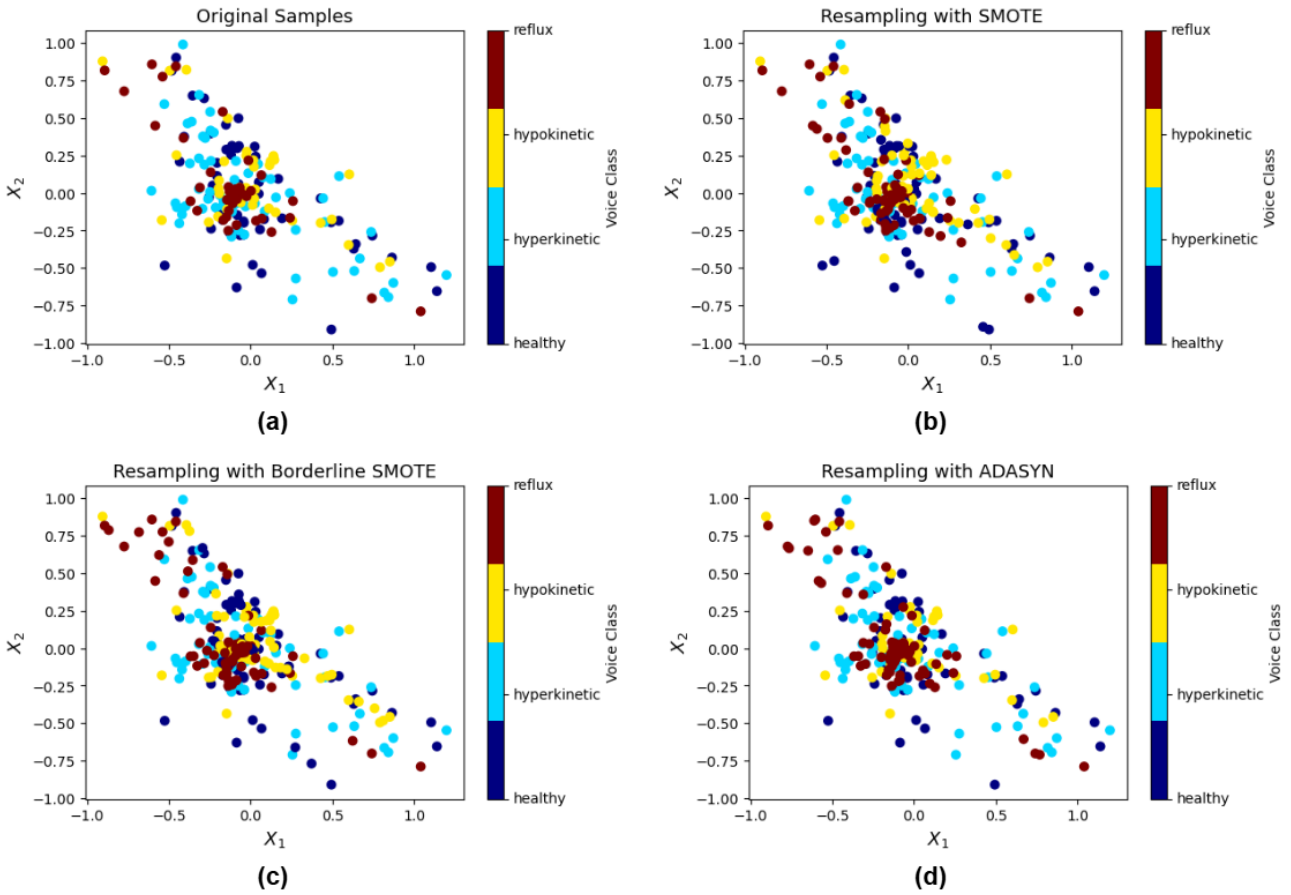


**Figure 5.** Illustration of over-sampling methods for VOICED dataset, (a) original samples, (b) resampling with SMOTE, (c) resampling with Borderline SMOTE, (d) resampling with ADASYN

In this study, the SMOTE algorithm was applied to features of voice signals to generate balanced-data classes. The $k$ nearest neighbor values for healthy and pathological voice data in class-balanced were set as 5 and 1.265, respectively. Thus, the number of healthy and pathological voice data was increased from 57 and 151 to 342, respectively. In addition, data numbers for each of the hyperkinetic dysphonia, hypokinetic dysphonia, and reflux laryngitis pathology types were increased from 72, 41, and 38 to 114, respectively. Hence, class-balanced data were produced by equating the amount of data for the pathology types. The training and testing process for all classifiers was carried out on the new minority artificial dataset.

## Feature selection using ReliefF algorithm

Feature selection algorithms are crucial to the machine learning process since it is vital to choose the most distinctive features in order to produce high-performance classification models. The execution time required for the models can be shortened by using feature selection methods. In this study, ReliefF feature selection algorithm proposed by Kira and Rendell [33] was utilized to choose the most effective features. ReliefF is a feature selection technique that gives each feature in the dataset a weight that may be adjusted gradually. This algorithm ensures that the significant features have high weights [34]. Thus, in this study, the reduced selected feature set was obtained with the ReliefF algorithm using the 10 nearest neighbors.

## Support vector machines

The voice/speech signals can be categorized using a variety of machine-learning techniques. Among these techniques, it has been proven that the support vector machine (SVM) classification algorithm provides high and effective performance in many studies. The SVM method developed by Vapnik [35] is a supervised approach and is widely used in classification. This method operates by creating the boundary at which the groups are divided by an ideal hyperplane [36]. An ideal hyperplane that satisfies the maximum margin conditions is selected in order to maximize the distance between the closest data points on either side of the plane. Hyperplanes, also called support vectors, are determined for each class. The kernel of an SVM algorithm is an ensemble of mathematical operations that accepts data as input and transforms it into the necessary forms.

The maximum margin and optimal hyperplane for a two-class SVM algorithm are shown in Figure 6. In the SVM approach, data separation is usually performed with a linear kernel function. Commonly used non-linear kernel functions for data that cannot be separated linearly are given in Table 3. In the SVM model, penalty parameter $C$ {1, 10, 100, and 1000}, kernel functions {linear, cubic and quadratic} and gamma parameter {0.1 to 0.9 at intervals of 0.1} were selected by the grid-search algorithm. The optimal values for $C$ and gamma parameters were set to 10 and 0.1, respectively. Also, the kernel scale and box constraint were set to 1.

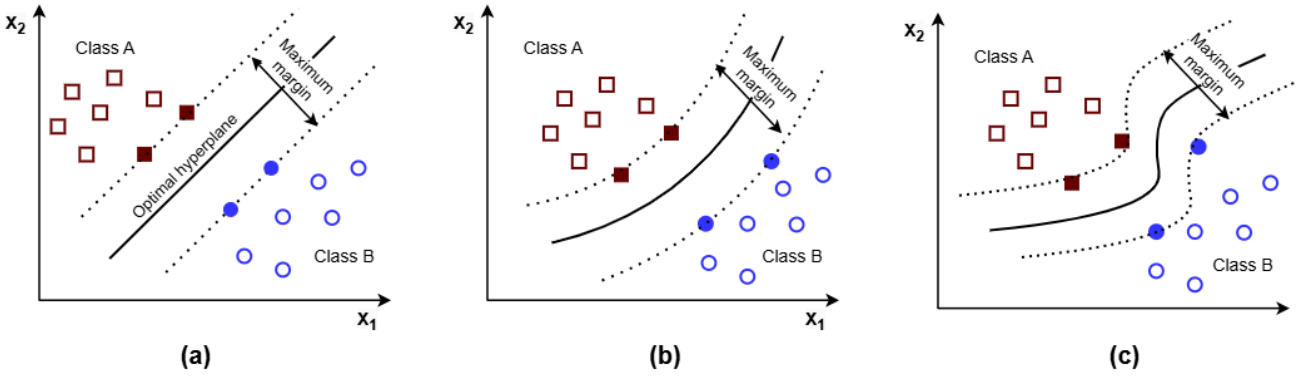


**Figure 6.** General structure of SVM algorithm for kernel functions (a) linear, (b) quadratic, (c) cubic

**Table 3.** Linear and polynomial kernel functions for SVM classifier

| Kernel Function | Mathematical Expression | Description |
|---|---|---|
| Linear | $k(x_1, x_2) = x_1 \cdot x_2$ | Decision boundary is linear |
| Polynomial | $k(x_1, x_2) = (x_1 \cdot x_2 + 1)^d$ | $d$ is the degree of the polynomial |
| Quadratic Polynomial | $k(x_1, x_2) = (x_1 \cdot x_2 + 1)^2$ | Degree of the polynomial is 2 |
| Cubic Polynomial | $k(x_1, x_2) = (x_1 \cdot x_2 + 1)^3$ | Degree of the polynomial is 3 |

## Performance metrics

In this study, the accuracy, precision, recall, specificity, F1-score, and Matthews correlation coefficient (MCC) metrics were used to evaluate the performance of the classification models. These metrics are defined as follows:

$$Accuracy(Acc) = \frac{TP + TN}{TP + TN + FP + FN}, \tag{16}$$

$$Precision(Pre) = \frac{TP}{TP + FP}, \tag{17}$$

$$Recall(Rec) = \frac{TP}{TP + FN}, \tag{18}$$

$$Specificity(Spe) = \frac{TN}{TN + FP}, \tag{19}$$

$$F1 - score(F1) = \frac{2TP}{2TP + FN + FP}, \tag{20}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{21}$$

There are four possible outcomes in the evaluation of classification results: It is considered true positive (TP) if the sample is positive and classified as positive, and false negative (FN) if the sample is classified as negative. The sample is considered true negative (TN) if negative and classified as negative, and false positive (FP) if classified as positive. MCC is used to obtain the Pearson product-moment correlation between actual and predicted values [37, 38]. MCC values are in the range of [-1, +1], and extreme values of +1 and -1, respectively, are obtained in the situations of perfect classification and perfect misclassification.

The data were divided into training and testing groups using the $k$-fold cross validation (CV) method in order to obtain the most effective, reliable, and acceptable model. The $k$-fold CV approach splits the data set into $k$ parts, $k-1$ of which is utilized for training while the other remains for testing. Each fold's result is acquired once this method is carried out on all datasets. The overall performance of the classifier is then evaluated by averaging all folds. The 5-fold CV was used to assess the performance of the classifiers in this study.

## 3 Experimental results

In this study, detection and classification models were obtained using the class-balanced synthetic VOICED dataset, which includes 342 healthy and 342 pathological data, each of which consists of three disorder types containing 114 data. The analysis of the voice signals was carried out in two stages: the cepstral domain analysis of the direct raw signals and the analysis of the IMFs obtained by the multivariate EMD, EEMD and CEEMDAN methods. The mode and IMFs of the decomposition methods to be used in the extraction of features were performed over temporal energy. Figure 7 shows temporal energy distribution according to IMFs for a healthy and pathological voice. Thus, only IMFs in the modes with the highest energy value were selected. In the pathological voice analysis shown in Figure 7a, the first 5 IMFs for EMD and EEMD, and the first 6 IMFs for CEEMDAN have the highest energy. Similarly, the healthy voice energy distribution shown in Figure 7b is similar to the pathological voice. Therefore, IMFs at the first 5 decomposition levels of EMD, EEMD and CEEMDAN methods were used for feature extraction.

In obtaining pathological voice detection and pathology type classification models, feature extraction was performed in two different ways. First, MFCC and its derivatives, LPC and LPCC cepstral techniques were applied directly to the raw signals and a feature vector was obtained. In the

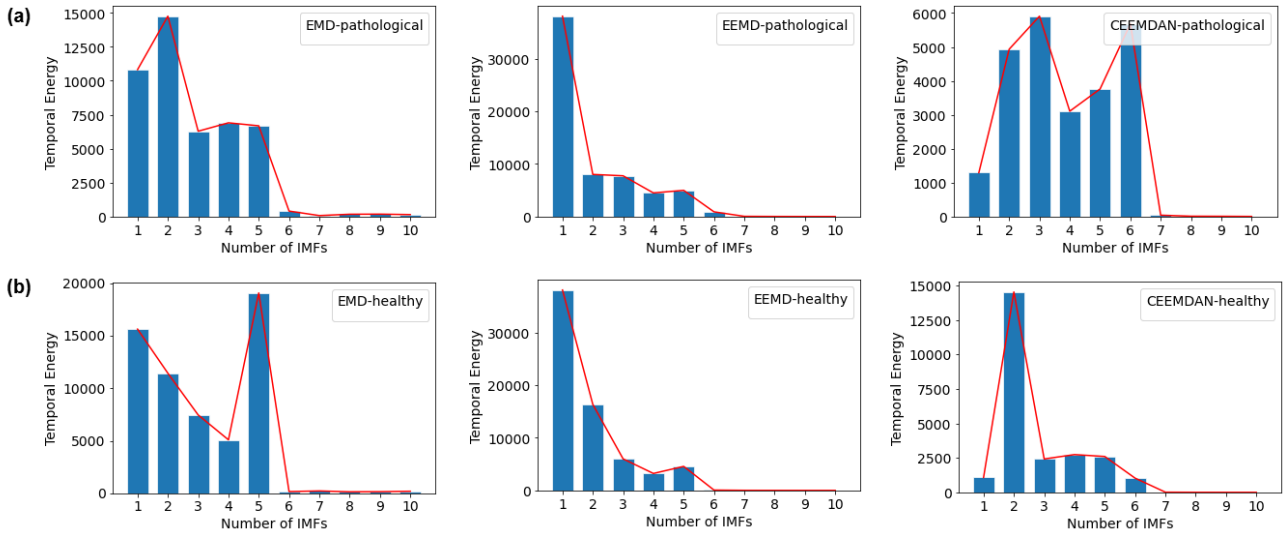


**Figure 7.** Temporal energy distribution of IMFs from voice decomposed by EMD, EEMD and CEEMDAN (from left to right) (a) for pathological, (b) for healthy

second way, cepstral techniques were applied to each IMFs in the first 5 modes obtained by EMD, EEMD and CEEMDAN methods and a vector with more features was achieved. The cepstral-domain features extracted from raw signals and EMD-based, EEMD-based and CEEMDAN-based cepstral features obtained from decomposition methods were used with the SVM classifier algorithm, and detection of pathological voice and classification of three pathology types were performed. The linear, quadratic, and cubic were used as kernel functions for the SVM classifier to obtain the detection and classification models.

The results of the pathological voice detection models for all features and the selected feature set are given in Table 4. The use of cepstral features based on EMD as the input vector for SVM-cubic provides the highest performance with 99.85% accuracy, 100% precision, 99.71% recall, 100% specificity, 99.85% F1-score, and 0.997 Matthews correlation coefficients. The results showed that features based on EEMD and CEEMDAN provide similar performance to EMD, while traditional cepstral features can achieve lower performance with 94.30% accuracy, 94.12% F1-score, and 0.888 MCC. Thus, a comprehensive review shows that methods based on adaptive decomposition outperform traditional methods in the detection of pathological voices.

In the pathological voice detection model based on EMD, the first three features with the highest distinctiveness among the selected features are shown in Figure 8. These features are listed as the 8th ΔΔ-MFCCs extracted from the IMFs in the 5th mode, the 3rd ΔΔ-MFCCs in the 3rd mode, and the 13th MFCCs in the 4th mode, respectively. It has been observed that MFCCs and its derivatives acoustic parameters obtained by EMD are highly effective features in the detection of pathological voice.

The confusion matrix obtained by traditional and adaptive decomposition-based approaches, which provide the highest performance in the detection of healthy and pathological voices, is illustrated in Figure 9. The confusion matrix shown in Figure 9a shows that 30 of the pathological voices and 9 of the healthy voices were detected incorrectly when using traditional cepstral domain features extracted from the raw signal. The confusion matrix that provides the best performance in the approach where the cepstral domain features based on EMD are the input of an SVM-cubic classifier is shown in Figure 9b. In this approach, it was concluded that all healthy voices were detected correctly and 1 of the pathological voices was detected incorrectly.

Statistical significance analysis was performed using McNemar's Chi-square test for EMD-based

**Table 4.** Comparative performance results of all features and selected features by ReliefF algorithm for healthy and pathological detection

| | Performance for all features | | | | | | Performance for selected features | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods/Algorithm | Acc(%) | Pre(%) | Rec(%) | Spe(%) | F1(%) | MCC | Acc(%) | Pre(%) | Rec(%) | Spe(%) | F1(%) | MCC |
| Cepstral/SVM Linear | 66.08 | 68.33 | 59.94 | 72.22 | 63.86 | 0.324 | 67.25 | 69.41 | 61.70 | 72.81 | 65.33 | 0.347 |
| Cepstral/SVM Quadratic | 84.50 | 91.55 | 76.02 | 92.98 | 83.07 | 0.700 | 92.25 | 95.58 | 88.60 | 95.91 | 91.96 | 0.847 |
| Cepstral/SVM Cubic | 89.77 | 92.77 | 86.26 | 93.27 | 89.39 | 0.797 | 94.30 | 97.20 | 91.23 | 97.37 | 94.12 | 0.888 |
| EMD/SVM Linear | 89.33 | 92.97 | 85.09 | 93.57 | 88.85 | 0.789 | 93.57 | 98.38 | 88.60 | 98.54 | 93.23 | 0.876 |
| EMD/SVM Quadratic | 97.66 | 100 | 95.32 | 100 | 97.60 | 0.954 | 99.71 | 100 | 99.42 | 100 | 99.71 | 0.994 |
| EMD/SVM Cubic | 98.39 | 100 | 96.78 | 100 | 98.37 | 0.968 | 99.85 | 100 | 99.71 | 100 | 99.85 | 0.997 |
| EEMD/SVM Linear | 84.21 | 87.50 | 79.82 | 88.60 | 83.49 | 0.687 | 91.08 | 98.95 | 83.04 | 99.12 | 90.30 | 0.832 |
| EEMD/SVM Quadratic | 97.66 | 100 | 95.32 | 100 | 97.60 | 0.954 | 99.12 | 100 | 98.25 | 100 | 99.12 | 0.983 |
| EEMD/SVM Cubic | 98.68 | 100 | 97.37 | 100 | 98.67 | 0.974 | 99.71 | 100 | 99.42 | 100 | 99.71 | 0.994 |
| CEEMDAN/SVM Linear | 82.60 | 89.68 | 73.68 | 91.52 | 80.90 | 0.663 | 88.30 | 97.12 | 78.95 | 97.66 | 87.10 | 0.780 |
| CEEMDAN/SVM Quadratic | 96.78 | 100 | 93.57 | 100 | 96.68 | 0.938 | 98.54 | 100 | 97.08 | 100 | 98.52 | 0.971 |
| CEEMDAN/SVM Cubic | 97.66 | 100 | 95.32 | 100 | 97.60 | 0.954 | 99.27 | 100 | 98.54 | 100 | 99.26 | 0.985 |

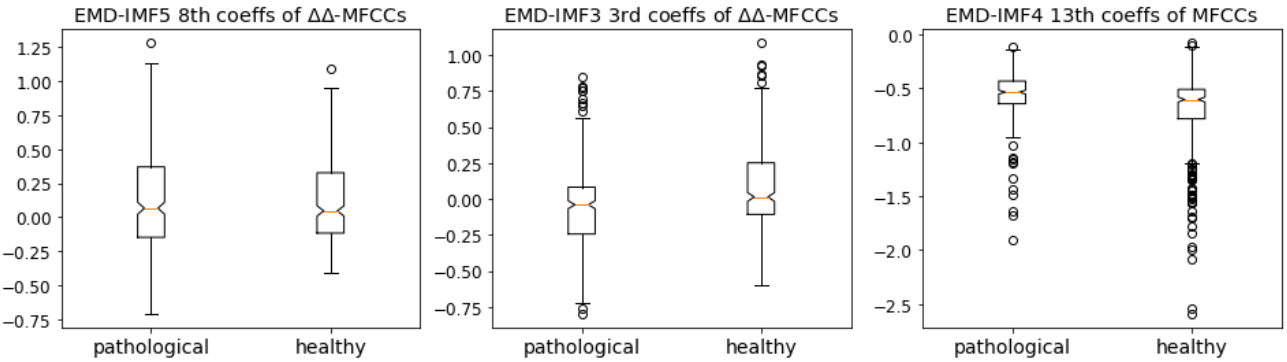


**Figure 8.** Statistical distribution representation of the top three features with the highest distinctiveness for the healthy and pathological detection model

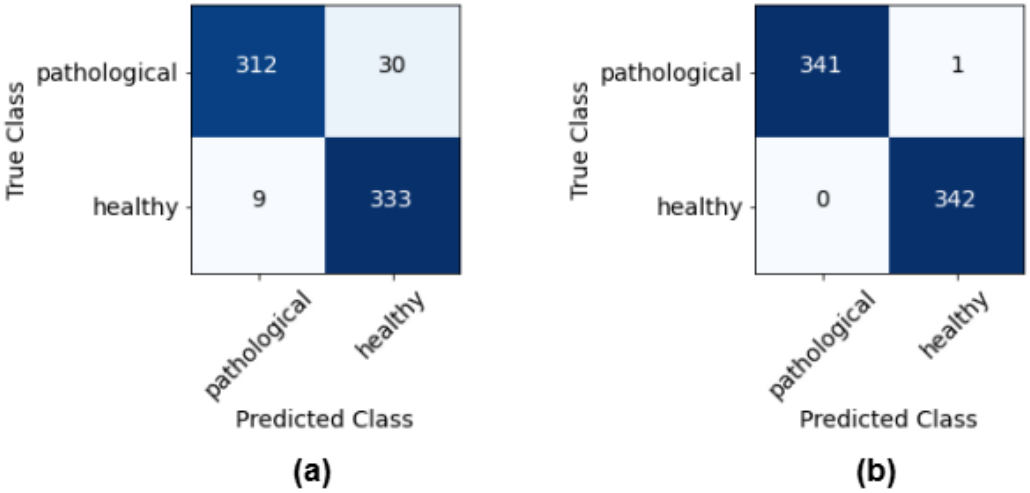


**Figure 9.** Confusion matrix of the highest-performing model in healthy and pathological voice detection (a) for traditional cepstral domain features (b) for cepstral domain features based on EMD

MFCCs, Δ-MFCCs, ΔΔ-MFCCs, LPCs and LPCCs cepstral domain features in pathological voice detection. Table 5 gives the results of McNemar's test using the Chi-square and p-values ($\alpha = 0.05$) for the features. All feature cases, with the exception of LPCs, are rejected by the test which was performed. As seen in Table 5, feature vectors with p-values near zero can distinguish between pathological and healthy voices and are statistically significant.

**Table 5.** McNemar's Chi-square statistical test for pathological voice detection

| Feature Set | $\chi^2$ | p-value | SD (p<0.05) |
|---|---|---|---|
| MFCCs | 6.857 | 0.0088 | Yes |
| Δ-MFCCs | 12.96 | 0.0003 | Yes |
| ΔΔ-MFCCs | 7.578 | 0.0059 | Yes |
| LPCs | 1.928 | 0.1649 | No |
| LPCCs | 22.40 | 0.0001 | Yes |

$\chi^2$ : Chi-square value, SD : Significant difference

In the pathology type detection experiment, the VOICED dataset's pathologies for reflux laryngitis, hyperkinetic dysphonia, and hypokinetic dysphonia had been detected. Table 6 shows the classification results for the three pathology types utilizing both traditional and decomposition-based cepstral features. Comprehensive results demonstrate that the SVM-quadratic method

combined with the cepstral features extracted from IMFs acquired by EEMD resulted in improved classification performance. The EEMD-based classification model, which was more successful than other approaches, yielded 96.49% accuracy, 96.74% precision, 96.49% recall, 98.25% specificity, 96.46% F1-score, and 0.949 MCC values. In addition, the maximum values of 85.09% accuracy, 84.84% F1-score, and 0.779 MCC were achieved when cepstral features were used with SVM-cubic. The results revealed that decomposition-based approaches are more effective than traditional cepstral features.

Figure 10 shows the top three distinctive features of the EEMD-based cepstral features for pathology type detection. These features are the 3rd, 1st, and 6th LPCs coefficients obtained from IMF2, IMF3, and IMF4, respectively. Thus, it was concluded that LPCs cepstral coefficients extracted from IMFs obtained by EEMD are quite effective in detecting pathological voice types.

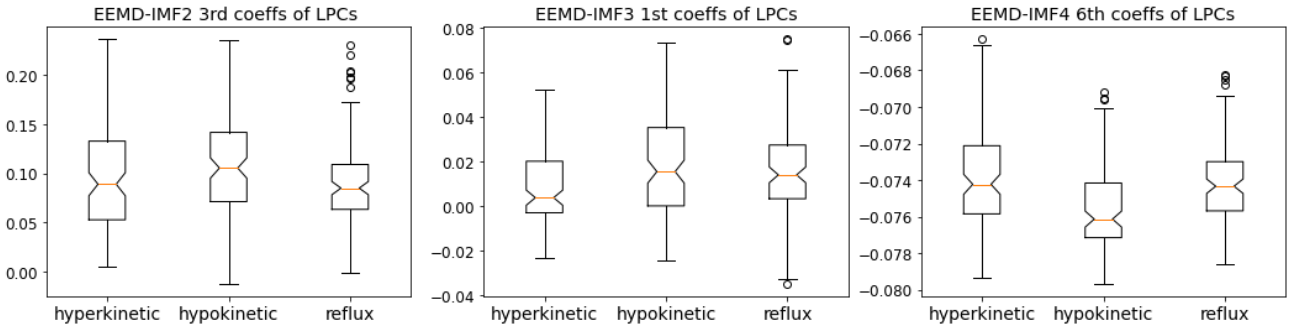


**Figure 10.** Statistical distribution representation of the top three features with the highest distinctiveness for the healthy and pathological detection model

The confusion matrices of classification models that provide the highest performance, obtained by traditional and EEMD-based cepstral domain approaches in the detection of hyperkinetic dysphonia, hypokinetic dysphonia and reflux laryngitis voice pathology types, are illustrated in Figure 11. The confusion matrix obtained with the traditional cepstral domain approach (Figure 11a) showed that hyperkinetic, hypokinetic and reflux pathological voices were classified with 71.93%, 89.47% and 93.86% accuracy rates, respectively. The confusion matrix of the model based on EEMD (Figure 11b) showed that accuracy rates of 89.47% for hyperkinetic and 100% for hypokinetic and reflux were achieved. Thus, it was concluded that the efficiency of traditional cepstral features was increased with the EEMD approach.

The highest performance results from the binary classification of pathological voice types were examined to demonstrate the ability of the proposed approach. The results showed that EEMD for hyperkinetic dysphonia-hypokinetic dysphonia (99.08% accuracy and 99.03% F1-score), CEEMDAN for hyperkinetic dysphonia-reflux laryngitis (98.18% accuracy and 98.08% F1-score), EEMD and CEEMDAN features for hypokinetic dysphonia-reflux laryngitis (100% accuracy and F1-score) provide the highest performance for classification of pathological voice types.

The statistical significance analysis was performed using McNemar's Chi-square test for the feature sets used in these high-performance binary pathological voice type classification models. Table 7 gives the Chi-square and p-values ($\alpha = 0.05$) obtained for the features. The test results show that all features except LPCCs for hyperkinetic dysphonia-hypokinetic dysphonia, all features for hyperkinetic dysphonia-reflux laryngitis and only LPCs for hypokinetic dysphonia-reflux laryngitis classes reject the null hypothesis. As shown in Table 7, feature vectors with p-values close to zero are statistically significant and highly effective in distinguishing pathological voice types.

**Table 6.** Comparative performance results of all features and selected features by ReliefF algorithm for pathological voice (hyperkinetic, hypokinetic and reflux laryngitis) classification

| | Performance for all features | | | | | | Performance for selected features | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods/Algorithm | Acc(%) | Pre(%) | Rec(%) | Spe(%) | F1(%) | MCC | Acc(%) | Pre(%) | Rec(%) | Spe(%) | F1(%) | MCC |
| Cepstral/SVM Linear | 64.91 | 65.39 | 64.91 | 82.46 | 64.61 | 0.477 | 65.79 | 66.11 | 65.79 | 82.89 | 65.55 | 0.489 |
| Cepstral/SVM Quadratic | 76.02 | 76.57 | 76.02 | 88.01 | 75.61 | 0.644 | 81.58 | 81.89 | 81.58 | 90.79 | 81.36 | 0.726 |
| Cepstral/SVM Cubic | 82.16 | 82.33 | 82.16 | 91.08 | 81.94 | 0.734 | 85.09 | 85.43 | 85.09 | 92.54 | 84.84 | 0.779 |
| EMD/SVM Linear | 74.85 | 74.77 | 80.41 | 74.85 | 87.43 | 0.623 | 90.64 | 91.10 | 90.64 | 95.32 | 90.37 | 0.863 |
| EMD/SVM Quadratic | 89.77 | 90.10 | 89.77 | 94.88 | 89.52 | 0.849 | 95.61 | 95.67 | 95.61 | 97.81 | 95.56 | 0.935 |
| EMD/SVM Cubic | 92.40 | 92.63 | 92.40 | 96.20 | 92.27 | 0.888 | 92.98 | 93.65 | 92.98 | 96.49 | 92.75 | 0.899 |
| EEMD/SVM Linear | 77.49 | 77.58 | 77.49 | 88.74 | 77.18 | 0.664 | 91.81 | 92.68 | 91.81 | 95.91 | 91.52 | 0.883 |
| EEMD/SVM Quadratic | 91.52 | 91.52 | 91.52 | 95.76 | 91.45 | 0.873 | 96.49 | 96.74 | 96.49 | 98.25 | 96.46 | 0.949 |
| EEMD/SVM Cubic | 93.86 | 93.92 | 93.86 | 96.93 | 93.79 | 0.908 | 94.15 | 94.72 | 94.15 | 97.08 | 94.02 | 0.916 |
| CEEMDAN/SVM Linear | 82.16 | 82.23 | 82.16 | 91.08 | 82.11 | 0.733 | 90.35 | 90.53 | 90.35 | 95.18 | 90.11 | 0.857 |
| CEEMDAN/SVM Quadratic | 93.27 | 93.45 | 93.27 | 96.64 | 93.24 | 0.900 | 95.91 | 96.01 | 95.91 | 97.95 | 95.86 | 0.939 |
| CEEMDAN/SVM Cubic | 94.15 | 94.31 | 94.15 | 97.08 | 94.12 | 0.913 | 94.44 | 94.88 | 94.44 | 97.22 | 94.31 | 0.919 |

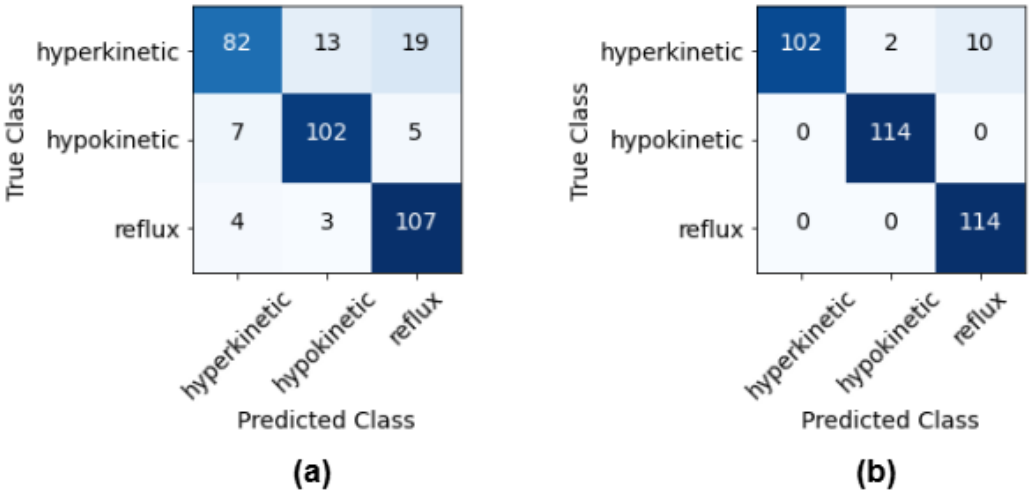


**Figure 11.** Confusion matrix of the highest-performing model in pathology types classification (a) for traditional cepstral domain features (b) for cepstral domain features based on EEMD

**Table 7.** McNemar's Chi-square statistical test for classification of pathology types

| | Hyperkinetic Dysphonia vs. Hypokinetic Dysphonia | | | Hyperkinetic Dysphonia vs. Reflux Laryngitis | | | Hypokinetic Dysphonia vs. Reflux Laryngitis | | |
|---|---|---|---|---|---|---|---|---|---|
| Features Set | $\chi^2$ | p-value | SD (p<0.05) | $\chi^2$ | p-value | SD (p<0.05) | $\chi^2$ | p-value | SD (p<0.05) |
| MFCC | 6.500 | 0.0107 | Yes | 13.136 | 0.0002 | Yes | 1.125 | 0.2888 | No |
| Δ-MFCC | 12.12 | 0.0005 | Yes | 19.862 | 0.0001 | Yes | 0.100 | 0.7518 | No |
| ΔΔ-MFCC | 14.81 | 0.0001 | Yes | 6.6657 | 0.0098 | Yes | 0.166 | 0.6830 | No |
| LPC | 19.36 | 0.0001 | Yes | 12.892 | 0.0003 | Yes | 8.000 | 0.0371 | Yes |
| LPCC | 6.500 | 0.0107 | No | 11.130 | 0.0008 | Yes | 4.442 | 0.0504 | No |

$\chi^2$ : Chi-square value, SD : Significant difference

## 4 Discussion

Voice disorders that cause deterioration of people's communication skills are pathological events caused by vocal cord paralysis, intensive drug abuse and inappropriate use of voice. Pathological voices are difficult to identify and detect, and many researchers have proposed approaches based on machine and deep learning to overcome this difficulty. The machine learning technique consists of feature extraction and analysis in the identification of abnormal voices. It is very important to choose distinctive and effective voice features and to decide on the best classification algorithm. Pathological voices are discovered utilizing raw signals or images obtained by a transformation process in deep learning algorithms that do not involve feature extraction techniques. However, in order to achieve outstanding performance with deep learning applications, an extensive amount of data is required. Therefore, machine learning-based algorithms that can provide high performance with small dataset in pathological voice detection and classification studies stand out in the literature. There are four datasets that are heavily used in state-of-the-art VPD studies, and they are as follows: Massachusetts Eye and Ear Infirmary (MEEI) [39], Saarbruecken Voice Database (SVD) [40], Hospital Universitario Príncipe de Asturias (HUPA) [41] and VOice ICar fEDerico II (VOICED) [23]. All of these datasets are class-unbalanced in terms of the healthy and pathological voice data included. The comparisons of studies in the field of VPD, which mostly use traditional feature extraction techniques and machine learning approaches, are given in Table 8.

Chen et al. [15] obtained 93.30% accuracy and 94% F1 score by using EMD and LPCC features

**Table 8.** Comparison of proposed approach with state-of-the-art pathological voice detection systems

| Author and Year | Dataset | Class (# of voices) | Features | Classifier | Results (%) |
|---|---|---|---|---|---|
| Chen et al. [15], 2020 | VOICED | Healthy (58)<br>Pathological (150) | LPCCs/EMD | KNN | Acc: 93.30<br>Sen: 95.00<br>Pre: 93.00<br>F1: 94.00 |
| Hammami et al. [19], 2020 | Self-dataset | Healthy (30)<br>Pathological (28) | Higher Order Statistics/EMD-DWT | SVM | Acc: 94.82<br>Sen: 92.85<br>Spe: 96.66 |
| Al-Dhief et al. [12], 2021 | SVD | Healthy (687)<br>Pathological (1354) | MFCCs | OSELM | Acc: 91.17<br>Pre: 94.00<br>Rec: 91.00 |
| Omeroglu et al. [42], 2022 | SVD | Healthy (687)<br>Pathological (1354) | MFCCs, LPCs, Pitch, and<br>CNN-based Electroglottography (EGG) | SVM | Acc: 90.10<br>Sen: 92.90<br>Spe: 84.60<br>F1: 92.57 |
| Zhou et al. [43], 2022 | HUPA | Healthy (197)<br>Pathological (169) | Gammatone spectral latitude (GTSL) | SVM | Acc: 97.40 |
| Abdulmajeed et al. [21], 2023 | SVD | Healthy (687)<br>Pathological (1354) | MFCCs, zero-crossing rate (ZCR)<br>and spectrograms | LSTM | Acc: 99.30<br>Sen: 99.00<br>Spe: 99.50<br>Pre: 99.00<br>F1: 99.00 |
| Lee and Lee [44], 2023 | Synthetic<br>SVD | Healthy (1354)<br>Pathological (1354) | LPCs | CNN | Acc: 98.89<br>Rec: 100.0<br>Spe: 97.00<br>Rec: 99.00 |
| This study | Synthetic<br>VOICED | Healthy (342)<br>Pathological (342) | EMD-based cepstral features | SVM | Acc: 99.30<br>Pre: 100.0<br>Rec: 99.71<br>Spe: 100.0<br>F1: 99.85<br>MCC: 0.997 |
| This study | Synthetic<br>VOICED | Hyperkinetic (114)<br>Hypokinetic (114)<br>Reflux (114) | EEMD-based cepstral features | SVM | Acc: 96.49<br>Pre: 96.74<br>Rec: 96.49<br>Spe: 98.25<br>F1: 96.46<br>MCC: 0.949 |

with KNN. Hammami et al. [19] used the SVD dataset in the training of the algorithm in the proposed approach and the RABTA Hospital of Tunisia (30 healthy and 28 pathological data) containing its own private data in the testing. The authors achieved 94.82% accuracy, 92.85% sensitivity and 96.66% specificity rates by using higher-order statistics and EMD-DWT features with SVM algorithm. In another study, Al-Dhief et al. [12] proposed an approach based on MFCCs cepstral features and online sequential extraction learning machine (OSELM). The authors reported that 91.17% accuracy, 94% precision and 91.00% recall performance values were achieved in the proposed model. Omeroglu et al. [42] proposed a hybrid approach using cepstral features and CNN-based Electroglottography (EGG) signal with SVM classifier. The authors achieved 90.10% accuracy, 92.9% sensitivity, 84.6% specificity and 92.57% F1-score performance values. Zhou et al. [43] proposed a model based on Gammatone spectral latitude (GTSL) and RF classifier, using containing 197 healthy and 169 pathological voice, and achieved an accuracy rate of 97.40%. Abdulmajeed et al. [21], who proposed a model based on deep learning, used MFCCs, zero-crossing rate (ZCR) and spectrogram images with long-short term memory (LSTM), which is a type of recurrent neural networks. The authors reported that they achieved over 99% performance for all metrics in pathological sound detection. In the study using synthetic data with SMOTE technique, Lee and Lee [44] performed an approach based on LPCs features and CNN deep learning model. It was reported that 98.89% accuracy, 100% recall, 97% specificity and 99% F1 values were achieved in the study.

In this study, an approach based on healthy and pathological voice detection and classification of three pathology types such as hyperkinetic dysphonia, hypokinetic dysphonia and reflux laryngitis is proposed. The traditional cepstral features and features based on mode decomposition (EMD, EEMD and CEEMDAN) were used with the SVM classifier which has been proven effective in voice/speech recognition, to obtain detection and classification models. The feature sets were extracted using the raw signal (1x60), each IMFs (1x60) and IMFs of the first five levels (5x60). In addition, class-balanced synthetic data were created by applying the SMOTE technique to these features. All features and the selected features with the highest distinctiveness obtained with the ReliefF algorithm were used with the SVM classifier, which includes linear, quadratic and cubic kernel functions for the detection and classification models. Extensive results show that the highest performance is achieved with selected cepstral features based on EMD and SVM-cubic algorithms in pathological voice detection. The proposed detection model has 99.85% accuracy, 100% precision, 99.71% recall, 100% specificity, 99.85% F1-score and 0.997 MCC. In the other approach in which three pathological voice types are classified, the highest performance model with 96.49% accuracy, 96.74% precision, 98.25% specificity, 96.46% F1 and 0.949 MCC values was achieved with selected cepstral features based on EEMD and SVM-quadratic classifier. It can be clearly stated that the feature extraction approach based on mode decomposition has a higher performance than traditional features obtained from the raw signal, both in pathology detection and classification. In addition, MFCC and its derivatives stand out as the acoustic parameters that provide the highest distinctiveness in the detection of pathological and healthy voice, and LPC parameters in the detection of the pathology type.

The main points that distinguish the proposed approaches in this research apart from previous studies are outlined in the following order: (1) Detection and classification of pathology based on voice signals have been performed with cepstral feature extraction approaches based on mode decomposition as an alternative to traditional cepstral feature extraction. (2) The efficiency of the IMFs obtained by EMD, EEMD and CEEMDAN methods in feature extraction was analyzed. (3) The class-imbalanced data problem was overcome with SMOTE technique and its effects on model performance were examined. (4) High distinctiveness features were selected using the ReliefF algorithm. (5) Two high-performance models based on the SVM classifier were obtained for

pathology detection and classification. (6) A computer-aided decision system has been proposed that can help experts as an alternative to existing approaches for pathological voice detection. Despite promising results, limitations such as model generalizability require further investigation. This study primarily focused on the VOICED dataset, which includes three types of vocal pathologies. The generalizability of models for detecting voice pathologies depends on the diversity of the datasets, the representational capacity of the extracted features, and the design of the model. Factors such as imbalances in datasets, variations in language and accent, and environmental noise can significantly influence the model's performance under different conditions. Furthermore, inconsistencies in data collection processes and the limited availability of pathological samples constrain the generalizability of the models. Future research should aim to expand the dataset to encompass a broader range of materials to further validate the generalization capabilities of the proposed models. Additionally, while mode decomposition-based cepstral features and SVM models have proven effective in detecting voice pathologies, there remains an opportunity to further enhance detection performance by exploring more advanced architectures, such as CNN, LSTM networks, and attention mechanism-based models.

## 5 Conclusion

In this study, it has been evaluated the performances of the proposed new adaptive cepstral features extracted by mode decomposition to distinguish between pathological and normal voices. The study utilized adaptive feature extraction derived from EMD, EEMD, and CEEMDAN as well as the conventional cepstral feature extraction method applied directly to the raw signal. The class-balanced data were obtained by applying the SMOTE technique to the acquired feature sets for VPD model. ReliefF algorithm was applied to MFCCs and derivatives, LPCs and LPCCs cepstral features extracted from voice signals and five-level IMFs, and the performance improvement was achieved with reduced selected features. The selected features were used as input to the SVM and, the VPD models that provided the best performance were implemented. Comprehensive results showed that EMD-based features and SVM-cubic for VPD, and EEMD-based features and SVM-quadratic approaches for multi-classification of pathology type provided the highest accuracy performance. It can be said that this high-performance automatic decision support system, developed as an alternative to traditional approaches, can be used as an auxiliary tool for pre-diagnosis in VPD.

## Declarations

### Use of AI tools

The author declares that he has not used Artificial Intelligence (AI) tools in the creation of this article.

### Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

### Ethical approval

The author states that this research complies with ethical standards. This research does not involve either human participants or animals.

### Consent for publication

Not applicable

## Conflicts of interest

The author declares that there is no conflict of interest related to this paper.

## Funding

There is no funding source for this study.

## Author's contributions

Ö.A.: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Visualization, Writing - Original Draft, Writing - Review & Editing. The author has read and agreed to the published version of the manuscript.

## Acknowledgements

Not applicable

## References

[1] Hegde, S., Shetty, S., Rai, S. and Dodderi, T. A survey on machine learning approaches for automatic detection of voice disorders. *Journal of Voice*, 33(6), 947.e11-947.e33, (2019). [CrossRef]

[2] Ding, H., Gu, Z., Dai, P., Zhou, Z., Wang, L. and Wu, X. Deep connected attention (DCA) ResNet for robust voice pathology detection and classification. *Biomedical Signal Processing and Control*, 70, 102973, (2021). [CrossRef]

[3] Verde, L., De Pietro, G. and Sannino, G. Voice disorder identification by using machine learning techniques. *IEEE Access*, 6, 16246-16255, (2018). [CrossRef]

[4] Islam, R., Abdel-Raheem, E. and Tarique, M. Voice pathology detection using convolutional neural networks with electroglottographic (EGG) and speech signals. *Computer Methods and Programs in Biomedicine Update*, 2, 100074, (2022). [CrossRef]

[5] Chen, L. and Chen, J. Deep neural network for automatic classification of pathological voice signals. *Journal of Voice*, 36(2), 288.e15-288.e24, (2022). [CrossRef]

[6] Al-Nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z., Mesallam, T.A., Farahat, M. et al. An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification. *Journal of Voice*, 31(1), 113.e9-113.e18, (2017). [CrossRef]

[7] Brockmann, M., Drinnan, M.J., Storck, C. and Carding, P.N. Reliable jitter and shimmer measurements in voice clinics: the relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task. *Journal of Voice*, 25(1), 44-53, (2011). [CrossRef]

[8] Ferrand, C.T. Harmonics-to-noise ratio: an index of vocal aging. *Journal of Voice*, 16(4), 480-487, (2002). [CrossRef]

[9] Neto, B.G.A., Fechine, J.M., Costa, S.C. and Muppa, M. Feature estimation for vocal fold edema detection using short-term cepstral analysis. In Proceedings, *IEEE 7th International Symposium on BioInformatics and BioEngineering*, pp. 1158-1162, Boston, USA, (2007, October). [CrossRef]

[10] Gelzinis, A., Verikas, A. and Bacauskiene, M. Automated speech analysis applied to laryngeal disease categorization. *Computer Methods and Programs in Biomedicine*, 91(1), 36-47, (2008). [CrossRef]

[11] Anusuya, M.A. and Katti, S.K. Front end analysis of speech recognition: a review. *International Journal of Speech Technology*, 14, 99-145, (2011). [CrossRef]

[12] Al-Dhief, F.T., Baki, M.M., Latiff, N.M.A.A., Malik, N.N.N.A., Salim, N.S., Albader, M.A.A. Voice pathology detection and classification by adopting online sequential extreme learning machine. *IEEE Access*, 9, 77293-77306, (2021). [CrossRef]

[13] Jothilakshmi, S. Automatic system to detect the type of voice pathology. *Applied Soft Computing*, 21, 244-249, (2014). [CrossRef]

[14] Majidnezhad, V. and Kheidorov, I. An ANN-based method for detecting vocal fold pathology. *ArXiv Areprint, ArXiv:1302.1772*, (2013). [CrossRef]

[15] Chen, L., Wang, C., Chen, J., Xiang, Z. and Hu, X. Voice disorder identification by using Hilbert-Huang transform (HHT) and K nearest neighbor (KNN). *Journal of Voice*, 35(6), 932.e1-932.e11, (2021). [CrossRef]

[16] Hemmerling, D., Skalski, A. and Gajda, J. Voice data mining for laryngeal pathology assessment. *Computers in Biology and Medicine*, 69, 270-276, (2016). [CrossRef]

[17] Ali, Z., Alsulaiman, M., Elamvazuthi, I., Muhammad, G., Mesallam, T.A., Farahat, M. and Malki, K.H. Voice pathology detection based on the modified voice contour and SVM. *Biologically Inspired Cognitive Architectures*, 15, 10-18, (2016). [CrossRef]

[18] Akbari, A. and Arjmandi, M.K. An efficient voice pathology classification scheme based on applying multi-layer linear discriminant analysis to wavelet packet-based features. *Biomedical Signal Processing and Control*, 10, 209-223, (2014). [CrossRef]

[19] Hammami, I., Salhi, L. and Labidi, S. Voice pathologies classification and detection using EMD-DWT analysis based on higher order statistic features. *Irbm*, 41(3), 161-171, (2020). [CrossRef]

[20] Wu, H., Soraghan, J., Lowit, A. and Di Caterina, G. Convolutional neural networks for pathological voice detection. In Proceedings, *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1-4, Honolulu, Hawaii, USA, (2018, July). [CrossRef]

[21] Abdulmajeed, N.Q., Al-Khateeb, B. and Mohammed, M.A. Voice pathology identification system using a deep learning approach based on unique feature selection sets. *Expert Systems*, 42(1), e13327, (2023). [CrossRef]

[22] Chaiani, M., Selouani, S.A., Boudraa, M. and Yakoub, M.S. Voice disorder classification using speech enhancement and deep learning models. *Biocybernetics and Biomedical Engineering*, 42(2), 463-480, (2022). [CrossRef]

[23] Cesari, U., De Pietro, G., Marciano, E., Niri, C., Sannino, G. and Verde, L. A new database of healthy and pathological voices. *Computers & Electrical Engineering*, 68, 310-321, (2018). [CrossRef]

[24] Huang, N.E. Introduction to Hilbert-Huang transform and some recent developments. In *The Hilbert-Huang Transform in Engineering*, (pp. 1-23). CRC Press, USA, (2005).

[25] Arslan, Ö. and Karhan, M. Effect of Hilbert-Huang transform on classification of PCG signals using machine learning. *Journal of King Saud University-Computer and Information Sciences*, 34(10), 9915-9925, (2022). [CrossRef]

[26] Zhang, T., Zhang, Y., Sun, H. and Shan, H. Parkinson disease detection using energy direction features based on EMD from voice signal. *Biocybernetics and Biomedical Engineering*, 41(1), 127-141, (2021). [CrossRef]

[27] Zhaohua Wu, N.E.H. Ensemble empirical mode decomposition: A noise-assited. *Biomed Tech*, 55, 193-201, (2010).

[28] Torres, M.E., Colominas, M.A., Schlotthauer, G. and Flandrin, P. A complete ensemble empirical mode decomposition with adaptive noise. In Proceedings, *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4144-4147, Prague, Czech Republic, (2011, May). [CrossRef]

[29] Chen, X., Hu, M. and Zhai, G. Cough detection using selected informative features from audio signals. In Proceedings, *14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* pp. 1-6, Shanghai, China, (2021, October). [CrossRef]

[30] Ghoraani, B. and Krishnan, S. Time–frequency matrix feature extraction and classification of environmental audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), 2197-2209, (2011). [CrossRef]

[31] Fang, S.H., Tsao, Y., Hsiao, M.J., Chen, J.Y., Lai, Y.H., Lin, F.C. and Wang, C.T. Detection of pathological voice using cepstrum vectors: A deep learning approach. *Journal of Voice*, 33(5), 634-641, (2019). [CrossRef]

[32] Wang, S., Dai, Y., Shen, J.and Xuan, J. Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports*, 11, 24039, (2021). [CrossRef]

[33] Kira, K. and Rendell, L.A. A practical approach to feature selection. In *Machine Learning Proceedings 1992*, (pp. 249-256). Morgan Kaufmann: USA, (1992). [CrossRef]

[34] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E. et al. Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. *IEEE Access*, 9, 19304-19326, (2021). [CrossRef]

[35] Vapnik, V. *The Nature of Statistical Learning Theory*. Springer Science & Business Media: New York, (1995). [CrossRef]

[36] Cortes, C. Support-vector networks. *Machine Learning*, 20, 273-297, (1995).

[37] Chicco, D. and Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, 1-13, (2020). [CrossRef]

[38] Powers, D.M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiv Preprint, ArXiv:2010.16061*, (2020). [CrossRef]

[39] Saenz-Lechon, N., Godino-Llorente, J.I., Osma-Ruiz, V. and Gómez-Vilda, P. Methodological issues in the development of automatic systems for voice pathology detection. *Biomedical Signal Processing and Control*, 1(2), 120-128, (2006). [CrossRef]

[40] Martínez, D., Lleida, E., Ortega, A., Miguel, A. and Villalba, J. Voice pathology detection on the Saarbrücken voice database with calibration and fusion of scores using multifocal toolkit. In Proceedings, *Advances in Speech and Language Technologies for Iberian Languages: IberSPEECH 2012 Conference*, pp. 99-109, Madrid, Spain, (2012, November). [CrossRef]

[41] Godino-Llorente, J.I., Osma-Ruiz, V., Sáenz-Lechón, N., Cobeta-Marco, I., González-Herranz, R. and Ramírez-Calvo, C. Acoustic analysis of voice using WPCVox: a comparative study with Multi Dimensional Voice Program. *European Archives of Oto-Rhino-Laryngology*, 265, 465-476, (2008). [CrossRef]

[42] Omeroglu, A.N., Mohammed, H.M. and Oral, E.A. Multi-modal voice pathology detection architecture based on deep and handcrafted feature fusion. *Engineering Science and Technology, an International Journal*, 36, 101148, (2022). [CrossRef]

[43] Zhou, C., Wu, Y., Fan, Z., Zhang, X., Wu, D. and Tao, Z. Gammatone spectral latitude features extraction for pathological voice detection and classification. *Applied Acoustics*, 185, 108417, (2022). [CrossRef]

[44] Lee, J.N. and Lee, J.Y. An efficient SMOTE-based deep learning model for voice pathology detection. *Applied Sciences*, 13(6), 3571, (2023). [CrossRef]

Mathematical Modelling and Numerical Simulation with Applications (MMNSA)
(https://dergipark.org.tr/en/pub/mmnsa)



**How to cite this article:** 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