Research Article GU J Sci 39(1): 113-128 (2026) DOI: 10.35378/gujs.1606193

Gazi University

Journal of Science

http://dergipark.org.tr/gujs

Detection of Voice Pathologies with Information Theory Based Features

Ecenur AKUREN[1,2], Derya YILMAZ[3,*]

[1]Gazi University, Graduate School of Natural and Applied Sciences, Program of Electrical and Electronics Engineering, 06500, Ankara, Türkiye
[2]Scientific and Research Council of Türkiye (TÜBİTAK), 06680, Ankara, Türkiye
[3]Gazi University, Faculty of Engineering, Department of Electrical and Electronics Engineering, 06500, Ankara, Türkiye

Highlights

•Detecting vocal cord pathologies in both binary and multiclass procedures.
• Using the recurrence plots quantities in voice analysis for the first time in literature.
• Efficient classification performances with only a few features from nonlinear domain.

Article Info

Received: 23 Dec 2024
Accepted: 03 Nov 2025

Keywords

Vocal cord diseases, Recurrence plots quantities, Discrete wavelet transform, Feature selection, Binary and multiclass classification

Abstract

Voice is produced by excited dynamic vocal tract system consisting of vocal cords, lips, tongue, with the air from lungs. Pathological cases in these components change the characteristics of the voice. Voice diseases impair quality of life, and some pathologies can be life-threatening, so early diagnosis is important. In literature, voice disorders can be generally investigated with traditional or acoustics signal analysis methods and detected by using machine learning. In this study, analysis and classification studies were realized for detecting vocal cord deformities such as cyst, polyp and sulcus vocalis, and healthy ones. Unlike literature, instead of acoustic features, features based on information theory such as recurrence plot quantities, entropies and dimensions coming from nonlinear approaches were used. Three feature selection and three classification methods were tried with nonlinear features calculated from whole signal and decomposed signal with discrete wavelet transform. Two classification procedures (binary and multiple class) with 10-fold cross validation were applied and tested with distinct groups. According to the results, in binary classification (healthy/ diseased), the best test accuracy of 99.2% with Coarse Tree classifier was obtained with only four features selected with MRMR algorithm. In multiclass (healthy/cyst/polyp/sulcus), the test accuracy of 91.6% was found as the best with only eight selected features by using ANOVA and Kernel Naive Bayes classifier. The results show that the use of a few features coming from nonlinear domain provided very effective and successful in classifying voice disorders in both binary and multiclass when compared to the studies in the literature.

1. INTRODUCTION

Speech is a basic communication tool, and the health of phonetic system components is most important for understandable speech because any pathological case called voice disorder disrupts the speech. Since the first symptom of pathologic conditions developed in organs that are effective in voice formation, such as the larynx, is voice disorder, and early detection of voice disorders is important for the treatment of these diseases, especially, oncological pathologies constitute a risk for human life [1]. Vocal cord pathologies which are among the causes of voice disorders are diagnosed by Ear, Nose and Throat experts with invasive procedures such as videolaryngoscopy and endoscopy [2]. These procedures have disadvantages such as being expensive and time-consuming. In addition, the importance of approaches based on non-invasive techniques, rather than invasive, in the diagnosis and treatment of diseases is increasing recently. Accurately identifying pathologies in the treatment of voice disorders is crucial. Since diagnoses made through voice measurements and evaluations, the accuracy of which depends on the experience of the specialist and involve subjectivity, so evaluating diseased conditions through voice analysis can increase decision accuracy. While there are programs that analyze the acoustic properties of the voice, there is no standard method or device based on voice analysis approaches that can be used in clinical practice for the diagnosis of voice disorders. In order to overcome these difficulties, studies on computer-aided diagnosis

*Corresponding author, e-mail: deryayilmaz@gazi.edu.tr

including various analyzing approaches have become an interest area in recent years [3, 4]. In this context, examining the voice by using new analysis approaches different from traditional acoustic methods and extracting different characteristics can significantly contribute to research in this field.

Voice analysis involves evaluating voice samples recorded from patients through various software tools consisting of traditional/acoustics signal analysis methods [3, 4]. The acoustics features of voice such as shimmer, jitter, fundamental frequency, energy etc. are extracted by these methods for analyzing the healthy/disorder conditions [3]. In recent years, studies related to the detection of voice disorder by processing the voice signal focus on new feature extraction and different machine learning models, so classifiers [4]. Mohamed et al. used Support Vector Machines (SVMs) to classify 78 voice recordings from Saarbruecken Voice Database (SVD) as healthy/pathological with the selected features of Zero Crossing Rate (ZCR), jitter, peaks and fundamental frequency (F0), and reached accuracy of 92% [5]. In the study of Gayathri and Priya's, laryngitis, dysphonia and leukoplakia disorders were examined with energy, entropy, root mean square, ZCR and Mel Frequency Cepstral coefficients (MFCCs) features calculated for the /a/ vowel voice from the SVD [6]. According to their results, MFCCs can distinguish between the different voice pathologies considered. Murthy et al. classified Spasmodic Dysphonia (SD) and healthy voices from SVD by using mean energy calculated for non-overlapping hamming window frames [7]. They used Random Forest (RF) classifier and achieved 93.5% accuracy. In the Miliaresi et al.'s study, four voice disorders (phonotrauma, dysphonia, unilateral vocal paralysis, laryngeal neoplasm) were classified by using MFCCs and Convolution Neural Networks (CNNs) with feed-forward architecture of Neural Networks (NNs) [8]. The test accuracy was found to be 57%. The effect of some different models and features for detecting pathological cases in SVD were evaluated in Ribas et al's study [9]. The progressive self-supported representation learning approach was examined in Deep Neural Networks (DNNs) based system for Automatic Voice Disorder Detection (AVDD), and 90% accuracy was obtained. Barche et al. evaluated different epoch location determination algorithms for voice disordered speech in SVD [10]. Their results showed that the majority of these methods give good performance in healthy cases, moreover zero phase zero frequency filtering algorithms are adequate with the accuracy of 90.37% for diseased cases. AnilKumar et al., used MFCCs, Constant-Q Cepstral Coefficients (CQCCs) and All-Pole Group Delay Function (APGDF) with Bidirectional Long Short-Term Memory (BI-LSTM) for classifying pathological voice data in SVD, and 92.7% was obtained as the best accuracy with CQCCs [11]. In Ali et al.'s study, automatic detection system for voice disorders was suggested for insufficient voice samples in training phase [12]. For this purpose, a reference model was developed for healthy voice, and some decision rules were determined for voice disorder definition. The reference model composed with Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and k-means method demonstrated 74.9% accuracy. Wang et al. investigated the capability of DNNs to distinguish benign voice disorders and glottic neoplasia with features from MFCCs [13]. Their result showed that DNNs is an efficient method for classifying voice disorders with 98% success rate. CNNs and DNNs were used for detecting dysarthria disorder with MFCCs, and these models were compared RF and SVMs by Suresh et al. [14]. They obtained better performance by DNNs with 97.6% accuracy. Dysarthric disorder was also studied by Sahane et al., with Discrete Fourier Transform (DFT) and 13 features selected from Multi-Taper Mel Frequency Cepstral Coefficients (MTMFCCs) [15]. They achieved 99.04% and 96.00% accuracies with k-Nearest Neighbors (KNNs) and SVMs, respectively. Rebecca et al., the vowels /a/, /i/, /u/ from SVD with NNs and DNNs techniques were used to detect voice disorders, and they found that DNNs classifier was more successful than NNs with 92.3% accuracy for the /u/ voice [16]. Vieira et al. analyzed the nonstationary acoustic changes in voice with laryngeal disorders with Non-Stationarity Index (NSI) to determine frame sizes, and used MFCCs, Gammatone Frequency Cepstral Coefficients (GFCCs), Linear Predictive Coding Coefficients (LPCCs), and Recurrence Quantification Measures (RQMs) for features [17]. They reached best results of 96.36% with classifier of Linear Discriminant Analysis (LDA) by using the ROMs as features. In Islam and Tarique's study, Relative Spectral-Perceptual Linear Predictive (RASTA-PLP) technique was used for dysarthric disorder in SVD [18]. They tried Artificial Neural Networks (ANNs) and some other classifiers and found 100% accuracy when female and male subjects were considered separately. Ankışhan and Inam, investigated SVD and Voice Disorder (VD) database, which includes Sulcus Vocalis (SV), Vocal Cyst (VC) and Vocal Fold Polyp (VFP) disorders, with some features of entropy, MFCCs, shimmer, kurtosis, skewness, jitter, ZCR and raw signals [19]. Two fusion models, which are Decision Level Fusion (DLF) and Feature Level Fusion (FLF) by using Long Short-Term Memory Networks (LSTMs) and CNNs were

proposed. For healthy/unhealthy classification, the best results of 99.58% and 93.05% were obtained for VD and SVD with DLF, respectively. In multi-class classification, the best results of 90.18% (for VD database in 4 class) and 85.45% (for SVD database in 3 class) were achieved with DLF.

In this study, voice signals from cyst (Vocal Cyst-VC), polyp (Vocal Fold Polyp-VFP) and sulcus (Sulcus Vocalis-SV) patients and healthy people were analyzed and classified in binary and multiclass. For this aim, VD database obtained from Turkish people was used for detecting the voice diseases. For analysis, unlike the studies in the literature, Discrete Wavelet Transform (DWT), and a few nonlinear features based on information theory were used. As a new approach in voice analysis, it was purposed to find out the effectiveness of non-linear features especially based on Recurrence Plots (RPs) quantities. The binary (healthy/disordered) and multiclass classification procedures were performed with three classifiers by using three feature extraction methods, and the performances of the features and classifiers in detecting the pathological cases were evaluated.

2. MATERIAL and METHODS

2.1. Structure of Proposed Study

The applied processes of proposed study for analysis and detection of voice diseases are given in Figure 1. In preprocessing step, normalization, and denoising which is a technique for noise reduction were applied to the voice signals and then framing them into 40 ms periods with 20% overlap. Speech voices have a non-stationary structure. They are segmented at values ranging from 5 to 100 ms for analysis, because they are considered stationary over short time intervals [3-9, 11, 13, 17, 19]. Furthermore, since voice is a continuous variable due to both its dynamics and the changing vocal tract behavior, overlapping process is generally used with the half the segment size (50%) to maintain their integrity during framing and this ratio can be increased depending on the problem being solved. In this study, taking into account literature, segment size was selected at the value of 40 ms that would capture the changing dynamics and avoid computational complexity in order to obtain the features of voice. An overlapping ratio of 20% was used to reduce the similarity of the obtained features and to mitigate classification bias that could arise from this process. Additionally, the first and last seconds of the recordings were not used to avoid being affected by instability and energy changes at the beginning and end of the voice.

A few time-domain and some nonlinear features that come from information theory were extracted for whole signals and 6-level DWT coefficients. Feature selection procedures were applied for improving classification performance. Two classification examinations were realized as binary and multiclass. For classification, data (subjects) were randomly split into train and test groups. Randomly selected two subjects from each group were used for testing. Three classifiers were used in both classification schedules with 10-folds cross validation process. The test performances of classifiers were evaluated by using classifying metrics.

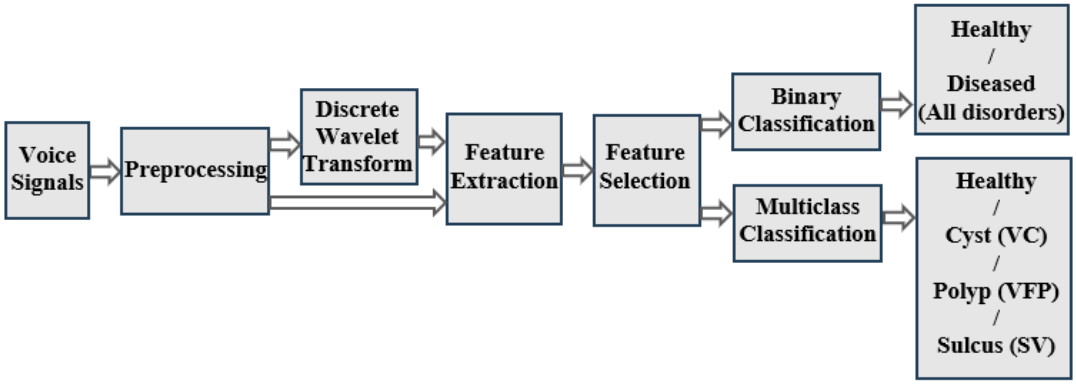


Figure 1. *Schematic representation of the study*

2.2. Data

VD database used in the study consists of 60 recordings of /a/ vowel taken form healthy people and cysts (VC), polyps (VFP) and sulcus (SV) patients by a specialist at the Department of Otorhinolaryngology, Faculty of Medicine in Sütçü İmam University. The dataset which was used in [19] consists of 28 male and 32 female subjects ages of between 15-72 (17-60 for healthy, 15-72 for disordered groups). The detail of the database is presented in Table 1. The recording took place in a calm environment with a background noise level ranging from 45 to 60 dB(A), a 16-bit resolution in single-channel pulse code modulation (PCM) format and rate of 44.1 kHz sampling. The duration of recordings is 10 seconds. Some signal samples of VD dataset are presented in Figure 2.

Table 1. *Information about the VD dataset*

	# of males	# of females	Total
Cyst (VC)	7	8	15
Polyp (VFP)	6	9	15
Sulcus (SV)	7	8	15
Healthy	8	7	15
Total	28	32	60

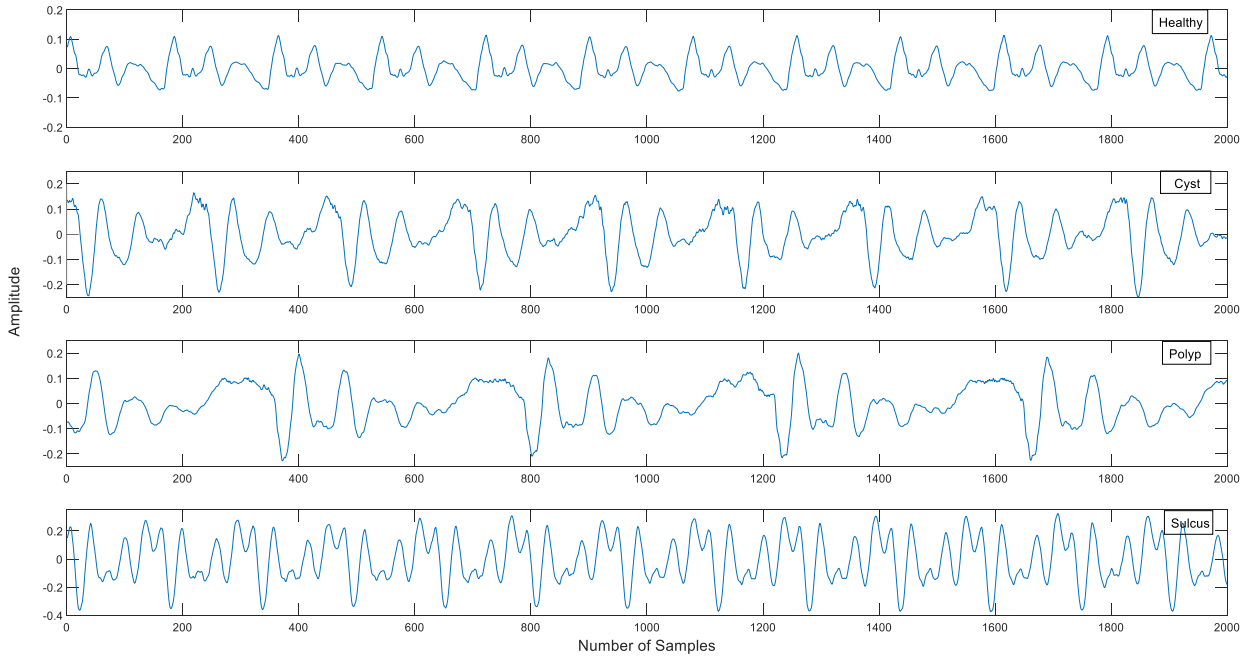


Figure 2. *Signal samples of vocal cord signals: From top to bottom in order healthy, cyst, polyp and sulcus*

2.3. Discrete Wavelet Transform (DWT)

Wavelet transform is a method of analysis that provides resolution in both domain time and frequency [20]. Wavelets are functions which dissociate data into particular frequency elements in different bands and analyze each element in these bands with a resolution corresponding to the particular component's scale. [20]. In wavelet analysis, the wavelet function selection (analyzing or main wavelet) is important. Time domain analysis is realized using the compressed and enlarged versions of the main wavelet. As a result, the data can be represented in relation to the wavelet coefficients. The mathematical definition of wavelet transform is given below; where s and Ψ are scale parameter and mother wavelet, respectively

$$F(\tau,s)=\frac{1}{\sqrt{|s|}}\int_{-\infty}^{\infty}f(t)\Psi^{*}\left(\frac{t-\tau}{s}\right)dt\,. \tag{1}$$

The Continuous Wavelet Transform (CWT) is the summation over time of the data obtained by using the mother wavelet, multiplied by its scaled and time translated versions. Wavelet coefficients are produced according to scale and position as a result of these processes. In DWT, translation and scaling are made as powers of two and DWT more efficient than CWT because of easy calculation [20]. DWT is preferred for analysing nonstationary signals such as voice. DWT coefficients are given for a *s(t)* as in Equation (2). In here, *j* and *k* are integers, and a and b are discretization parameters. The definition of DWT basis $\Psi_{j,k}(t)$ is given in Equation (3)

$$DWT_{j,k}=<s(t),\Psi_{j,k}>=\int_{-\infty}^{\infty}s(t)\Psi_{j,k}(t)dt \tag{2}$$

$$\Psi_{j,k}(t)=a_0^{-j/2}\Psi\left(a_0^{-j}t-kb_0\right). \tag{3}$$

Wavelet transform provides an effective analysis in time-frequency relationships [20]. In DWT analysis, high and low pass filters are implemented to the data to obtain the components of data in high and low frequency. The approximation and detail coefficients are obtained with the high (*h(n)*) and low (*g(n)*) pass filter, respectively. The approximation coefficients are high scale (low frequency) components of the data and give the original signal. On the other hand, the detail coefficients are low scale (high frequency) components of the signal and generally represent the noise elements of the data. Approximation and detail coefficients can be defined in the following equations

$$DWT_A=\sum_{n}s[k].f_0[-n+2k] \tag{4}$$

$$DWT_D=\sum_{n}s[k]f_1[-n+2k]\,. \tag{5}$$

In this study wavelet transformation was used for two aims. One is to reduce noise which is also called denoising and the other is the analysis of the signal components in different frequency bands for extracting the features from these bands.

Denoising is a noise reducing process by using wavelet method [21, 22]. In wavelet transform, the decomposing process of the signal into various bands of frequency for some decomposition levels is realized by sequentially filtering them through high and low pass filters with downsampling the signal. The original signal can be obtained by summing the filter outputs (approximation and detail coefficients of wavelet) at the same decomposition level by upsampling and passing through merging filters. This is a reconstruction process. For denoising, some thresholding methods are applied to the detail coefficient at different decomposition levels to reduce the noise. In this process, suitable choosing the wavelet type, decomposition level and threshold method is important to effective noise removal [21, 22]. In this study, wavelet of db8 from the Daubechies family, decomposition level of 3 and soft thresholding were selected for denoising [22, 23].

For analyzing the signal in different frequency bands by using wavelet transform, the signals were decomposed in 6 level with db8 wavelet. For these decomposition levels, approximation (A1, A2, A3,…,A6) and detail (D1, D2, D3,…,D6) components of signal were obtained. The features were also calculated for these components.

2.4. Features

Features provide important information to identify the signal. Voice signals are non-stationary and have nonlinear characteristics. So, in this study unlike literature, some features that come from information theory in nonlinear domain were used to obtain voice characteristics. Features were extracted for whole signal frames and the signal components obtained by using DWT in different frequency bands. The extracted features are explained below.

Log Energy Entropy (LEE): LEE characterizes the signal's energy distribution over time so gives information about the signal's irregularity and can be used for detecting some distortions in the signal [24]. For a discrete random variable of $x= x_0, x_1, x_2, \dots, x_{N-1}$having probability distribution function $p(x_i)$, and LEE is defined in Equation (6)

$$LEE(x) - \sum_{i=0}^{N-1} \left(log_2\left(p(x_i)\right)\right)^2 . \tag{6}$$

Short-Time Energy (STE): STE gives the amount of energy of a speech/vocal segment so easily detects endpoints. It is used for distinguishing between voiced and unvoiced segments [25]. The mathematical expression of the STE is given below. Where $x(m)$ is the signal, $w(n-m)$ is frame, N is the length of frame, and n is the center sample of the frame

$$E_n = \sum_{m=n-N+1}^{n} [x(m)w(n-m)]^2 . \tag{7}$$

Spectral Entropy (SE): SE is a property that represents the complexity of spectral distribution and so gives characteristics of frequency components of the signal and is calculated as in Equation 8. In here, $P(m)$ is the probability distribution of the signal's power spectrum [26]

$$SE = -\sum_{m=1}^{N} P(m)\, log_2\, P(m) . \tag{8}$$

Short-Term Zero Crossing Rate (STZCR): STZCR is a measure of zero crossings in the signal and given in Equation (9) [25]

$$STZCR = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn\,[x(m)]|\, w(n-m) . \tag{9}$$

In Equation (9), $sgn[x(n)] = \begin{Bmatrix} -1,\ x(n) < 0 \\ 1,\ x(n) \geq 0 \end{Bmatrix} and\ w(n) = \begin{Bmatrix} 1/2W, 0 \leq n \leq W-1 \\ 0, otherwise \end{Bmatrix}$, W:number of samples in the frame.

Higuchi Fractal Dimension (HFD): HFD is a quantity measuring the complexity of the signal, by taking the variability of signal [27]. HD comes from nonlinear analysis and computes the fractal dimension of system based on signal as given in Equation (10). In here, *d* is the state space dimension, *x* is signal and $L_m(d)$ is calculated while m goes from 1 to d. HFD is found by using Equation (11)

$$L_m(d) = \frac{\sum_{i=1}^{[(N-m)/d]\times d} |x(m+i\times d) - x[m+(i-1)\times d]|}{\frac{N-1}{\left[\frac{N-m}{d}\right]/d}} \tag{10}$$

$$L(d) = \frac{1}{d}\sum_{m=1}^{d} L_m(d) . \tag{11}$$

Katz Fractal Dimension (KFD): KFD is a nonlinear measure to identify the irregularity or complexity of a signal [28]. So, this quantity helps to present disordered cases. As its value increases, the situation of system or signal moves from simple to complex. KFD is calculated by Equation (12) for a signal. Here *L* is the

sum of the differences between the successive signal samples, d is the largest difference between the signal's first sample value and the other samples values, and N is the number of the total samples in the signal

$$KFD = \frac{\log(N)}{\log(d/L) + \log(N)}. \tag{12}$$

2.4.1. Recurrence Plots (RPs) Quantities

RPs is a nonlinear analysis method that comes from information theory [29]. To analysis the signal with RPs, firstly reconstruction the state space reconstruction is employed by using the method of time-delay. For time delay embedding, the time-delayed vector in the m dimensional state space, is given in Equation (13) which includes the embedding dimension (k) and time delay (lag) (T)

$$\vec{X}_n = \left(x_n, x_{n-1}, \dots, x_{n-(k-1)T}\right) \epsilon R^k . \tag{13}$$

RPs visualize the dynamic structure of the signal. It is used to determine similar situations in the signal and to obtain recurring patterns. RP quantities are metrics that identify the visual information provided by RPs. These quantities give the properties of repeating patterns and are explained below [29].

Recurrence Rate (REC): REC quantifies the frequency of recurring patterns in the signal [29]. For obtaining REC given in Equation (14), calculate the distances between time delayed vectors in state space and it is checked whether these distances are below or above a threshold value. Distances below the threshold indicate the existence of a repeating case, so, the same or similar pattern is repeated over a period. REC gives the frequency of these recurring patterns. A higher value of REC says regularities, otherwise irregularities. In Equation (14), $R(i,j)$ is the number of recurrent points

$$REC = \frac{1}{N^2}\sum_{i,j=1}^{N} R(i,j) . \tag{14}$$

Determinism (DET): DET metric indicates the extent to what recurrent points create repeated data sequences, disclosing the distinctive structure of the signal [29]. DET gives only the ratio of repeated points located along diagonal line, not on the main crossline. But diagonal lines that do not valid a specific length criterion are not considered. The mathematical formula of DET is presented in Equation (15), in here, $P(l)$ is histogram of the diagonal lines' lengths l, and N is the number of time-delayed vectors on the trajectory in the state space

$$DET = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{i=1}^{N} lP(l)}. \tag{15}$$

Entropy (ENT): ENT identifies the irregularity of the dynamic structure within the signal and given in Equation (16), $P(l)$ and l are explained above [29]

$$ENT = -\sum_{l=l_{min}}^{N} p(l)\, ln\, p(l) . \tag{16}$$

Maximal Diagonal Line Length (LMAX): LMAX is a metric of mathematical stability which is related to the initial conditions of dynamical system and represents the length of the longest diagonal line in the recurrence graphic given as Equation (17) [29]. The lower value of LMAX shows chaotic behavior of the signal. In Equation (17), l_i is diagonal line, N_l is the diagonal lines number at recurrence plot

$$LMAX = max(\{l_i; i = 1, \dots, N_l\}) . \tag{17}$$

2.5. Feature Selection

Feature selection approach is used for efficient classification. In this procedure, expressive features are selected, and others are removed. Three different methods of Relief, Minimum Redundancy Maximum Relevance (MRMR) and one-way Analysis of Variance (ANOVA) algorithms were used in this study.

Relief: Relief is a weighting or ranking algorithm used for feature selection for classification [30]. In this algorithm, the difference between two similar samples is used for determining the weight of each feature. Relief ranks the importance of features by evaluating the obtained weights. Features with higher weights contribute to greater classification performance so they are considered more important.

ANOVA: One way ANOVA is a statistical method used for comparing means across three or more groups. It determines whether there exists a statistically significant difference among the groups [31]. To make the variance analysis, each group has a normal distribution, and the samples are independent, and their variances are equal. ANOVA uses F statistics which expresses the ratio of between group variance to within group variance. F statistics evaluates whether this ratio shows a significant difference. If the value of F exceeds a significant level, the distributions of the groups have significant differences.

Minimum Redundancy Maximum Relevance (MRMR): MRMR is a filtering method that treats each feature and class labels as discrete random variables and uses mutual information to obtain the relationship between features and classes. It calculates how much each feature contributes to the classification and how similar it is to others. The selected features are highly correlated with classes, and low in redundancy with each other [32].

2.6. Classification Methods

Classification algorithms determine to identify which class an unknown pattern belongs to with a decision approach using the properties of that pattern. Classification success depends on features and classification methods. In general, the aim is to increase classification success by using different or new features and suitable classifiers. There are many classification methods in literature and commonly used of them are KNN, SVM, ANN, etc. in voice studies. Therefore, in this study, Decision Trees (DTs), Naive Bayes (NB) and Ensemble Learning (EL) methods were evaluated to observe the performances of different classifiers.

DTs have an approach that develops a tree structure including decision and leaf nodes based on features, and targets, for classification [33]. The DTs algorithm starts with root node on the tree and evolves recursively, so a decision node may have one or more branches. In DTs, the gain of information by using entropy which indicates uncertainty, and the potential for unexpected outcomes, is calculated for each feature to identify the most distinctive one. NB classifier is a simple and probability based (Bayes Theorem) and especially effective on complex data [34]. NB works with independence assumption so, the relationships between features are not considered. Therefore, NB independently evaluates the contribution of each feature to the classification and is suitable for high dimensional feature vectors, and the distribution of features can affect the classification performance. EL evaluates different algorithms, combines these algorithms to create a powerful model and give better results than these methods which are used separately [35]. EL includes bagging (bootstrap), boosting and stacking strategies. Bagging method trains separately multiple models by using random subgroups (bootstrap samples) of data set, and then the results are combined (such as Random Forest). Boosting approach (for example AdaBoost and Gradient Boosting) develops a strong model by combining weak models. Each weak model is trained according to the errors obtained from previous models. Stacking brings together multiple different models by combining the strengths of them, and a main model decides final predictions using the other's predictions.

The metrics used for evaluating the classifier performance are: True Positive Rate (TPR, sensitivity, recall), True Negative Rate (TNR, specificity), Positive Predictive Value (PPV, precision), Accuracy (ACC), F1 score (F1), and Area Under the Curve (AUC). They are calculated by using the Equations (18) to (22), where true positives (TP, correctly identified) and true negatives (TN, correctly rejected) are the numbers

of intances accurately predicted, and false positives (FP, incorrectly identified) and false negatives (FN, incorrectly rejected) refer to the numbers of instances wrongly predicted

$$TPR = \frac{TP}{TP + FN} \tag{18}$$

$$TNR = \frac{TN}{TN + FP} \tag{19}$$

$$PPV = \frac{TP}{TP + FP} \tag{20}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{21}$$

$$F1 = \frac{2 \times TPR \times PPV}{TPR + PPV}. \tag{22}$$

In this study, the statistical comparison of the results of classifiers was done by using McNemar's test for binary class [36]. McNemar's statistics, which is a non-parametric test, compares the two binary classification models performances by determining the disagreement between the predictions of these two models using a same dataset having N instances. For implementation of McNemar's test, 2×2 contingency matrix (Table 2) is constructed for two models (M1 and M2). N_{M1M2} shows the numbers of instances predicted by the M1 and M2 models, where 1 represents true and 0 is false. (so; N_{11}:M1(1) and M2(1), N_{10}:M1(1) and M2(0), N_{01}:M1(0) and M2(1), N_{00}:M1(0) and M2(0)). McNemar's test statistic x^2 is calculated by using Equation (23) and then corresponding *p*-value is found by using chi-square distribution table with one degree of freedom (df). Thus, the hypotheses "H0 (null hypotesis): M1 and M2 are the same", and "H1: M1 and M2 are different" are tested for a certain statistical significance level of *p*.

$$x^2 = \frac{(|N_{10} - N_{01}| - 1)^2}{(N_{10} + N_{01})} \tag{23}$$

Table 2. *2×2 Contingency table*

M1/M2	M2 model true (1)	M2 model false (0)	Total
M1 model true (1)	N_{11}	N_{10}	$N_{11}+ N_{10}$
M1 model false (0)	N_{01}	N_{00}	$N_{01}+ N_{00}$
Total	$N_{11}+ N_{01}$	$N_{10}+ N_{00}$	$N= N_{11}+ N_{10}+ N_{01}+ N_{00}$

3. RESULTS

In this study, a total of 60 voice samples of /a/ vowel were analysis, including 15 healthy, 15 cyst (VC), 15 polyp (VFP) and 15 sulcus (SV). After the framing process, 3200 voice segments were obtained from each group, so 12800 segments were used in total. 10 features (LEE, STE, SE, STZCR, HFD, KFD, REC, DET, ENT, LMAX) were calculated and some of them selected by using three different algorithms in feature selection step. And then two classification procedures were realized with three different classification methods: binary and multiclass. The classification schedules are presented in Figure 3. In the classification process, the data set was divided into two separate groups as test and training. Two randomly selected subjects from each group were used for testing, and the other 13 subjects were used for training. Therefore, the test data consisted of 458 segments obtained from each group (so 458 healthy and 1374 disordered, totaling 1832 segments) and the training data consisted of 2742 segments obtained from each group (so 2742 healthy and 8226 disordered, totaling 10968 segments). All models of three classification approaches were applied with the MATLAB R2022b Classification Learner with 10-fold cross validation. The test results were given for the models with the highest accuracy values.

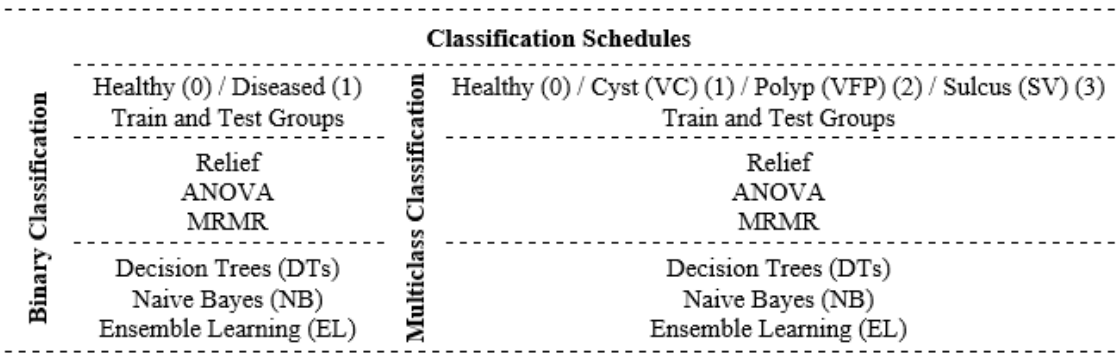


Figure 3. *Classification procedures in this study*

Binary Classification: For this application, all pathological groups were evaluated as a disordered group. So healthy/diseased classification was made and the classification performances in testing with only 4 selected features (in order of importance) are given in Table 3. The mean ± Standard Deviation (SD) values of selected features for healthy and diseased groups are given in Table 4. According to Table 3, the difference between the LMAX and REC values obtained for D6 components of the signal for the healthy and diseased groups is remarkable (Tables 3 and 4). REC quantity gives about the regularity based on repeated pattern in the signal. LMAX represents the stability of the signal. These two quantities have lower values in healthy case with respect to the disordered situation (Table 4). These results can be interpreted as the regularity and stability of pathological voice signals is higher than healthy voices.

Table 3. *The binary classification test results*

Feature Selection	Classification Method	ACC (%)	TPR (%)	TNR (%)	PPV (%)	F1 (%)	AUC
Relief (LMAX, SE, HFD, STE)	Coarse Tree	85.9	97.1	52.2	85.9	91.2	0.89
	Kernel Naive Bayes	85.7	73.1	89.9	70.7	71.9	0.92
	Subspace Discriminant Ensemble	78.8	15.3	100	100	26.5	0.97
ANOVA (LMAX, REC, LEE, STE)	Coarse Tree	89.4	86.2	98.7	99.5	92.4	0.92
	Kernel Naive Bayes	83.2	77.5	92.4	77.3	77.4	0.95
	Subspace Discriminant Ensemble	88.5	63.3	96.9	92.9	73.4	0.94
MRMR (LMAX, REC, KFD, STE)	Coarse Tree	**99.2**	**98.3**	**99.6**	**99.4**	**98.5**	**0.99**
	Kernel Naive Bayes	86.6	96.3	83.3	65.8	78.2	0.83
	Subspace Discriminant Ensemble	87.5	50	100	100	66.7	0.99

Table 4. *The mean ± SD values of selected features for healthy and diseased groups*

Feature	Healthy	Diseased
$LMAX_{(D6)}$	550.89±248.19	1193.94±391.19
$REC_{(D6)}$	4.03±0.78	7.31±1.62
$LEE_{(D4)}$	-14233.4±781.25	-14824.8±2408.11
$STE_{(D4)}$	0.389±0.268	0.215±0.341
$KFD_{(D4)}$	1.00±1.05E-06	1.00±4.78E-07
$HFD_{(A5)}$	1.026±0.0025	1.014±0.0035
$SE_{(D4)}$	0.624±0.026	0.638±0.027

According to the test results of binary classification (Table 3), Coarse Tree showed the highest accuracy rate (99.2%), with only 4 features selected by MRMR feature selection algorithm. The confusion matrix for this best classification test outcome is presented in Figure 4. It is shown that 6 diseased cases (1) were classified as healthy (0) and 8 healthy (0) cases were classified as disorder (1) (Figure 4).

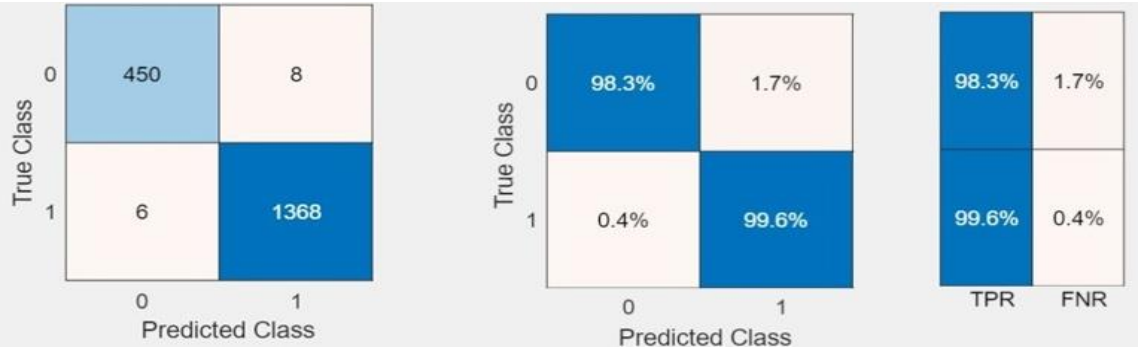


Figure 4. *The test confusion matrix for binary class with Coarse Tree and MRMR algorithm (0: healthy, 1: diseased)*

The McNemar's test was applied to see whether the results obtained by machine learning methods for binary classification are statistically significant at the level of $p<0.05$. The results of this test made for the classifications with MRMR in Table 3 are presented in Table 5 and the classifiers having higher accuracy are indicated with arrows. In Table 5, all p values are statistically significant, so null hypothesis (H0) is rejected. Hence, these classifiers are statistically different each other, in other words, the variations in their performances are not random. According to the results in Table 5, the performance of Coarse Tree model has the superior significant.

Table 5. *McNemar's test results (p values)*

Classifier	Coarse Tree	Kernel Naive Bayes	Subspace Discriminant Ensemble
Coarse Tree		5.95×10^{-52} ←	3.04×10^{-48} ←
Kernel Naive Bayes			0.0001 ↑
Subspace Discriminant Ensemble			

Multiclass classification: In multiclass, four class classification (healthy and three disordered groups) was conducted. The test results for three classification algorithms by using only 7 or 8 selected features (in order of importance) are given in Table 6. The mean ± SD values of selected features for healthy and diseased groups are also presented in Table 7. As we analyze Table 7, it is seen that, especially the ENT, LMAX and REC values which are calculated for D6 coefficients are effective in distinguishing groups from each other, and so healthy and polyp voices have low level stability than others.

In multiclass, the Kernel Naïve Bayes classifier showed a 91.6% best accuracy rate by using only 8 features selected with ANOVA, and the confusion matrix belongs to this result is given in Figure 5. It is understood that from Figure 5, the cyst disorder was detected with a 99.8% accuracy, showing the highest success rate compared to other groups. This rate is followed by polyp with 99.3%, healthy with 97.2%, and sulcus with 70.3%, so 135 sulcus data were classified as cyst disorder (Figure 5).

Table 6. *The multiclass classification test results*

Feature Selection	**Classification Method**	**ACC (%)**	**TPR (%)**	**TNR (%)**	**PPV (%)**	**F1 (%)**	**AUC**
Relief	Coarse Tree	55.9	56.0	85.3	56.4	56.2	0.85

(ENT, LMAX, REC, STZCR, SE, STE, HFD)	Kernel Naive Bayes	76.9	76.9	92.4	77.4	77.1	0.81
	Subspace Discriminant Ensemble	83.8	83.8	94.5	86.7	85.2	0.98
ANOVA (ENT, LMAX, DET, REC, STZCR, SE, STE, HFD)	Coarse Tree	89.2	89.3	96.4	92.3	90.8	0.96
	Kernel Naive Bayes	**91.6**	**91.7**	**97.2**	**93.4**	**92.5**	**0.96**
	Subspace Discriminant Ensemble	75.4	71.7	90.7	75.9	73.7	0.93
MRMR (ENT, LMAX, REC, SE, STE, HFD, KFD)	Coarse Tree	88.8	75.0	96.2	92.2	82.7	0.96
	Kernel Naive Bayes	79.5	79.5	93.0	78.8	79.1	0.84
	Subspace Discriminant Ensemble	75.4	57.6	93.5	52.7	55.0	0.89

Table 7. *The mean ± SD values of selected features for groups in multiclass classification*

Feature	Healthy	Cyst	Polyp	Sulcus
ENT$_{(D6)}$	4.7821±0.11	4.9582±0.15	4.7975±0.16	4.9814±0.21
LMAX$_{(D6)}$	550.89±248.19	1388.50±372.94	611.39±498.85	1581.92±301.77
REC$_{(D6)}$	4.03±0.78	8.54±1.26	4.47±1.46	8.72±2.14
DET$_{(D6)}$	97.84±0.19	98.03±0.27	98.93±0.25	99.28±0.29
STZCR$_{(A5)}$	0.0246±0.0035	0.0223±0.0038	0.0208±0.0053	0.0202±0.0053
SE$_{(D4)}$	0.6244±0.0256	0.6217±0.0291	0.6025±0.0269	0.6086±0.0251
STE$_{(D4)}$	0.3885±0.2676	0.2254±0.2874	0.1311±0.2481	0.2894±0.4873
HFD$_{(A5)}$	1.0261±0.0025	1.0144±0.0031	1.0135±0.0032	1.0129±0.0041
KFD$_{(D4)}$	1.00±1.05E-06	1.00±1.26E-06	1.00±4.41E-07	1.00±8.68E-07
LEE$_{(D4)}$	-20658±2118.89	-19950.3±3080.73	-19298±3855.22	-17925.9±122.7

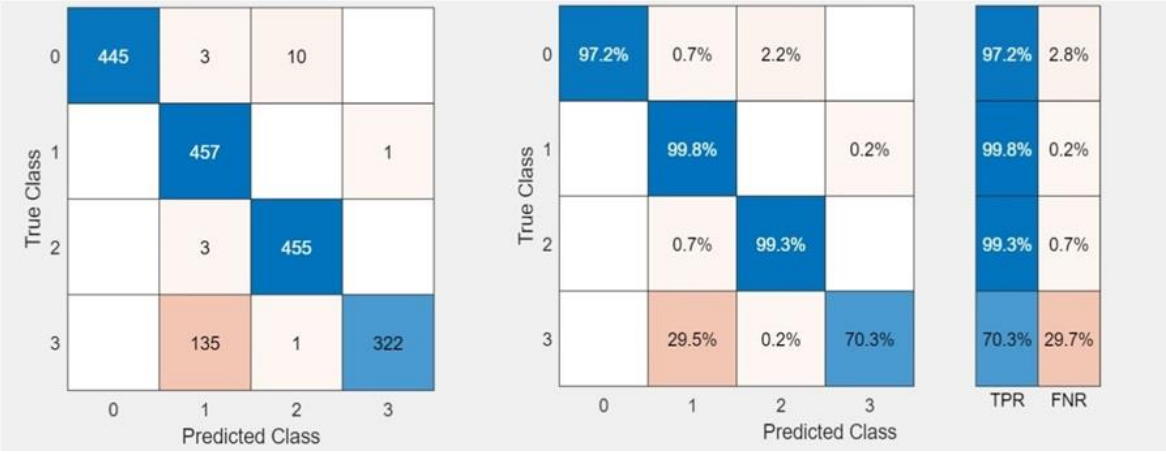


Figure 5. *The test confusion matrix for multiclass with Kernel Naive Bayes and ANOVA algorithm (0: healthy, 1: cyst, 2: polyp, 3: sulcus)*

4. CONCLUSIONS

In this study, the Coarse Tree showed a 99.2% high accuracy rate with only 4 features selected by MRMR feature selection algorithm, in binary classification performed to detect healthy and diseased (all disorders) voice data. In multiclass classification implemented for distinguishing healthy, cyst, polyp, and sulcus voice disorders by using only 8 selected features with ANOVA, the Kernel Naive Bayes classifier obtained the highest success rate at 91.6% accuracy.

The study has differences in terms of the used features, classifiers and analysis approach compared to the studies conducted on different data sets of voice diseases in the literature. In this study, features from information theory and nonlinear approaches were selected, while features obtained from acoustic voice analysis were generally used in literature. The classification performances obtained for both classification procedures, especially in binary classification, with a very small number of features are quite high compared to the results of previous studies where a lot of features (generally above the 20 features) were used [5-17]. Our study is the first to use RPs metrics in the detection of voice disorders for VD dataset. According to the results of the only study in the literature using the VD database used in this study, the training accuracies were obtained as 99.8% and %90.15 for binary (healthy/unhealthy) and multiclass classifications, respectively [19]. In [19], 21 features and also the raw signals were used together. LSTM and CNN were used for two feature fusion methods (DLF and FLF). In addition, data was not split between train and test groups, so distinct test data was not used, and the test performances were not given [19]. This case is one of the factors that increase classification accuracy. The binary classification test accuracy of our study (99.2%) is approximately same the result in [19]. The multiclass classification test result obtained in our study was 91.6%, and higher than the multiclass result obtained in [19]. Considering all these situations, the use of a test group in our study made the results more realistic and these results were achieved with very few features obtained with a simple and new approach (RPs). It is considered that our study has the important outcomes.

In the literature, RPs measurements for voice signals (for Disordered Voice Database obtained by Kay Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Lab) were used only in Vieira et al.'s study in 2023 [17]. In their study, they showed that the use of RP features improves classification accuracy compared to the classification results obtained with traditional acoustic features. Similarly, in our study, it is shown that the quantities obtained from RP analysis give some useful/efficient information for identifying the voice and distinguishing the voice disorders. In this context, RP quantities have the potential to be used in other areas related to voice, such as voice/speech detection and recognition [37, 38].

One of the limitations of this study is the small number of subjects in the groups, particularly in multi-class classification. Repeating the study with larger datasets constructed by increasing number of subjects, would provide useful results for validating the method. Due to the challenges of collecting diseased voice data, the use of data augmentation techniques may be considered. While the dataset obtained from Turkish speakers is a considerable feature of the study, it is important to evaluate the method for other languages to investigate its widespread use. This study used samples of the /a/ voice, which is routinely used in voice analysis studies and three voice disorders (except healthy) are analyzed. The proposed method's performance could also be examined for different samples of vowel/consonant voices and speech to achieve more effective outcomes and could be investigated for other voice diseases.

In conclusion, our results are valuable in showing that nonlinear analysis approaches provide effective solutions in revealing the characteristic features of voice, especially disordered cases. Future studies could include evaluating the proposed method with DNNs to obtain performance-enhancing results. As a result of examining the performance of the proposed method on large data sets, population and data diversity, it can be said that it has the potential to be incorporated into voice analysis programs to be developed for the diagnosis and monitoring of diseases for experts and thus could be adapted in clinical use. Additionally, the use of the proposed method in automatic voice disease diagnosis, voice recognition, voice-activated systems, etc. could yield beneficial results.

ACKNOWLEDGMENTS

This study is derived from a master's thesis titled "Detection of voice diseases with signal processing methods and machine learning" in the Program of Electrical and Electronics Engineering, Graduate School of Natural and Applied Sciences, Gazi University.

The authors would like to thank Prof. Dr. M. Akif Kılıç for giving permission the using data.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] Bhattacharyya, N., "The prevalence of voice problems among adults in the United States", The Laryngoscope, 124: 2359-2362, (2014). DOI: https://doi.org/10.1002/lary.24740

[2] Hossain, M.S., Muhammad, G., Alamri, A., "Smart healthcare monitoring: A voice pathology detection paradigm for smart cities", Multimedia Systems, 25: 565-575, (2017). DOI: https://doi.org/10.1007/s00530-017-0561-x

[3] Kılıç, M., "Ses problemi olan hastanın objektif ve subjektif yöntemlerle değerlendirilmesi", Kulak Burun Boğaz Baş Boyun Cerrahisi'nde Güncel Yaklaşım, 6, (2010).

[4] Hegde, S., Shetty, S., Rai, S., Dodderi, T., "A survey on machine learning approaches for automatic detection of voice disorders", Journal of Voice, 33(6): 947.e11-947.e33, (2019). DOI: https://doi.org/10.1016/j.jvoice.2018.07.014

[5] Mohamed, A.N., Salama, A.A., Darwish, S.H., "Classification of voice pathology using SVM classifier", International Telecommunications Conference (ITC-Egypt), Alexandria, Egypt, 419-422, (2023). DOI: https://doi.org/10.1109/ITC-Mısır58155.2023.10206116

[6] Gayathri, S., Priya, E., "Identification of voice pathology from temporal and cepstral features for vowel 'a' low intonation", International Conference on Electronic Systems and Intelligent Computing (ICESIC), Chennai, India, 345-350, (2022). DOI: https://doi.org/10.1109/ICESIC53714.2022.9783484

[7] Murthy, G.L.N., Iswarya, V., Ramya Sri, K., Harshitha, K., Prasanth, Ch., "A novel algorithm for detecting spasmodic dysphonia voice pathology using random forest framework", International Conference on Edge Computing and Applications (ICECAA), Tamilnadu, India, 800-804, (2022). DOI: https://doi.org/10.1109/ICECAA55415.2022.9936567

[8] Miliaresi, I., Poutos, K., Pikrakis, A., "Combining acoustic features and medical data in deep learning networks for voice pathology classification", 28th European Signal Processing Conference (EUSIPCO), Amsterdam, Netherlands, 1190-1194, (2021). DOI: https://doi.org/10.23919/Eusipco47968.2020.9287333

[9] Ribas, D., Pastor, M.A., Miguel, A., Martinez, D., Ortega, A., Lleida, E., "Automatic voice disorder detection using self-supervised representations", IEEE Access, 11: 14915-14927, (2023). DOI: https://doi.org/10.1109/ACCESS.2023.3243986

[10] Barche, P., Gurugubelli, K., Vuppala, A.K., "Comparative study of different epoch extraction methods for speech associated with voice disorders", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada, 6923-6927, (2021). DOI: https://doi.org/10.1109/ICASSP39728.2021.9413518

[11] AnilKumar, V., Venkata Siva Reddy, R., "Classification of voice pathology using different features and Bi-LSTM", 2023 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES), Tumakuru, India, 1-4, (2023). DOI: https://doi.org/10.1109/ICSSES58299.2023.10200529

[12] Ali, Z., Herrera, A.G.S., Mesallam, T.A., Muhammad, G., "Computer-based blind diagnostic system for classification of healthy and disordered voices", IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS), L'Aquila, Italy, 654-659, (2023). DOI: https://doi.org/10.1109/CBMS58004.2023.00295

[13] Wang, C.T., Chuang, Z.Y., Hung, C.H., Tsao, Y., Fang, S.H., "Detection of glottic neoplasm based on voice signals using deep neural networks", IEEE Sensors Letters, 6(3): 1-4, (2022). DOI: https://doi.org/10.1109/LSENS.2022.3152738

[14] Suresh, M., Rajan, R., Thomas, J., "Dysarthria speech disorder classification using traditional and deep learning models", Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichirappalli, India, 01-06, (2023). DOI: https://doi.org/10.1109/ICEEICT56924.2023.10157285

[15] Sahane, P., Pangaonkar S., Khandekar S., "Dysarthric speech recognition using multi-taper Mel frequency cepstrum coefficients", International Conference on Computing, Communication and Green Engineering (CCGE), Pune, India, 1-4, (2021). DOI: https://doi.org/10.1109/CCGE50943.2021.9776318

[16] Rebecca, M.G., Anjana, A., Arthi, V., Dhanalakshmi, M., Babu, R.D., "Machine learning based vowel analysis in severity level detection of voice pathology", 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 1005-1008, (2022). DOI: https://doi.org/10.1109/ICECA55336.2022.10009604

[17] Vieira, V.J.D., Costa, S.C., Correia, S.E.N., "Non-stationarity-based adaptive segmentation applied to voice disorder discrimination", IEEE Access, 11: 54750-54759, (2023). DOI: https://doi.org/10.1109/ACCESS.2023.3281191

[18] Islam, R., Tarique, M., "Robust assessment of dysarthrophonic voice with RASTA-PLP features: A nonlinear spectral measures", 2nd International Conference on Mechatronics and Electrical Engineering (MEEE), Abu Dhabi, United Arab Emirates, 74-78, (2023). DOI: https://doi.org/10.1109/MEEE57080.2023.10126695

[19] Ankışhan, H., İnam, S.Ç., "Voice pathology detection by using the deep network architecture", Applied Soft Computing, 106, (2021). DOI: https://doi.org/10.1016/j.asoc.2021.107310

[20] Mallat, S.G., "A theory for multiresolution signal decomposition: The wavelet representation", IEEE Transactions on Pattern Analysis and Machine Intelligence, 11: 674-693, (1989). DOI: https://doi.org/10.1109/34.192463

[21] Mihoc, S.G., Popov, A.N., Ivanov, R., "Denoising speech signals by wavelet transform", Annual Journal of Electronics, (2009).

[22] Dautov, Ç.P., Özerdem, M.S., "Wavelet transform and signal denoising using wavelet method", 26th Signal Processing and Communications Applications Conference (SIU), (2018). DOI: https://doi.org/10.1109/SIU.2018.8404418

[23] Daubechies, I., "Orthonormal bases of compactly supported wavelets", Communications on Pure and Applied Mathematics, 41: 909-996, (1988). DOI: https://doi.org/10.1002/cpa.3160410705

[24] Shannon, C.E., "Mathematical theory of communication", Bell Labs Technical Journal, 27: 379-423, 623-656, (1948). DOI: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[25] Lau, Y.K., Chan, C.K., "Speech recognition based on zero crossing rate and energy", IEEE Transactions on Acoustics, Speech, and Signal Processing, 33(1): 320-3233, (1985). DOI: https://doi.org/10.1109/TASSP.1985.1164503

[26] Misra, H., Ikbal, S., Bourlard, H., Hermansky, H., "Spectral Entropy Based Feature for Robust ASR", IEEE International Conference on Acoustics, Speech, and Signal Processing, (2004). DOI: https://doi.org/10.1109/ICASSP.2004.1325955

[27] Higuchi, T., "Approach to an irregular time series on the basis of the fractal theory", Physica D, 31: 277-283, (1988). DOI: https://doi.org/10.1016/0167-2789(88)90081-4

[28] Katz, M.J., "Fractals and the analysis of waveforms", Computers in Biology and Medicine, 18(3): 145-56, (1988). DOI: https://doi.org/10.1016/0010-4825(88)90041-8

[29] Norbert Marwan, M., Romano, C., Thiel, M., Kurths, J., "Recurrence plots for the analysis of complex systems", Physics Reports, 438: 237-329, (2007). DOI: https://doi.org/10.1016/j.physrep.2006.11.001

[30] Robnik-Sikonja, M., Kononenko, I., "Theoretical and empirical analysis of ReliefF and RReliefF", Machine Learning, 53: 23-69, (2003). DOI: https://doi.org/10.1023/A:1025667309714

[31] Hogg, R.V., Ledolter, J., "Engineering Statistics", New York: MacMillan, (1987).

[32] Darbellay, G.A., Vajda, I., "Estimation of the information by an adaptive partitioning of the observation space", IEEE Transactions on Information Theory, 45(4): 1315-1321, (1999). DOI: https://doi.org/10.1109/18.761290

[33] Quinlan, J.R., "Induction of decision trees", Machine Learning, 1: 81-106, (1986). DOI: https://doi.org/10.1007/BF00116251

[34] Manning, C.D., Raghavan, P., Schütze, H., "Introduction to Information Retrieval", NY: Cambridge University Press, (2008). DOI: https://doi.org/10.1017/CBO9780511809071

[35] Polikar, R., "Ensemble based systems in decision making", IEEE Circuits and Systems Magazine, 6(3): 21-45, (2006). DOI: https://doi.org/10.1109/MCAS.2006.1688199

[36] Atilkan, Y., Kirik, B., Acici, K., Benzer, R., Ekinci, F., Guzel, M.S., Benzer, S., Asuroglu, T., "Advancing crayfish diseased detection: A comparative study of deep learning and canonical machine learning techniques", Applied Sciences, 14, 6211, (2024). DOI: https://doi.org/10.3390/app14146211

[37] Tombaloğlu, B., Erdem, H., "Turkish speech recognition techniques and applications of recurrent units (LSTM and GRU)", Gazi University Journal of Science, 34(4): 1035-1049, (2021). DOI: https://doi.org/10.35378/gujs.816499

[38] Disken, G., "Robust spoofed speech detection with denoised I-vectors", Gazi University Journal of Science, 36(4): 1553-1561, (2023). DOI: https://doi.org/10.35378/gujs.1062788