IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 September 2023, accepted 18 October 2023, date of publication 26 October 2023, date of current version 3 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3327902

RESEARCH ARTICLE

A Preliminary Study on OSA Severity Levels Detection by Evaluating Speech Signals Nonlinearities With Multi-Class Classification

TUĞÇE KANTAR UĞUR[1], **(Member, IEEE), DERYA YILMAZ**[2], **METİN YILDIZ**[3], **AND SİNAN YETKİN**[4]

[1]Department of Biomedical Engineering, Faculty of Engineering, Başkent University, 06790 Ankara, Turkey
[2]Department of Electrical and Electronics Engineering, Faculty of Engineering, Gazi University, 06570 Ankara, Turkey
[3]Department of Biomedical Engineering, Faculty of Engineering, İzmir Democracy University, 35140 İzmir, Turkey
[4]Department of Psychiatry Clinic, Gülhane Training and Research Hospital, University of Health Sciences, 06018 Ankara, Turkey

Corresponding author: Tuğçe Kantar Uğur (tkantar@baskent.edu.tr)

This work was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) through the 1001 Program under Project 121E496.

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the ethics committee of Gülhane Training and Research Hospital (document number 46418926/decision number 19/87).

ABSTRACT Diagnosis of obstructive sleep apnea (OSA) from speech has become a popular research area in recent years, which can be an alternative way to the application difficulties in polysomnography (PSG). The promising results obtained in our previous study, in which we tried to detect apnea using nonlinear analysis of speech, gave rise to the thought that it is possible to detect OSA and OSA severity by diversifying speech samples and nonlinear features. The principal aim of this study, for the first time in the literature, is to detect the OSA severity levels as mild, moderate, and severe as in the clinic use (multi-class classification) using nonlinear analyses of speech while the patient is awake. In addition, healthy/OSA classification (binary classification) was also carried out. The feature selection method of ANOVA was applied to 336 features (28 voices × 12 features) for each subject, 14 and 5 features were used in multi-class and binary classifications, respectively. As a result of the classifications made with various KNN and SVMs models, the best results were obtained by SVMs in both classifications for OSA severities (with one-vs-all classification scheme and the Gaussian kernel) and OSA detection (with the quadratic kernel) as 82% and 95.1% accuracies, respectively. The proposed study showed that OSA and OSA severity can be determined with the small number of nonlinear features calculated from a few different speech samples, in nearly 15 minutes, consistent with PSG results (simple snorer, mild, moderate, and severe OSA). In conclusion, the highest OSA/healthy classification accuracy rate in the literature was achieved. Furthermore, OSA severity detection in four-class performed quite well as a preliminary study.

INDEX TERMS AHI level, multi-class classification, nonlinear analysis, obstructive sleep apnea, OSA severity, speech.

I. INTRODUCTION

Obstructive sleep apnea (OSA), one of the most common sleep diseases, is defined as the cessation of breathing for more than 10 seconds during sleep and can cause problems up to death in sleep. Polysomnography (PSG) examination is the gold standard diagnostic method for OSA [1]. The diagnosis of OSA is made by recording many physiological parameters from the patients during a night's sleep and evaluating these records. The PSG method has several challenges for the

The associate editor coordinating the review of this manuscript and approving it for publication was Santosh Kumar.

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

hospital and the patient [1]. The biggest challenge among these is that the patient has to sleep for long hours in a sleep lab with multiple electrode connections.

There have been studies attempting to diagnose OSA with fewer signal recordings. For this purpose, OSA has been tried to be detected by using ECG, respiratory signals, SPO2, and snoring/breathing sounds which recorded outside the hospital environment [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18]. Although these studies have yielded good results for the detection of OSA/OSA severity, the cost of detecting OSA could not be reduced to the desired level due to the recording time requiring a night's sleep, the need for multiple signal recordings, and the long-term complex evaluation.

On the other hand, deformations on the vocal tract of OSA patients have been shown to affect speech, and recently studies trying to reduce the diagnosis cost (shortening of diagnosis time and easy application) of OSA have started to focus on speech/voice analysis [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32].

In the apnea detection study by Goldshtein et al. [23], the vowels of /a/, /e/, /i/, /o/, /u/ and nasal consonants of /n/ and /m/ were analyzed by using many features coming from the voice analysis and Gaussian mixture model (GMM), 79% and 84% sensitivities, and 83% and 86% specificities were achieved for male and female respectively. Some studies researched the performance of different classifiers and the effect of recording position on acoustic features for detecting apnea [24], [25]. In another study conducted on the detection apnea-hypopnea index (AHI) level and OSA, speech and respiratory sounds were used together, and 75%, 79%, and 77.14% sensitivity, specificity, and accuracy performances were achieved on average [26]. Espinoza-Cuadros et al. [27] realized the analysis of sound together with image and reached 79.4% accuracy for OSA detection. The researchers obtained a sensitivity of 92.92%, but in another study, they obtained a specificity of 20.6% and a correct recognition rate of 71.06% for apnea detection [28]. Blanco et al. [29] used Mel-Frequency cepstral coefficients (MFCCs) and some nonlinear features with the subject groups of AHI>30 (OSA) and AHI<10 (control). Their OSA detection performances were 86.7%, 90.6%, and 88.5% sensitivity, specificity, and accuracy, respectively. In Ding et al.'s study, [30] linear prediction cepstral coefficients (LPCCs) and support vector machines (SVMs) were used for AHI=30 and AHI=10 thresholds and different recording positions such as sitting and lying. They achieved 78.8% OSA classification accuracy in both cases at the highest. In another study by Ding et al. [31], the severity of OSA was determined by evaluating the Chinese syllables for AHI=10 and AHI=30 threshold values and using the 30 LPCCs and decision tree classifier. They reached the accuracies of 81.7% and 80.3% for AHI thresholds of 10 and 30, respectively. In our previous study, for the first time in the literature, OSA detection was realized with only nonlinear analysis features such as the largest Lyapunov exponent, time delay, and embedding dimension from speech [32]. As speech samples, 4 vowels and 24 syllables (6 consonants with 4 vowels) were analyzed, and binary classification (healthy/OSA) was made with SVMs and K-nearest neighbor (KNN) classifiers. OSA was detected using only 6 features derived from consonants, and SVMs with 100%, 65%, and 82.5% sensitivity, specificity, and accuracy, respectively [32]. The most important difference of this study from our previous study is that the classification process was carried out in a multi-class manner as simple snoring, mild OSA, moderate OSA, and severe OSA, as in its clinical use (the same in diagnosis procedure). In addition, while the healthy group in our previous study consisted of subjects who were not included in the PSG test, all subjects were applied PSG, in this study. Along with the multi-class (four-class) classification procedure, the healthy/OSA classification (binary classification) was also made in our study. In this procedure, subjects who were diagnosed with simple snoring by PSG test were considered healthy.

Studies in this field have generally used traditional acoustic analysis methods and a small variety of voice/speech samples. It is necessary to perform new studies in OSA/OSA severity detection on different analysis methods and many speech samples in order to see the potential in this field. In this study, considering our previous study [32] that revealed the presence of chaotic behavior in the structure of the voice, we focused on the determination of the OSA severity levels as the same in clinical use (simple snoring, mild, moderate, severe) by using the features under the nonlinear analysis. For the first time in the literature, multi-class classification is used to diagnose the OSA severities with speech signals from the consonant voices affected by OSA. Additionally, a binary classification (OSA/healthy) procedure was also implemented for comparing with literature.

II. MATERIAL AND METHODS

A. SPEECH SAMPLES

Speech occurs as a result of the compressed air coming from the lungs being shaped by structures such as vocal cords, larynx, mouth, teeth, lips, and nose located on the vocal tract. Vocal cords, tongue, chin, and lips are effective in the formation of vowels. Unvoiced consonants are formed by the throat, mouth, teeth, lips, and nose in the vocal tract, and the vocal cords are open, while the vocal cords are also effective in voiced consonants. In this study, some vowels and consonants in the Turkish alphabet were used since the speech recordings were obtained from native Turkish speakers.

In this study, /n/ from the nasals of /m/ and /n/ used in previous studies was selected. Among the voices /b/, /d/, /g/, /k/, /p/, /t/ produced by the tongue and posterior palate and come out of the mouth in the form of sudden bursts, /g/and /k/ were chosen. The voices of /c/, /ç/, /h/, /s/, /ş/, /v/, /y/, and /z/ are produced by the friction of air to the vocal tract (larynx, palate, uvula, tongue root). Additionally, to analyze frictional voices, unvoiced consonant of /s/ (from /ç/, /f/, /h/, /k/, /p/, /s/, /ş/, /t/), and voiced consonant of /c/

and /r/ (from /b/, /c/, /d/, /g/, /j/, /l/, /m/, /n/, /r/, /v/, /y/, /z/) were also selected. Considering that the subjects would have difficulties in vocalizing the consonants, they were turned into syllables with the vowels /a/, /i/, /ı/, /u/ for recording. Among the 8 vowels in the Turkish alphabet, the vowels /i/ and /ı/, in which the front and middle parts of the tongue are effective during their production, were chosen, respectively. The vowels /a/ and /u/ were chosen because the posterior part is where the vocal tract narrows the least and the most, respectively, during the production of these vowels. As a result, seven consonants (/c/, /g/, /h/, /k/, /n/, /r/, /s/) and four vowels (/a/, /i/, /ı/, /u/) related to the vocal tract affected from OSA were selected for this study and given in Table 1. This selection is the same as in our previous study except /r/ [32].

TABLE 1. Selected voices (syllables).

Consonants (with vowels)			
/ca/	/ci/	/cı/	/cu/
/ga/	/gi/	/gı/	/gu/
/ha/	/hi/	/hı/	/hu/
/ka/	/ki/	/kı/	/ku/
/na/	/ni/	/nı/	/nu/
/ra/	/ri/	/rı/	/ru/
/sa/	/si/	/sı/	/su/

B. EXPERIMENTAL STUDIES

The processes of this study were represented as a block diagram in Fig. 1. For the detection of OSA/OSA severity, 28 different consonant voices (syllables) were recorded from the subjects. The interface program (details were given in our previous study [32]) was used for recordings, and registration of subjects with demographic information. The recordings were repeated five times by the subjects for each voice sample. After that, the speech signals were preprocessed, and the features were calculated by using the nonlinear approaches. Experiments were implemented using a laptop computer with the Windows 10 Pro operating system, a 2.8 GHz processor speed, and 16GB RAM (3200 MHz). An external cardioid condenser microphone (Behringer C-1U) with a low noise level was used for voice recordings.

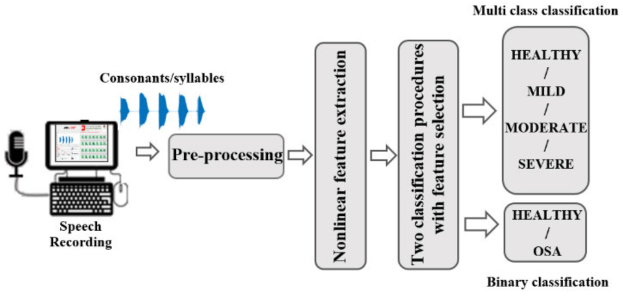

Multi class classification
Consonants/syllables
Speech Recording
Pre-processing
Nonlinear feature extraction
Two classification procedures with feature selection
HEALTHY / MILD / MODERATE / SEVERE
HEALTHY / OSA
Binary classification

FIGURE 1. The processes of the study (by changing from [32]).

Two experimental procedures of classifications were realized in this study. Firstly, the multi-class classification was made, so the classification process was carried out in four class as simple snoring (AHI<5), mild OSA ($5 \leq$ AHI < 15), moderate OSA ($15 \leq$ AHI < 30), and severe OSA (AHI ≥ 30), which are used in the diagnosis of OSA disease (clinical use). In the second experimental procedure of classification, healthy/OSA detection (binary classification) was also performed. In binary classification, subjects diagnosed as simple snoring by PSG test were evaluated as the healthy group, and all other subjects having different OSA severities were taken into the OSA group. SVMs and KNN methods were used as classifiers in both procedures since they are reliable classifiers that are frequently used in the literature and allow us to compare the results of this study with the results of previous studies.

Speech recordings were obtained from 61 people, including 16 mild OSA, 16 moderate OSA, 16 severe OSA, and 13 simple snorers who were considered healthy, at the Sleep Center of Gülhane Training and Research Hospital. The speech signals were sampled at 96000 Hz and recorded at 16- bit resolution. The recording data included 28 speech samples from each subject, so a total of 364 (13 subjects × 28 voices) recording files were collected from simple snorers and 1344 files (48 subjects × 28 voices) from OSA patients (448 files for each severity group). Then, the recordings made repeatedly 5 times were separated from each other and saved as separate audio files. Considering the dynamic structure and variability of the voice, it was repeated 5 times to reduce the effect of this case on the features to be calculated. As a result, 140 files (28 syllables × 5 repeats) were obtained for 28 consonants from each subject. A total of 8540 files (6720 ($28 \times 5 \times 48$) of them from 48 OSA patients (2240 files for each severity group) and 1820 ($28 \times 5 \times 13$) of them from 13 healthy subjects) were obtained. The sampling rate of the signals was down sampled to 44100 Hz (CD quality). DC removal and normalization were applied to the signals. Framing procedure and pre-emphasis filtering were not used. The noise denoising using discrete wavelet transform was applied before the feature extraction. In denoising, Daubechies "db8" wavelet, level of 3 for decomposition, and "heursure thresholding" were utilized [33]. The processes mentioned above are the preprocessing part of our analysis and are the same as in the previous study [32].

C. DATA

Permission was obtained from the ethics committee of Gülhane Training and Research Hospital (document number 46418926/decision number 19/87) for the use of audio recordings and data. Subjects participated in the study voluntarily by signing the informed consent form. The clinical

TABLE 2. Demographic information.

	Healthy Mean ± Std	Mild Mean ± Std	Moderate Mean ± Std	Severe Mean ± Std
Age	43.5 ± 7.7	49.1 ± 9.4	50.5 ± 11.8	47.8 ± 11.1
Weight (kg)	78.4 ± 16.2	87.4 ± 13.8	82.8 ± 9.6	87.4 ± 9.1
Height (cm)	171.5 ± 7.6	173.4 ± 11.6	171.2 ± 11.7	171.1 ± 3.7
BMI (kg/m^2)	26.7 ± 5.7	29.1 ± 3.6	28.3 ± 2.5	30 ± 3.9
AHI	2.6 ± 1.4	10.4 ± 2.8	20.4 ± 4.5	61.5 ± 27.4
Gender	5 female 8 male	4 female 12 male	5 female 11 male	5 female 11 male

diagnoses of subjects were determined with PSG (Philips - Alice 6LDe, Natus Grass Technologies - Comet) by an expert physician. Sleep stages and sleep events were scored manually according to AASM scoring manual version 2.5 [34]. Apnea was scored when there is a drop in the peak signal excursion by $\geq$ 90% of the pre-event baseline using an oronasal thermal sensor or an alternative apnea sensor, for $\geq$ 10 seconds. Hypopnea was scored when the peak signal excursions drop by $\geq$ 30% of the pre-event baseline using nasal pressure or an alternative sensor, for $\geq$ 10 seconds in association with either $\geq$ 3% arterial oxygen desaturation or an arousal. The inclusion criterion for subjects was age 30-65 years, and exclusion criteria were surgery in the vocal tract or tonsillar region, upper respiratory and vocal tract disorders, and symptoms of nasal congestion or sinusitis. In addition, subjects were also asked not to smoke, just prior to the experiment. 13 healthy (simple snoring) subjects (5 females, 8 males) and 48 patients with OSA (14 females, 34 males) participated in the study. The ranges of body mass index (BMI) and age of the mild, moderate, severe OSA, and healthy group were 29.1±3.6 kg/m2, 49.1±9.4 years, 28.3±2.5 kg/m2, 50.5±11.8 years, 30±3.96 kg/m2, 47.8±11.1 years, and 26.7±5.7 kg/m2, 43.5±7.7 years respectively. According to PSG results, 16 of the subjects with OSA had mild, 16 had moderate, and 16 had severe OSA. The demographic information is given in Table 2. The student-t test ($p<0.05$) was applied to the demographic information, and no statistically significant results were obtained among all groups.

D. NONLINEAR FEATURES

1) ATTRACTOR RECONSTRUCTION

The system that produces the voice considers many physiological and anatomical mechanisms which interact with each other; therefore, it is a dynamic system. To examine its dynamics and some chaotic properties of the voice, the attractor of the system is reconstructed in a state space. The time delay embedding method is used to obtain the attractor by using a single signal in the multidimensional state space [35]. According to Takens' theorem, the time delay vectors $\vec{Y}_i$ are constituted as given in (1). In equation (1), N is the number of samples, $i = 1, \ldots, N - (m-1)$, T and m are the time delay (Lag) and embedding dimension (ED), respectively.

$$\begin{aligned}\vec{Y}_1 &= \left(x_1, x_{1+T}, x_{1+2T}, \ldots, x_{1+(m-1)T}\right)\\ \vec{Y}_2 &= \left(x_2, x_{2+T}, x_{2+2T}, \ldots, x_{2+(m-1)T}\right)\\ &\vdots\\ \vec{Y}_i &= \left(x_i, x_{i+T}, x_{i+2T}, \ldots, x_{i+(m-1)T}\right)\end{aligned} \tag{1}$$

The average mutual information (AMI) function is used for defining the T and gives the quantity of information obtained from one random variable about another one. AMI function gives the amount of mutual information between x_i and x_{i+T} (2). The value of AMI function's first minimum is selected as time delay T [36].

For attractor reconstruction, the state space dimension (ED) must be large enough to get the attractor details. The false nearest neighborhood (FNN) method can be used to find ED [37].

$$I(T) = \sum_{i=1}^{N} P(x_i, x_{i+T}) \log_2 \left[\frac{P(x_i, x_{i+T})}{P(x_i)\, P(x_{i+T})}\right], I(T) \geq 0 \tag{2}$$

According to the FNN method, the dimension of the state space where the points on the attractor of the system, which show the states of the system, are no longer false neighbors with each other, gives the ED value.

2) DIMENSIONS (CORRELATION, BOX COUNTING, HIGUCHI)

Fractal is a type of geometric pattern that repeats its own structure over a wide range of scales and does not have a specific size. It is considered to evaluate or compare the geometric complexity of objects or systems of different sizes, shapes, and structures. The fractal dimension can help distinguish between chaos and randomness. Whereas chaos is a type of dynamic behavior of a system. In chaos theory, fractal dimensions are treated as scaling exponents, and an exponent usually remains constant over a set of scale dimensions and a quantitative value. An attractor in state space also has a shape that has a finite and measurable dimension. If the attractor is chaotic, it has a non-integer dimension. If the system has random behavior, the attractor has no definable shape, and its size is not finite. On the other hand, the size of an attractor is related to the minimum number of variables or quantities required to describe or model the system [38].

There are some types of dimensions in the concepts of chaos theory such as information, capacity, correlation, and box-counting. Since correlation dimension (CD) gives information on both geometry and probabilistic aspects of the system attractor, it is a prominent measure. Calculation of CD considers that the trajectories visit some state space regions more often than other regions. Two data points that close together on the attractor in state space are highly correlated spatially. The same two points can be totally unrelated regarding time, depending on the route of trajectory between them. The CD examines the attractor's points for their spatial interrelations [39], [40]. In CD calculation, the number of points within the distance of ε radius from the reference point on the attractor is counted. This procedure is repeated for each point, and then the sum of counts and the sum called as correlation sum is normalized with the total number of pairs of points on the attractor. The correlation sums, $C(\varepsilon)$, are also obtained for larger values of ε. The CD is a scaling exponent according to the relation of $C(\varepsilon) \sim \varepsilon^{CD}$, so it is found as the slope of the straight central region of the log-log plot of correlation sum versus ε (3). After that, the entire process is made for larger EDs, and CD values are plotted versus EDs. For chaotic data, CD values are reached a nearly constant number. This number gives the CD of this attractor of the system [39], [40].

$$\mathrm{CD} = \lim_{\varepsilon \to 0} \frac{\log C(\varepsilon)}{\log \varepsilon} \tag{3}$$

Box-counting dimension (BD) is one of the fractal dimensions; it is widely used and easy to calculate and takes boxes (or cubes) as the measure [38]. Calculation is made on the attractor of the signal reconstructed in the m-dimensional state space. This m-dimensional space is covered by m-dimensional boxes (cubes), each side of which is r. Then, by counting how many m cubes contain the attractor point, N(r) is found. This process continues by changing (reducing) the side length r of m-dimensional boxes (cubes), and the total amount of boxes, N(r), changes as r changes. From the $N(r) \sim r^{-BD}$ relationship, the box-counting size is calculated as the slope of the log (N(r)) divided by log (1/r) (as r goes to 0) curve (4). 1/r is called as a scaled ratio in each spatial direction.

$$\mathrm{BD} = \lim_{r \to 0} \frac{\log N(r)}{\log(1/r)} \tag{4}$$

Higuchi dimension (HD) gives the quantity of the signal complexity in the time domain [41]. HD provides rapid evaluation of signal nonlinearity by using logarithmic calculation. For the signal x, N is the sample number of signal, d is the state space dimension, and the length $L_m(d)$ is calculated for each d, while *m* goes from 1 to the *d*, by (5).

$$L_m(d) = \frac{N-1}{\left\lfloor \frac{N-m}{d} \right\rfloor d^2} \sum_{i=1}^{\left\lfloor \frac{N-m}{d} \right\rfloor} \left| x_N(m+id) - x_N(m+(i-1)d) \right| \tag{5}$$

The length (L(d)) is found by averaging these Lm(d) lengths for each d as in (6).

$$L(d) = \frac{1}{d} \sum\nolimits_{i=1}^{d} L_m(d) \tag{6}$$

Then HD of the signal x is estimated by the slope of the curve plotted on the double logarithmic axis, (ln (L (d)) ,ln (d)).

3) THE LARGEST LYAPUNOV EXPONENT (LLE)

Lyapunov exponent is a quantity that indicates the sensitivity of a system to initial conditions and is calculated by evaluating the divergence or convergence of the trajectories forming the system attractor embedded in the m-dimensional state space [37], [41]. For each dimension in the state space of a system, the Lyapunov exponent can be found, but since at least one of them is positive and is considered an indicator of chaos, it is usually sufficient to find the largest Lyapunov exponent (LLE). In this study, Rosenstein's method, which is fast and easy to implement, and is not affected by the changes in the size of data, noise level and reconstructed parameters, is used [42]. According to Rosenstein's algorithm, the LLE (λ_1) is connected to the relationship $d(t) = Ce^{\lambda_1 t}$ [37]. Here, C is the initial and d(t) is the average divergence value of two adjacent points on different trajectories on the attractor at the t time. The natural logarithm of these divergences is obtained as a function of time, and then the slope of this average line is estimated as LLE (7).

$$y(i) = \frac{1}{t} \langle \ln d(i) \rangle \tag{7}$$

4) ENTROPIES (SHANNON, APPROXIMATE, SAMPLE, LOG ENERGY)

In information theory, the entropy of a random variable is a measure related to the information or uncertainty about the possible outcomes [37], [42]. The Shannon entropy (Shn) is specified the average rate the information obtained from data by produced the complex system. The higher value shows the bigger information gained by a new data value in the process. The Shn is defined for a signal as in (8). In (8), the amount of information is given as $h(P_i) = log_2(1/P_i)$ depending on the statistical distribution of the signal.

Approximate entropy (AEn) and sample entropy (SEn) methods are used to determine the regularity or randomness of the signal without any former information related to the system obtaining the signal. Since these methods have limitless applicability, they are preferred in most research areas [43]. For calculating these entropies, firstly time delay vectors are reconstructed in m dimensional state space and the differences between the two vectors are calculated as scalar values. B_i and A_i are found by counting the similar vectors, having the distance between them is less than the r tolerance value, for consecutive state space dimensions. The difference between these entropies calculations is in the state of the self-counting case. This is taken into account for AEn to avoid calculating the natural logarithm of zero at each iteration. AEn and SEn are given in (9) and (10), respectively.

The entropy of log energy (ELog) is described by Coifman et al. [44] and estimated based on the energy of the signal. ELog is calculated as in (11) since square of signal gives the whole information obtained from the signal. In (11) $x = (x_1 x_2 x_3, \ldots, x_N)$ is the signal which has N samples. ELog which is one of the wavelet entropy types in wavelet packet was also calculated without wavelets and decomposition.

$$\mathrm{Shn} = \sum_{i=1}^{N} P_i \log_2 \left(\frac{1}{P_i}\right) \tag{8}$$

$$\mathrm{AEn}\,(m, r, N) = -\frac{1}{N-m} \sum_{i=1}^{N-m} \ln \left(\frac{A_i}{B_i}\right) \tag{9}$$

$$\mathrm{SEn}\,(m, r, N) = -\log \left(\sum_{i=1}^{N-m} A_i \Big/ \sum_{i=1}^{N-m} B_i \right) \tag{10}$$

$$\mathrm{ELog}(x) = \sum_{i}^{El} \log(x_i^2) \tag{11}$$

5) SQUARE ROOT DIFFERENCE OF TIME DELAY-BASED SIGNAL COPIES (SRDIFF)

The time delay embedding method is used for reconstructing the system attractor and presents the system states [35]. According to the Takens' method, the attractor is obtained by using the time delayed (lagged) copies of the signal with time delay T (Lag). If T is chosen as the first minimum of the AMI function, the difference between x_i and x_{i+T} (first differencing with T) can give some new information about the system [33], [34]. As a result of this approach, the root of the sum of the squares of the differences (SRDiff) between the original signal x and its T delayed copies is used as the feature [32]. SRDiff is calculated as in (12), where N is the total number of samples in the signal.

$$\mathrm{SRDiff} = \sqrt{\sum_{i=1}^{N-T} (x_i - x_{i+T})^2} \tag{12}$$

6) HIGH ORDER STATISTICS (SKEWNESS AND KURTOSIS)

In recent years, high order statistics (HOS) have been a popular field of statistical signal processing used in many areas especially in digital signal processing. HOS quantities are extensions of second-order measures to higher orders [45]. The second, third, and fourth-order measures are related to the variance, skewness (Skew), and kurtosis (Kurt), respectively. The second-order measures, such as the autocorrelation function and power spectrum, are proper for the signals with Gaussian distribution. On the other hand, real-life data are non-Gaussian and non-stationary, so using higher order moments for the analysis and specification of nonlinearities in the signal can give useful findings [45]. Speech signals or voices are produced by systems having nonlinear dynamics [32]. The Fourier spectra of these signals take place in a very wide range. The higher order properties of the distributions can give extra information about the signals. The third order (Skew) and fourth order (Kurt) are measures giving the asymmetry and (or symmetry) peaked (or flat) in a probability distribution relative to a normal distribution, respectively. The formulas of Skew (13) and Kurt (14) measures are given below. In these equations, $\bar{x}$ is the mean of signal and N is the number of data points.

$$\mathrm{Skew} = \frac{\frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2}\right)^3} \tag{13}$$

$$\mathrm{Kurt} = \frac{\frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^4}{\left(\frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2\right)^2} \tag{14}$$

E. FEATURE SELECTION

The process of feature selection using various methods is an important and desirable step to reduce the number of input variables in classification applications for improving the model performance and decreasing the computational cost of modeling. Statistics-based methods statistically evaluate the relationship between each input variable and the target variable and score the inputs according to the strength of the relationship, where larger scores indicate greater feature importance. In multiclass classification, the one-vs-rest method can be preferred such as analysis of variance (ANOVA) to feature ranking [46]. The one-way ANOVA is implemented for each predictor variable, grouped by class, and the features are ranked using the p values. ANOVA cumulatively compares the arithmetic means of the groups for each feature. When at least one of these comparisons is statistically significant, the ANOVA result is also significant. Hypotheses are: H_0, there is no difference between the means, and H_1, there is a significant difference between at least two averages. The p value is obtained from the F measure of ANOVA. The ranked scores are calculated as –log(p). In this study, the features that have ranking score is greater than 5 were selected from the ranked feature list yielded by ANOVA for classification.

F. CLASSIFICATION STUDIES

In this study, the robust classifiers of KNN and SVMs, with 5-fold cross validation, which we used in our previous study [32], were preferred for classifications and MATLAB R2022b Classification Learner Application was used. Classification applications were carried out to show that the proposed analysis approach can determine OSA and OSA severity levels. Therefore, KNN and SVMs classifiers were attempted by changing only their major variants. In applications, the other classifier parameters that are assigned as default values in MATLAB R2022b Classification Learner Application were not changed. We did not prefer too many adjustments for classifiers because the aim of this study is not to reach highest accuracy rates, but simply to presented a result that could be useful and generalized. Thus, the hyperparameter optimization or tuning algorithms were not used for classifications. For the KNN classifier, attempts were made for Euclidean distance, 1 and 10 nearest neighbors, and cosine and cubic distance 10 nearest neighbors.

Linear, quadratic, cubic, and Gaussian kernel functions were used for SVMs classifier. Two multi-class classification approaches which are one-vs-one and one-vs-all were used for determining the severity level of OSA with SVMs.

Cross-validation is used in machine learning to predict the ability of the classifier model on test data to be encountered later (which we do not have yet). Cross-validation (k-fold) is a resampling procedure used to evaluate the performance of classifier models on a limited data sample. This approach randomly divides the observation set into k groups/folds with approximately equal sizes. Each of these groups is used as the test and the remaining k-1 group as training, respectively, and k classification processes are performed. Thus, the model is tried k times with distinct training and test groups. Model performance is obtained as the average of the results of these k classifications. Since the train and test groups are randomly determined by the method, it provides a less biased, less optimistic estimation of model performance than other methods such as manual training/test separation [47]. In our study, k was chosen as 5 considering the number of subjects in each group. There is one measurement (a single value for each feature) obtained from each subject, and the measurements of a subject in each classification trial were included in either the training or the test set. Thus, it was ensured that our results were independent of the subjects.

In the dataset, the healthy group is labeled with the number 0, mild OSA group 1, moderate OSA group 2, and severe OSA group 3. The classification performance during the determination of healthy/OSA and OSA severity was evaluated with the precision, recall, and F1-score parameters displayed on the confusion matrix, as well as the accuracy parameter.

III. RESULTS

In our study, 28 voice samples were recorded by repeating each of them 5 times. For each of these voice samples, 12 features were evaluated (Lag, CD, BD, HD, LLE, Shn, AEn, SEn, ELog, SRDiff, Skew, Kurt). The same feature was calculated 5 times for each voice sample (28 voices × 5 repeats × 12 features = 1680 features for each subject). Considering the variability of the dynamics in the formation of the voice, the changes in the vocal performance of the subject during the recording and possible environmental factors, the average value of the features, which were calculated 5 times for each voice, was taken to reduce the effect of these undesirable situations affecting the voice recording on the calculated features. This process is presented in Fig. 2. Therefore, since only one value was taken from each feature for each voice, 28 × 1×12=336 features were obtained for each subject. In this case, the size of our feature dataset was 336 × 61=20496. The ANOVA feature selection method included in the MATLAB 2022b Classification Learner application was applied to these 336 features, and the features selected according to their importance for both classification procedures are given in Table 3. The results in Table 3 showed that selected features belong to /g/, /h/, /n/ and /r/ consonant groups in both classification procedures. In addition, none of the entropy values are included in the selected features.

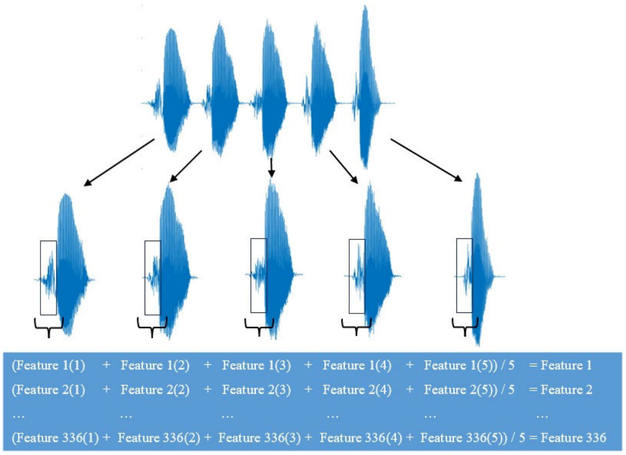

(Feature 1(1) + Feature 1(2) + Feature 1(3) + Feature 1(4) + Feature 1(5)) / 5 = Feature 1
(Feature 2(1) + Feature 2(2) + Feature 2(3) + Feature 2(4) + Feature 2(5)) / 5 = Feature 2
… … … … … …
(Feature 336(1) + Feature 336(2) + Feature 336(3) + Feature 336(4) + Feature 336(5)) / 5 = Feature 336

FIGURE 2. Reducing the number of features by averaging.

TABLE 3. Selected features for multi-class and binary classifications with Anova.

Features for multi-class classification (ranked value >5)				Features for binary class classification (ranked value >5)	
1	ELog /ca/	8	HD /na/	1	Lag /ri/
2	CD /ra/	9	CD /ni/	2	Skew /ga/
3	Lag /ri/	10	CD /nu/	3	LLE /nı/
4	Skew /ga/	11	Kurt /ha/	4	Kurt /ha/
5	SRDiff /gı/	12	ELog /ha/	5	Lag /hu/
6	BD /gı/	13	SRDiff /ha/		
7	CD /na/	14	HD /ha/		

The mean ± standard deviation (SD) values of selected features in Table 3 for both classifications are presented in Table 4 and the graphical representation of a few of them, which are remarkable in terms of distinguishing the groups from each other, is also given in Fig. 3 by using box plots. In this study, it was tried to reveal the ability of the features under nonlinear analysis approaches to determine the severity levels of OSA, taking into account the promising results of our previous study [32]. Therefore, it is important to evaluate the calculated features and their relationship with the severity of OSA. These considerations are useful for developing some physiological interpretations of how OSA severity changes the nonlinear dynamics of voice. For this purpose, it has been tried to visually reveal how the variability or distribution of

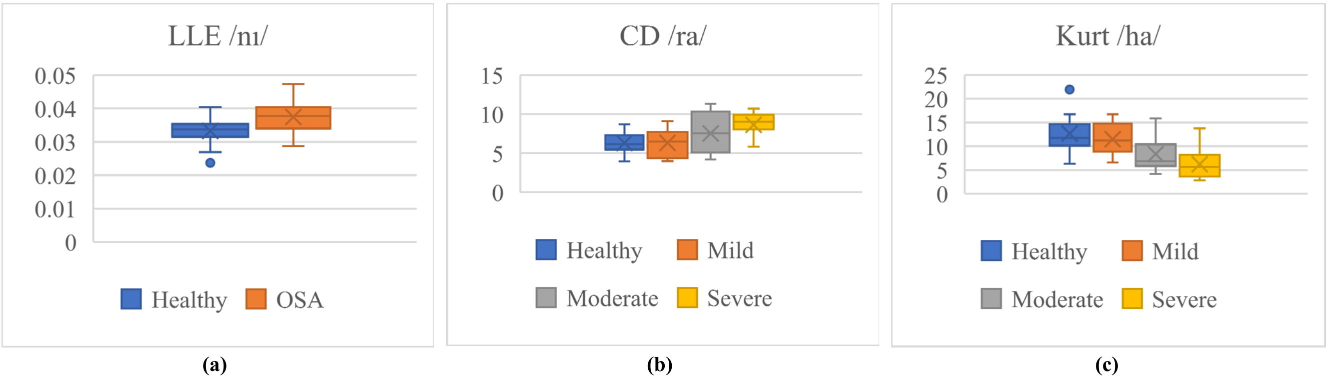


FIGURE 3. The graphical representation of mean±SD values some selected features given in Table 3. a) LLE /nɪ/ (binary class) b) CD /ra/ (multi-class) c) Kurt /ha/ (multi-class).

TABLE 4. The mean ± SD values of selected features in both classification procedures for groups.

Multi-Class Classification				
Selected features	**Healthy**	**Mild OSA**	**Moderate OSA**	**Severe OSA**
ELog /ca/	-12853.4 ±4569.26	-12682.8 ±3301.89	-13291.9 ±2932.42	-19267.9 ±6026.82
CD /ra/	6.31±1.33	6.34±1.73	7.54±2.567	8.64±1.58
Lag /ri/	9.54±0.59	9.91±0.35	9.94±0.14	9.95±0.15
Skew /ga/	-0.67±0.47	0.07±0.41	-0.19±0.41	0.053±0.4
SRDiff /gɪ/	6.32±1.47	5.95±1.36	4.101±1.73	6.24±1.43
BD /gɪ/	3.95±0.01	3.92±0.11	3.76±0.11	3.92±0.01
CD /na/	3.27±0.48	3.43±0.80	3.91±0.4	5.578±2.62
HD /na/	1.08±0.03	1.07±0.02	1.065±0.02	1.178±0.17
CD /ni/	3.83±0.48	3.38±0.66	3.929±0.41	4.147±0.50
CD /nu/	3.38±0.53	3.66±0.71	3.54±0.59	4.2±0.55
Kurt /ha/	12.65±3.93	11.49±3.29	8.36±3.64	6.21±3.01
ELog /ha/	-27127 ±4392.01	-24456.5 ±4380.3	-29631.7 ±8864.6	-34453.1 ±20738.5
SRDiff /ha/	7.20±1.91	6.49±1.49	9.31±3.64	11.49±2.80
HD /ha/	1.62±0.05	1.57±0.07	1.64±0.08	1.49±0.12

Binary Classification		
Selected features	**Healthy**	**OSA**
Lag /ri/	9.54±0.59	9.93±0.23
Skew /ga/	-0.67±0.47	-0.02±0.42
LLE /nɪ/	0.030±0.01	0.037±0.01
Kurt /ha/	12.65±3.93	8.69±3.92
Lag /hu/	8.92±1.09	9.62±0.66

some features selected to be used in classification changes for OSA groups with box plot representations (Fig. 3). In Fig. 3, the nonlinear states of the groups can be observed by evaluating the values of some of the features given in Table 4. Positive LLE value, which is accepted as an indicator of nonlinear behavior (chaos), was found to be lower in the healthy group than in the OSA group (Fig. 3a). When the CD values are examined for all classes in Table 4, it is understood that CD increases from the healthy group to the severe OSA group (Fig. 3b). In kurtosis (Fig. 3c), an increasing trend from healthy to severe OSA was obtained, which means that the distribution of signal components is getting flatter. These results are interpreted in the Discussion section.

In Fig. 4, to visually examine the effectiveness of the selected features for both classification procedures, the 61 subjects were projected on a two-dimensional plane by using the t-SNE algorithm which is preferred for visualization and dimension reduction [14], [48]. The 4 classes (healthy,

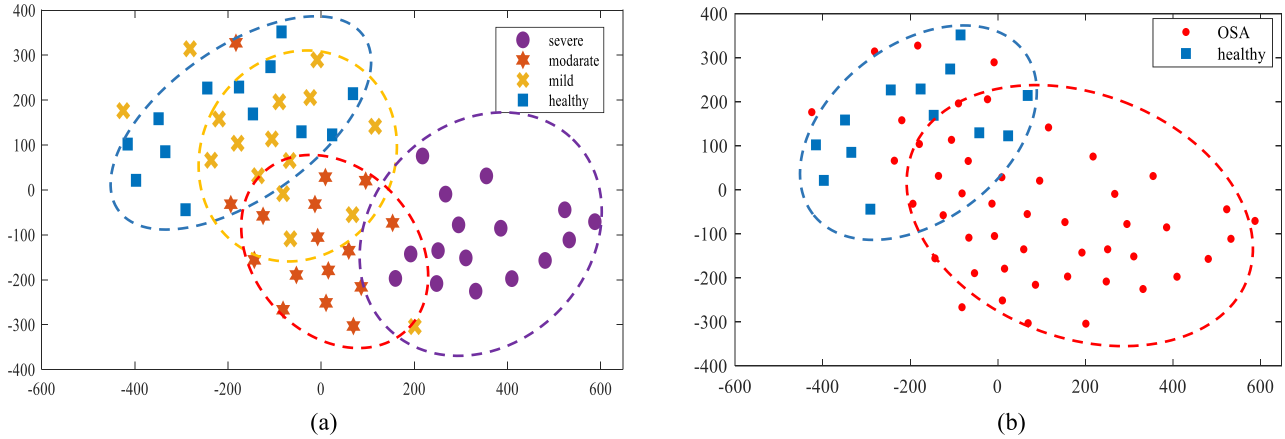


FIGURE 4. Visualization of the distributions of the subject groups obtained by applying the data reduction method (t-SNE) to the selected features. a) Four groups for multi-class classification b) Two groups for binary classification.

mild OSA, moderate OSA, severe OSA) (Fig. 4a) and 2 class (healthy and OSA) (Fig. 4b) were represented as group regions with dashed circles. Fig. 4 shows that the groups are separated by some small overlaps. This discrimination case affects the performances of classification.

In both procedures of multi-class and binary classifications, it has been attempted to see which one will provide the highest classification performance when the default parameters of KNN and SVMs classifiers were used (Table 5 and Table 6). The results of multi-class classification studies using 14 selected features in Table 5 are summarized according to the accuracy parameter. The performance values of the classifiers obtained for each group (healthy:0, mild OSA:1, moderate OSA:2, severe OSA:3) are presented in Table 5.

According to Table 5, the best result obtained with the KNN was found as 78.7% accuracy with cosine distance and 10 nearest neighborhood parameters. Among the classifications made with SVMs, the best accuracy result was obtained at 82% with the one-vs-all classification scheme and the Gaussian kernel.

While the accuracy gives an overall result for the performance of multiclass classifiers, the confusion matrix is used to evaluate the performance for each class separately. Table 5 presents the confusion matrices of the best accuracies obtained with KNN (Fig. 5) and SVMs (Fig. 6).

According to the best classification result obtained with the KNN in Fig. 5. most of the subjects were placed in the correct class within 4 classes. Most of the misclassifications were placed in the clinically closest class. Only 2 patients in the mild OSA (1) group were classified as severe and 1 patient in the moderate OSA (2) group was classified as healthy. Hence, only 3 patients were placed in a patient group that was not clinically close to their own. Considering the TPR values, the detection accuracy of over 80% was achieved in the healthy, moderate, and severe groups, while the accuracy of 56.2% was maintained in the mild group. Fig. 6 shows the confusion matrix of the best result for SVMs. It can be seen that the results are similar to the KNN results (Fig. 5), except for one subject who is not close to his/her clinical group. It is also observed that the SVMs classifier gives better results for the mild OSA group with 62.5% TPR.

The accuracy results of binary classification using 5 selected features (Table 3) are given in Table 6. Table 6 shows that the best result obtained with the KNN classifier was found as 88.5% accuracy with cosine distance and 10 NN parameters. The best result with SVMs was found to be 95.1% accuracy with the quadratic kernel. Confusion matrices are also presented for the best accuracies (Table 6) in Fig. 7 and Fig. 8 for KNN and SVMs, respectively. From Fig. 7 and Fig. 8, it can be seen that a total of 7 and 3 subjects are incorrectly detected in KNN and SVMs, respectively.

IV. DISCUSSION

The principal aim of this study is to detect the OSA severity level as mild, moderate, and severe as in the clinic use using nonlinear analyses from speech for the first time in the literature. In addition, healthy/OSA classifying was also carried out in order to compare the results with respect to previous studies. In this study, the speeches produced by the vocal tract organs deformed due to apnea were evaluated with nonlinear approaches to determine the severity of OSA and healthy status.

Speech recordings (28 voices/syllables, 5 repetitions) were made from subjects diagnosed by the PSG test. (13 simple snorer /healthy, AHI<5), 16 mild OSA ($5 \leq$ AHI < 15), 16 moderate OSA ($15 \leq$ AHI < 30) and 16 severe OSA (AHI ≥ 30)). Lag, CD, BD, HD, LLE, Shn, AEn, SEn, ELog, SRDiff, Skew, and Kurt features reflecting nonlinear dynamics, which give nonlinear dynamics, were used. Some of these features have not been used in OSA detection in the literature before, while others have been used in our previous study [32]. The features that were calculated 5 times for each voice sample recorded by repeating 5 times were reduced to a single feature value by taking their averages. Thus, a single value was obtained for each feature for each voice sample. 28 voices × 12 features, totally 336 features, were obtained for each subject. In the study, two classification procedures were realized by using the SVMs

TABLE 5. The multi-class classification results (%) with 5-fold cross validation.

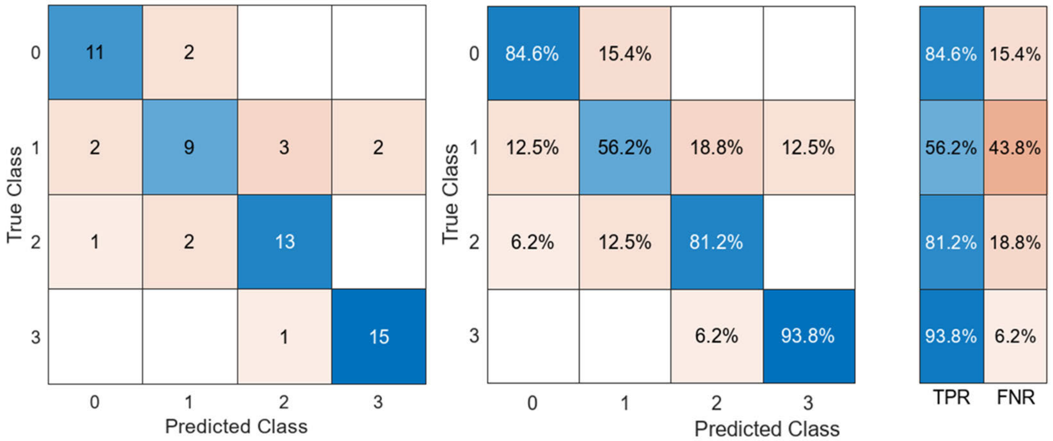
	Classifier	Class	0	1	2	3	Overall	Acc [a]
KNN	Linear-1 NN	Precision [a]	69.2	54.5	75	92.9	72.9	70.5
		Recall [a]	69.2	75	56.3	81.3	70.5	
		F1 [a]	69.2	63.2	64.3	86.7	70.9	
	Linear-10 NN	Precision	73.3	50	71.4	100	73.7	70.5
		Recall	84.6	62.5	62.5	75	71.2	
		F1	78.6	55.6	66.7	85.7	71.7	
	Cosine-10 NN	Precision	78.6	69.2	76.5	88.2	78.1	78.7
		Recall	84.6	56.3	81.3	93.8	79	
		F1	81.5	62.1	78.8	90.9	78.3	
	Cubic-10 NN	Precision	90.9	48.0	71.4	100	77.6	70.5
		Recall	76.9	75	62.5	68.8	70.8	
		F1	83.3	58.5	66.7	81.5	72.5	
SVMs kernel / one-vs-one class	Linear	Precision	78.6	81.8	70.0	87.5	79.5	78.7
		Recall	84.6	56.3	87.5	87.5	79.0	
		F1	81.5	66.7	77.8	87.5	78.4	
	Quadratic	Precision	80.0	64.3	66.7	100	77.8	77
		Recall	92.3	56.3	75.0	87.5	77.8	
		F1	85.7	60.0	70.6	93.3	77.4	
	Cubic	Precision	73.3	60.0	75.0	100	77.1	75.4
		Recall	84.6	75.0	56.3	87.5	75.9	
		F1	78.6	66.7	64.3	93.3	75.7	
	Gaussian	Precision	75.0	62.5	76.5	93.8	77.0	77
		Recall	69.2	62.5	81.3	93.8	76.7	
		F1	72.0	62.5	78.8	93.8	76.8	
SVMs kernel / one-vs-all class	Linear	Precision	73.3	83.3	65.0	100	80.4	78.7
		Recall	84.6	62.5	81.3	87.5	79.0	
		F1	78.6	71.4	72.2	93.3	78.9	
	Quadratic	Precision	75.0	75.0	73.3	100	80.8	80.3
		Recall	92.3	75.0	68.8	87.5	80.9	
		F1	82.8	75.0	71.0	93.3	80.5	
	Cubic	Precision	76.9	55.0	69.2	93.3	73.6	72.1
		Recall	76.9	68.8	56.3	87.5	72.4	
		F1	76.9	61.1	62.1	90.3	72.6	
	Gaussian	Precision	85.7	83.3	76.5	83.3	82.2	**82**
		Recall	92.3	62.5	81.3	93.8	82.5	
		F1	88.9	71.4	78.8	88.2	81.8	

[a] Precision (PPV: Positive Predictive Value) (%), Recall (Sensitivity, TPR: True Positive Rate) (%), F1: F1- score (%), Acc: Accuracy (%), 0: healthy, 1: mild OSA, 2: moderate OSA, 3: severe OSA.

TABLE 6. The binary classification results (%) with 5-fold cross validation.

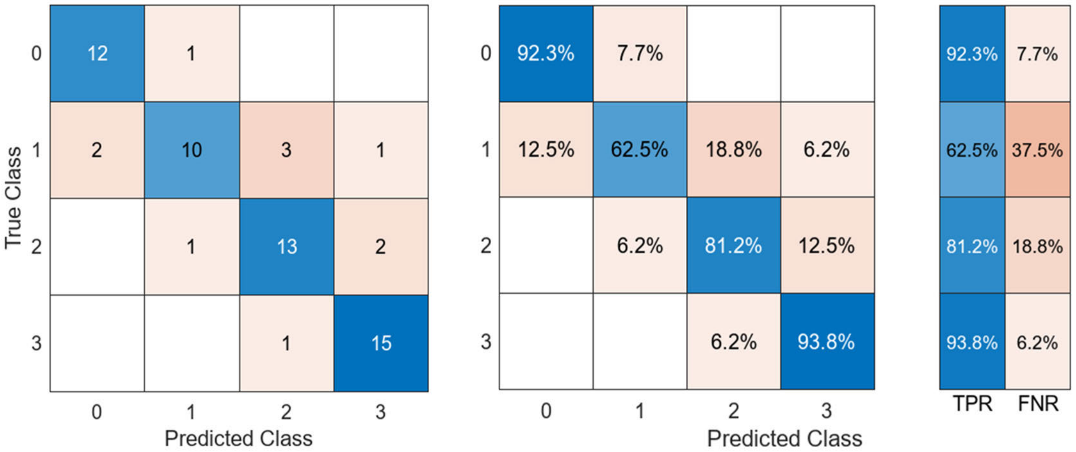
	Classifier	Precision [a]	Recall [a]	F-1 score [a]	Acc [a]
KNN	Linear-1 NN	70	53.8	60.9	85.2
	Linear-10 NN	100	23.1	37.5	83.6
	Cosine-10 NN	80	61.5	69.6	88.5
	Cubic-10 NN	100	15.4	26.7	82
SVMs kernel	Linear	90	69.2	78.3	91.8
	Quadratic	**91.7**	**84.6**	**88**	**95.1**
	Cubic	80	92.3	85.7	93.4
	Gaussian	83.3	38.5	52.6	85.2

[a] Precision (PPV: Positive Predictive Value) (%), Recall (Sensitivity, TPR: True Positive Rate) (%), F1: F1- score (%), Acc: Accuracy (%).



(a) (b)

FIGURE 5. The confusion matrices of for multi-class with KNN (0: simple snoring 1: mild 2: moderate 3: severe) a) Number of observations. b) Decision percentages, TPR and false negative rate (FNR) values.



(a) (b)

FIGURE 6. The confusion matrices of for multi-class with SVMs (0: simple snoring 1: mild 2: moderate 3: severe) a) Number of observations. b) Decision percentages, TPR and FNR value.

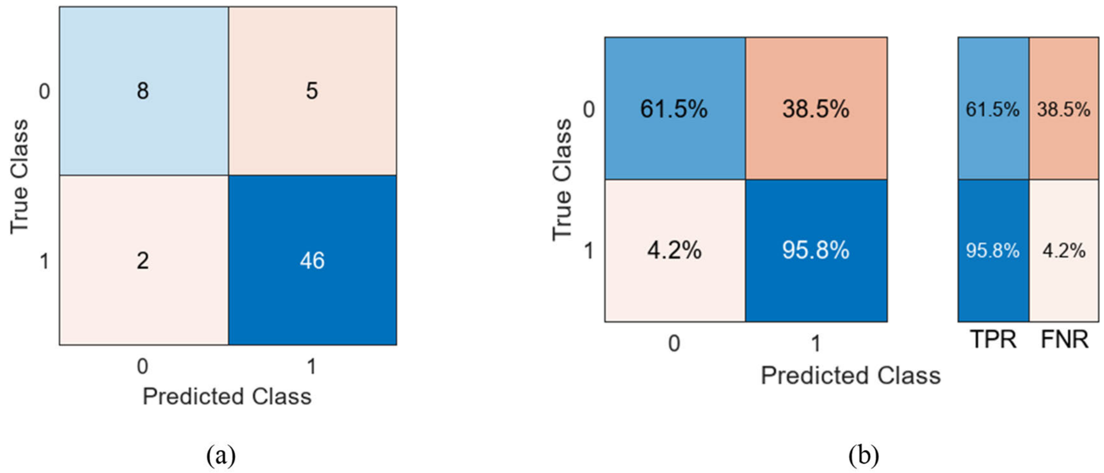


FIGURE 7. The confusion matrices of for binary class with KNN (0: healthy 1: OSA a) Number of observations. b) Decision percentages, TPR and FNR values.

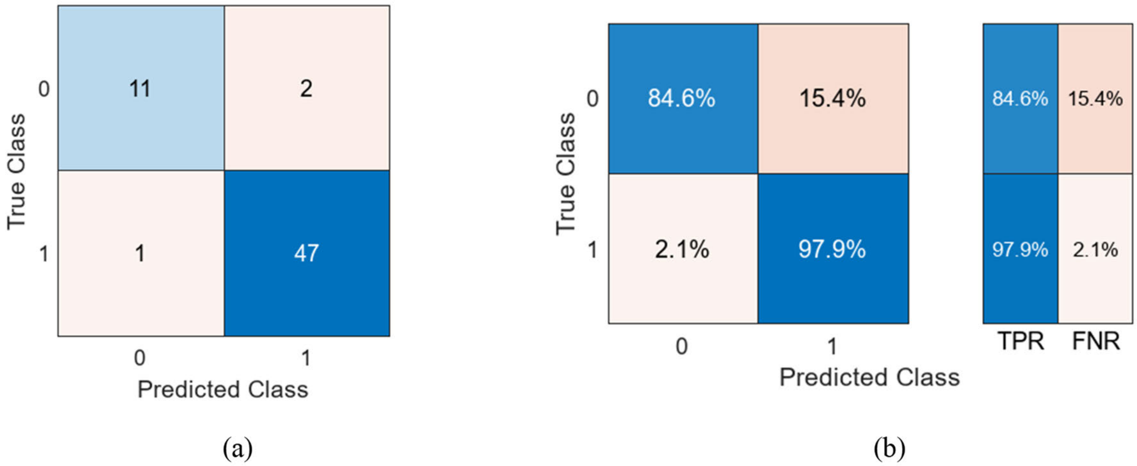


FIGURE 8. The confusion matrices of for binary class with SVMs (0: healthy 1: OSA a) Number of observations. b) Decision percentages, TPR and FNR values.

and KNN classifiers: multi-class and binary classifications. In the multi-class classification, the classes were simple snoring/mild OSA/moderate OSA/severe OSA, while in the binary classification the classes were healthy (simple snoring, AHI<5)/OSA (all severity groups of OSA (AHI $\geq$ 5)). Classifications were performed by using 5-fold cross validation process, and 14 and 5 features selected from 336 features with ANOVA were used in multi-class, and binary classifications, respectively. In the classifications carried out using SVMs and KNN, the classifier features with the highest detection accuracies, and results are given in Table 7. As can be seen from Table 7 the best results were obtained with SVMs in both classification procedures; the classes were separated from each other with 82% accuracy in multi-class classification, and OSA was detected with 95.1% accuracy in binary classification.

The results of our study can be compared, firstly, with studies suggesting the detection of OSA at a lower cost (effort, time, complexity) than PSG. Using only respiratory parameters (airway flow, respiratory effort, pulse oximetry) [2], [3], [4], single-channel ECG [5], [6], [7], [8], [9], breathing/respiratory/snore sounds [13], [14], [15], [16], [17], [18], a success rate of around 90% has been achieved in the detection of OSA or OSA severity. However, considering the all-night recording, the difficulties experienced by the subjects in recording on their own, and the difficulty of analysis

TABLE 7. The properties of classifiers for best results in both classifying procedures.

Class Type	Classifier Characteristics	Acc[a]	Sens	Spec
Multi-Class	SVMs Gaussian kernel one-vs-all class	82.0	82.5	94.0
Binary	SVMs Quadratic kernel	95.1	97.9	84.6

[a] Acc: Accuracy (%), Sens: Sensitivity (%), Spec: Specificity(%).

in the relevant studies, the goal of easier and more comfortable detection of OSA could not be fully achieved. In studies using speech/voice recording, getting results with a few minutes of recordings and analysis provides a great advantage over the systems developed to overcome the above-mentioned difficulties. When Table 8 is examined, it is seen that the successes of detecting OSA in speech/voice related studies are in the range of 80%-90%. However, in none of these studies, OSA severity was determined by multi-class classification, so OSA severity was tried to be determined by binary classifications for multiple AHI index threshold values (for example, AHI=10 and AHI=30 [30], [31]). Therefore, there is no study in the literature that classifies OSA as the same in clinical diagnosis (simple snorer/mild/moderate/severe) by using speech/voice. The results of our study have shown that our proposed method has the potential to be a simple and usable method that reduces the cost and employee burden of OSA severity diagnosis compared to previous studies.

To compare the results of our binary classifier with the literature, the data sets, methods, and results of other studies are presented in Table 8. The binary classification result, which we reached with 95.1% accuracy, is higher than the study of Blanco et al., which previously showed the highest performance (88.5% accuracy) in the literature. In Blanco et al.'s study, subjects with AHI$<$10 and AHI$>$30 were classified. Hence, the subjects having 10 $<$ AHI $<$ 30 values, this range roughly covers mild to moderate OSA severities, were excluded. In our study, a classification was made between healthy and all OSA groups in accordance with the clinical use.

In this study, features based on the hypothesis that nonlinear dynamics during voice formation have distinctive features in people with apnea were selected, similar to our previous study [32]. In Table 8, it is seen that other studies mostly work with classical analysis measures used in voice analysis. In this study, new ones were added to the nonlinear features (CD, BD, ELog, Skew, Kurt) that we used in our previous study [32]. Thus, the performance of new features that had not been examined before was evaluated and the diversity of features was increased. One of the limitations of our previous study was that the healthy group consisted of subjects not included in the PSG test. In this study, all our subjects were tested and diagnosed by PSG.

The degree of nonlinearity can be evaluated for the healthy and OSA severity groups by examining the selected features (Table 4, Fig. 3). Positive LLE value, which is accepted as an indicator of nonlinear behavior (chaos), was found to be lower in the healthy group than in the OSA group. This shows that the voice signals have chaotic behavior in both groups, and the level of nonlinearity in OSA is increased compared to the healthy ones. It is understood that CD increases from the healthy group to the severe OSA group, generally. CD is a feature associated with the degrees of freedom of the system, so it can be said that the number of variables managing the system increases from the healthy to the severe OSA. This can be interpreted in the sense that the complexity of the system, that is, its nonlinearity, has increased. In kurtosis, one of the HOSs, an increasing trend from healthy to severe OSA was observed, which means that the distribution of signal components is getting flatter in a general manner. A flatter spectrum indicates that the signal components are distributed over a wider range. A characteristic of chaotic signals is that their spectra show a wide distribution. Accordingly, it was evaluated that nonlinearity increased from healthy to severe OSA group. From these results, it is seen that the values and distributions of the features for the OSA groups have different nonlinearity levels from each other. These changes related to OSA severity in nonlinear features show that the voices of people with OSA have a more chaotic behavior than healthy ones, and the degree of chaoticity increases as the severity of OSA increases. This situation was interpreted as the deformed muscles and structures in OSA making the dynamic behavior of the voice even more chaotic.

For this study, a total of 28 different voice recordings were made, and 336 features were obtained from each subject. Selection by the ANOVA, 14 features were used to determine OSA severity, and 5 were used to detect OSA. When the selected features were examined, it was seen that these features belong to the voice samples of /ca/, /ra/, /ri/, /ga/, /gi/, /na/, /ni/, /nu/, and /ha/. Therefore, using these 9 voice samples from 28 voices was adequate for specifying the OSA severity. For the detection of OSA, it seems sufficient to record and analyze the 5 different voices of /ri/, /ga/, /nı/, /ha/, /hu/. Thus, in this case, there are 2 voice samples (/hu/, /nı/) in addition to the previous 9 voice samples. The time cost may be roughly calculated according to the results. If both procedures are desired to be applied, considering that all voice samples recording time is around 20 minutes, a recording time of 10 minutes is required for 11 different voices in total. If it takes 5 minutes to calculate a total of 16 features from 11 voices, it can be said that it is possible to determine OSA and OSA severity around 15 minutes duration, in addition to the clinical assesment. If the results of this study can show the same success in much larger groups, the proposed method can be the preferred as a practical solution instead of PSG.

Our results for OSA/healthy detection can be compared with the results of apnea screening tests. Godoy et al. [49], compared the Berlin (BQ) and STOP-Bang (S-Bang) questionnaires, which are the most commonly used in screening, and Epworth Sleepiness Scale (ESS) tests for patients over 65 years of age, and found that BQ was the most successful test (0.66 accuracy, 0.64 sensitivity, 0.71 specificity). E Silva et al. [50], compared these screening tests mentioned above for various age groups, and they found that S-Bang gave the best result (0.81 accuracy, 0.74 sensitivity, 0.87 specificity). Although these tests are applied with minimal cost, they contain subjectivities. In this study, a much higher performance was obtained by using an objective method (0.95 accuracy, 0.98 sensitivity, 0.85 specificity for

TABLE 8. Previous OSA detection studies on speech (by changing from [32]).

	Data set [a]	Features	Classifier Conditions	Results [a]
Goldshtein *et al.* (2011) [23]	26 healthy (12 m/14 f), 67 OSA (48 m/19 f) /a/, /e/, /i/, /o/, /u/ and/n/, /m/ taken from speech (Hebrew-speaking)	Acoustic features	LOO, Re-substitution, GMM	With short-term feature sets: For male: 83% (Spec), 79% (Sens) For female: 86% (Spec), 84% (Sens)
Blanco *et al.* (2013) [29]	AHI<10 control, AHI>30 OSA	12 MFCCs and velocity, energy, acceleration coefficients	LOO, GMM, MRMR	The highest results: 88.5% (Acc), 90.6% (Spec), 86.7% (Sens)
Espinoza-Cuadros *et al.* (2015) [27]	285 m with OSA 4 sentences and Spanish vowels [i, e, a, o, u] 570 photographs (frontal and profile) from 285 m subjects	MFCC+ΔMFCC, demographic information, neck circumference	LOO, 5-fold cross-validation, one subject is test, others are train with SVR	72.2% (Acc), 64.8% (Spec), 73.3% (Sens)
Espinoza-Cuadros *et al.* (2016) [28]	426 m, AHI <10: 125 non-OSA, AHI≥10: 301 OSA Spanish vowels [i, e, a, o, u], 4 sentences	Voice features (MFCC+ΔMFCC), clinical values: age, height, weight, BMI and neck circumference	10-fold cross-validation, one group is test, others are train with SVR	71.06 (Acc), 20.6 (Spec), 92.92 (Sens)
Simply *et al.* (2020) [26]	398 (208 OSA - 190 Non-OSA), AHI>15, 53 m/145 f (93 of them is diagnosed with Peripheral Arterial Tonometer-PAT systems, others with PSG) Hebrew protocol, /a/, /e/, /u/, /o/, /i/ vowels	Zero crossing rates (ZCRs), energies, MFCCs, kurtosis, pitch peak, age, BMI	Train: 149 OSA/136 non-OSA Validation: 40 OSA/31 non-OSA Test:19 OSA/23 non-OSA	The results by combining of 4 system: 77.14% (Acc), 79% (Spec), 75% (Sens)

Ding *et al.* (2020) [30]	151 m speakers with OSA
	AHI=10: 10 (AHI > 10), 41 (AHI ≤ 10)
	AHI=30: 75(AHI > 30), 76 (AHI ≤ 30)
	Chinese vowels /a/, /o/, /e/, /i/, /u/, /ü/, /en/, /eng/
		MFCC, LPCCs, formants (F1~F4), log energy, etc.	SVMs, LOO	78.8 (Acc) (for both task)
				79.1% (Sens) AHI=10, 77.3% (Sens) AHI=30
				78.0% (Spec) AHI=10, 80.3% (Spec) AHI=30

Ding *et al.* (2022) [31]	151 m speakers with OSA
	AHI>10 (n=117) AHI≤10 (n=41)
	AHI>30 (n=80) AHI≤30 (n=78)
		30 LPCCs	Decision tree, LOO	AHI=10: 81.7% (Acc), 81.8% (Sens), 81.3%(Spec)
				AHI=30: 80.3% (Acc), 78.1% (Sens), 82.6% (Spec)

Yilmaz *et al.* (2023) [32]	20 OSA (AHI>9), 20 healthy, 16 OSA/16 healthy for validation, 4 OSA/4 healthy for testing /a/, /ı/, /i/, /u/ vowels and consonants (syllable): /ca/, /cı/, /ci/, /cu/, /ga/, /gı/, /gi/, /gu/, /ha/, /hı/, /hi/, /hu/, /ka/, /kı/, /ki/, /ku/, /na/, /nı/, /ni/, /nu/,/sa/, /sı/, /si/, /su/	Lag, ED, HD, Shn, AEn, Sen, LLE, SRDiff	KNN, SVMs, MRMR 5-fold cross validation	The best result for testing: 82.5% accuracy for consonants with 6 features and SVMs
Proposed Study	61 patients: 16 simple snoring/healthy (5f - 8m) 16 mild OSA (4f - 12m), 16 moderate OSA (5f - 11m) 16 severe OSA (5f - 11m) Consonants (syllable): /ca/, /cı/, /ci/, /cu/, /ga/, /gı/, /gi/, /gu/, /ha/, /hı/, /hi/, /hu/, /ka/, /kı/, /ki/, /ku/, /na/, /nı/, /ni/, /nu/, /ra/, /rı/, /ri/, /ru/, /sa/, /sı/, /si/, /su/	Lag, CD, HD, BD, LLE, SRDiff, skewness, kurtosis	KNN, SVMs, ANOVA 5-fold cross validation	The highest accuracies: Multi-class classification: 82% overall with SVMs (Gaussian kernel one-vs-all class) Binary classification: 95.1% accuracy with SVMs (Quadratic kernel)

[a] Acc:accuracy, Spec:specificity, Sens:sensitivity, m: male, f: female

OSA/healthy detection). In addition, these screening tests do not provide any result for determining the severity of OSA. In our study, OSA severities could be determined with good performance (0.82 accuracy, 0.83 sensitivity, 0.94 specificity).

V. LIMITATIONS

Our most important limitation in this study is the small number of subjects in 4 classes. It was aimed to select the subjects in such a way that there would be no statistically significant difference between age, weight, and body mass indexes. Accordingly, the exclusion of subjects over 65 years and under 30 years and overweight or underweight subjects affected the size of the dataset. In the simple snoring group, which is considered healthy, there were fewer subjects than in the other groups because OSA was detected in the vast majority of people who were taken to PSG with the suspicion of OSA. Due to the small number of subjects, it could not be studied on the distinction between men and women.

Another limitation of this study is that the voices used in Turkish were studied since it was applied to people whose mother tongue was Turkish. It is not possible to translate the syllables used directly into English or another language, and the necessity of determining the appropriate voices in other languages is an important limitation. There is a need to repeat the proposed method in other languages by identifying the voices affected by OSA.

VI. CONCLUSION

This preliminary study demonstrates that OSA and OSA severity can be determined with only a small number of non-linear features calculated from a few different speech samples and with quite high accuracy, consistent with clinical assessment. Especially in the healthy/OSA classification, 95.1% accuracy is the highest value obtained in the literature. It is evaluated that a decision support system that can diagnose OSA in a few minutes can be developed in the clinic by testing the proposed method on larger populations.

In the study, some of the consonants formed by the vocal tract components, which are thought to affect voice formation by changing in people with apnea, were used. In future studies, the effect of increasing the number of consonants, using consonants and vowels together, and increasing the vocalization time of consonants on the performance of OSA diagnosis can be investigated. Additionally, optimized classifiers by using various optimization methods and deep learning approaches can be tried to increase the classification performance.

REFERENCES

[1] C. A. Kushida, M. R. Littner, T. Morgenthaler, C. A. Alessi, D. Bailey, J. Coleman, L. Friedman, M. Hirshkowitz, S. Kapen, M. Kramer, T. Lee-Chiong, D. L. Loube, J. Owens, J. P. Pancer, and M. Wise, "Practice parameters for the indications for polysomnography and related procedures: An update for 2005," *Sleep*, vol. 28, no. 4, pp. 499–523, Apr. 2005.

[2] P. Varady, T. Micsik, S. Benedek, and Z. Benyo, "A novel method for the detection of apnea and hypopnea events in respiration signals," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 9, pp. 936–942, Sep. 2002.

[3] A. S. Ng, J. W. Chung, M. D. Gohel, W. W. Yu, K. L. Fan, and T. K. Wong, "Evaluation of the performance of using mean absolute amplitude analysis of thoracic and abdominal signals for immediate indication of sleep apnoea events," *J. Clin. Nursing*, vol. 17, no. 17, pp. 2360–2366, Sep. 2008.

[4] N. Garg, A. J. Rolle, T. A. Lee, and B. Prasad, "Home-based diagnosis of obstructive sleep apnea in an urban population," *J. Clin. Sleep Med.*, vol. 10, no. 8, pp. 879–885, Aug. 2014.

[5] A. F. Quiceno-Manrique, J. B. Alonso-Hernandez, C. M. Travieso-Gonzalez, M. A. Ferrer-Ballester, and G. Castellanos-Dominguez, "Detection of obstructive sleep apnea in ECG recordings using time-frequency distributions and dynamic features," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sep. 2009, pp. 5559–5562.

[6] A. H. Khandoker, M. Palaniswami, and C. K. Karmakar, "Support vector machines for automated recognition of obstructive sleep apnea syndrome from ECG recordings," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 1, pp. 37–48, Jan. 2009.

[7] B. Yilmaz, M. H. Asyali, E. Arikan, S. Yetkin, and F. Özgen, "Sleep stage and obstructive apneaic epoch classification using single-lead ECG," *Biomed. Eng.*, vol. 9, no. 1, pp. 1–14, 2010.

[8] S. Babaeizadeh, D. P. White, S. D. Pittman, and S. H. Zhou, "Automatic detection and quantification of sleep apnea using heart rate variability," *J. Electrocardiol.*, vol. 43, no. 6, pp. 535–541, Nov. 2010.

[9] M. Bsoul, H. Minn, and L. Tamil, "Apnea MedAssist: Real-time sleep apnea monitor using single-lead ECG," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 3, pp. 416–427, May 2011.

[10] S. D. Pittman, N. T. Ayas, M. M. MacDonald, A. Malhotra, R. B. Fogel, and D. P. White, "Using a wrist-worn device based on peripheral arterial tonometry to diagnose obstructive sleep apnea: In-laboratory and ambulatory validation," *Sleep*, vol. 27, no. 5, pp. 923–933, Aug. 2004.

[11] S. Yalamanchali, V. Farajian, C. Hamilton, T. R. Pott, C. G. Samuelson, and M. Friedman, "Diagnosis of obstructive sleep apnea by peripheral arterial tonometry: Meta-analysis," *JAMA Otolaryngol.-Head Neck Surg.*, vol. 139, no. 12, pp. 1343–1350, 2013.

[12] L. Almazaydeh, K. Elleithy, and M. Faezipour, "Detection of obstructive sleep apnea through ECG signal features," in *Proc. IEEE Int. Conf. Electro/Inf. Technol.*, May 2012, pp. 1–6.

[13] A. Yadollahi and Z. Moussavi, "Acoustic obstructive sleep apnea detection," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sep. 2009, pp. 7110–7113.

[14] T. Kim, J.-W. Kim, and K. Lee, "Detection of sleep disordered breathing severity using acoustic biomarker and machine learning techniques," *Biomed. Eng. OnLine*, vol. 17, no. 1, pp. 1–19, Dec. 2018.

[15] H. Nakano, T. Furukawa, and T. Tanigawa, "Tracheal sound analysis using a deep neural network to detect sleep apnea," *J. Clin. Sleep Med.*, vol. 15, no. 8, pp. 1125–1133, Aug. 2019.

[16] Y. Castillo-Escario, I. Ferrer-Lluis, J. M. Montserrat, and R. Jané, "Entropy analysis of acoustic signals recorded with a smartphone for detecting apneas and hypopneas: A comparison with a commercial system for home sleep apnea diagnosis," *IEEE Access*, vol. 7, pp. 128224–128241, 2019.

[17] J. E. Hernandez and E. Cretu, "A wireless, real-time respiratory effort and body position monitoring system for sleep," *Biomed. Signal Process. Control*, vol. 61, Aug. 2020, Art. no. 102023.

[18] T. A. S. Marçal, J. M. dos Santos, A. Rosa, and J. M. R. Cardoso, "OSAS assessment with entropy analysis of high resolution snoring audio signals," *Biomed. Signal Process. Control*, vol. 61, Aug. 2020, Art. no. 101965.

[19] A. M. Benavides, R. F. Pozo, D. T. Toledano, J. L. B. Murillo, E. L. Gonzalo, and L. H. Gómez, "Analysis of voice features related to obstructive sleep apnea and their application in diagnosis support," *Comput. Speech Lang.*, vol. 28, no. 2, pp. 434–452, 2014.

[20] J. Solé-Casals, C. Munteanu, O. C. Martín, F. Barbé, C. Queipo, J. Amilibia, and J. Durán-Cantolla, "Detection of severe obstructive sleep apnea through voice analysis," *Appl. Soft Comput.*, vol. 23, pp. 346–354, Oct. 2014.

[21] J. A. Fiz, J. Morera, J. Abad, A. Belsunces, M. Haro, J. L. Fiz, R. Jane, P. Caminal, and D. Rodenstein, "Acoustic analysis of vowel emission in obstructive sleep apnea," *Chest*, vol. 104, no. 4, pp. 1093–1096, Oct. 1993.

[22] R. F. Pozo, J. L. B. Murillo, L. H. Gómez, E. L. Gonzalo, J. A. Ramírez, and D. T. Toledano, "Assessment of severe apnoea through voice analysis, automatic speech, and speaker recognition techniques," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, pp. 1–11, Dec. 2009.

[23] E. Goldshtein, A. Tarasiuk, and Y. Zigel, "Automatic detection of obstructive sleep apnea using speech signals," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 5, pp. 1373–1382, May 2011.

[24] M. Kriboy, A. Tarasiuk, and Y. Zigel, "Obstructive sleep apnea detection using speech signals," in *Proc. Annu. Conf. Afeka-AVIOS Speech Process.*, 2013, pp. 1–5.

[25] M. Kriboy, A. Tarasiuk, and Y. Zigel, "Detection of obstructive sleep apnea in awake subjects by exploiting body posture effects on the speech signal," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 4224–4227.

[26] R. M. Simply, E. Dafna, and Y. Zigel, "Diagnosis of obstructive sleep apnea using speech signals from awake subjects," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 251–260, Feb. 2020.

[27] F. Espinoza-Cuadros, R. Fernández-Pozo, D. T. Toledano, J. D. Alcázar-Ramírez, E. López-Gonzalo, and L. A. Hernández-Gómez, "Speech signal and facial image processing for obstructive sleep apnea assessment," *Comput. Math. Methods Med.*, vol. 2015, pp. 1–13, Nov. 2015.

[28] F. Espinoza-Cuadros, R. Fernández-Pozo, D. T. Toledano, J. D. Alcázar-Ramírez, E. López-Gonzalo, and L. A. Hernández-Gómez, "Reviewing the connection between speech and obstructive sleep apnea," *Biomed. Eng. OnLine*, vol. 15, no. 1, pp. 15–20, Dec. 2016.

[29] J. L. Blanco, L. A. Hernández, R. Fernández, and D. Ramos, "Improving automatic detection of obstructive sleep apnea through nonlinear analysis of sustained speech," *Cognit. Comput.*, vol. 5, no. 4, pp. 458–472, Dec. 2013.

[30] Y. Ding, J. Wang, J. Gao, Q. Fang, Y. Li, W. Xu, J. Wu, and D. Han, "Severity evaluation of obstructive sleep apnea based on speech features," *Sleep Breathing*, vol. 25, no. 2, pp. 787–795, Jun. 2021.

[31] Y. Ding, Y. Sun, Y. Li, H. Wang, Q. Fang, W. Xu, J. Wu, J. Gao, and D. Han, "Selection of OSA-specific pronunciations and assessment of disease severity assisted by machine learning," *J. Clin. Sleep Med.*, vol. 18, no. 11, pp. 2663–2672, Nov. 2022.

[32] D. Yilmaz, M. Yildiz, Y. U. Toprak, and S. Yetkin, "Obstructive sleep apnea detection with nonlinear analysis of speech," *Biomed. Signal Process. Control*, vol. 84, Jul. 2023, Art. no. 104956.

[33] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[34] R. B. Berry. (2018). *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications Version 2.5*. American Academy of Sleep Medicine, Darien, IL, USA. [Online]. Available: https://aasm.org/wp-content/uploads/2018/04/Summary-of-Updates-in-v2.5-1.pdf

[35] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Warwick*. Berlin, Germany: Springer, 1980, pp. 366–381.

[36] H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, and L. S. Tsimring, "The analysis of observed chaotic data in physical systems," *Rev. Modern Phys.*, vol. 65, no. 4, pp. 1331–1392, Oct. 1993.

[37] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Phys. Rev. A, Gen. Phys.*, vol. 45, no. 6, pp. 3403–3411, Mar. 1992.

[38] G. Williams, *Chaos Theory Tamed*. Washington, DC, USA: Joseph Henry Press, 1997.

[39] P. Grassberger and I. Procaccia, "Characterization of strange attractors," *Phys. Rev. Lett.*, vol. 50, no. 5, pp. 346–349, Jan. 1983.

[40] J. Theiler, "Efficient algorithm for estimating the correlation dimension from a set of discrete points," *Phys. Rev. A, Gen. Phys.*, vol. 36, no. 9, pp. 44–56, 1987.

[41] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Phys. D, Nonlinear Phenomena*, vol. 31, no. 2, pp. 277–283, Jun. 1988.

[42] M. T. Rosenstein, J. J. Collins, and C. J. De Luca, "A practical method for calculating largest Lyapunov exponents from small data sets," *Phys. D, Nonlinear Phenomena*, vol. 65, nos. 1–2, pp. 117–134, May 1993.

[43] A. Delgado-Bonal and A. Marshak, "Approximate entropy and sample entropy: A comprehensive tutorial," *Entropy*, vol. 21, no. 6, p. 541, May 2019.

[44] R. R. Coifman, Y. Meyer, S. Quake, and M. V. Wickerhauser, "Signal processing and compression with wavelet packets," in *Wavelets and Their Applications*. 1990, pp. 363–379.

[45] C. L. Nikias and A. P. Petropulu, *Higher-Order Spectra Analysis: A Nonlinear Signal Processing Framework*. NJ, USA: Prentice-Hall, 1993.

[46] R. V. Hogg and J. Ledolter, *Engineering Statistics*. New York, NY, USA: MacMillan, 1987.

[47] Z. Xiong, Y. Cui, Z. Liu, Y. Zhao, M. Hu, and J. Hu, "Evaluating explorative prediction power of machine learning algorithms for materials discovery using *k*-fold forward cross-validation," *Comput. Mater. Sci.*, vol. 171, Jan. 2020, Art. no. 109203.

[48] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explor. Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.

[49] P. H. Godoy, A. P. D. S. Nucera, A. D. P. Colcher, J. E. de-Andrade, and D. D. S. B. Alves, "Screening for obstructive sleep apnea in elderly: Performance of the Berlin and STOP-bang questionnaires and the Epworth sleepiness scale using polysomnography as gold standard," *Sleep Sci.*, vol. 15, pp. 203–208, 2022.

[50] D. V. E. Silva, D. D. S. B. Alves, M. D. R. R. Nóbrega, F. B. C. G. Ribeiro, L. Y. Y. Franco, I. R. D. Silva, L. Joffily, M. D. Rocha, and P. H. Godoy, "Obstructive sleep apnea screening in different age groups: Performance of the Berlin, STOP-bang questionnaires and Epworth sleepiness scale," *Brazilian J. Otorhinolaryngol.*, vol. 89, no. 4, Jul. 2023, Art. no. 101283.

TUĞÇE KANTAR UĞUR (Member, IEEE) was born in Ankara, Turkey, in 1992. She received the B.S. and M.S. degrees in biomedical engineering from Başkent University, Ankara, in 2014 and 2017, respectively, where she is currently pursuing the Ph.D. degree in biomedical engineering.

From 2014 to 2023, she was a Research Assistant with the Department of Biomedical Engineering, Başkent University. Her primary research interests include biomedical signal processing and artificial intelligence.

DERYA YILMAZ received the Ph.D. degree in electronics and computer technology from Gazi University, Ankara, Turkey, in 2008.

From 2008 to 2021, she was an Assistant Professor with the Department of Electrical and Electronics Engineering, Başkent University, Ankara. Since 2021, she has been an Associate Professor with the Department of Electrical and Electronics Engineering, Faculty of Engineering, Gazi University. Her main research interests include chaotic analysis, signal processing, image processing, biomedical systems, and machine learning.

METİN YILDIZ received the Ph.D. degree in electronic engineering from Selçuk University, Konya, Turkey, in 2006.

From 2007 to 2019, he was an Assistant Professor in biomedical engineering with Başkent University, Ankara, Turkey. He is currently an Associate Professor with the Department of Biomedical Engineering, İzmir Democracy University, İzmir, Turkey. His primary research interests include biomedical instrumentation and biomedical signal processing.

SİNAN YETKİN graduated from the Gülhane Military Medical Academy, in 1992. He completed a specialization qualification in psychiatry, in 1999, and later received training in polysomnography and sleep medicine from the Sleep Research Center, Gülhane Military Medical Faculty.

He is currently a Professor with the Gülhane Faculty of Medicine, University of Health Sciences. His primary research interests include electrophysiology of sleep and sleep disorders.