Journal of Biomedical Informatics 95 (2019) 103218

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Abnormal respiratory event detection in sleep: A prescreening system with smart wearables

Burçin Camcı[a,*], Cem Ersoy[a], Hakan Kaynak[b]

[a] NETLAB, Computer Networks Research Laboratory, Department of Computer Engineering, Bogazici University, 34342 Istanbul, Turkey
[b] Uykum Sleeping Center, 34365 Istanbul, Turkey

ARTICLE INFO

Keywords:
Respiratory event in sleep
Prescreening system
Smart phone
Smart watch
Machine learning
Pervasive health

ABSTRACT

Sleeping is an important activity to monitor since it has a crucial role in the overall health and well-being of the people and society. In order to diagnose the problems in sleep, different monitoring systems are developed in the literature. The unobtrusiveness, reduced cost, objectiveness, protection of privacy and user-friendliness are the main design considerations and the proposed system design achieves those objectives by utilizing smart wearables, smart watch and smart phone. The accelerometer and heart rate monitor sensors on smart watch and the sound level sensor on the smart phone are activated. The experiments with this system are performed with 17 subjects in a sleep clinic. The data collected from these subjects is used to generate various combinations by employing varied feature extraction, feature selection and sampling approaches. Five different machine learning algorithms are implemented and the classification results are generated using the various combinations of data, training and scoring strategies. The system performance is measured in two ways, the accuracy rate of distinguishing abnormal respiratory events is 85.95% and the classification success of subjects according to the problems in their respiration is one misclassification among 17 subjects. With all the methodology utilized in this study, the proposed system is a novel prescreening tool which recognizes the severity of problems in respiration during sleep.

1. Introduction

The raising awareness of the personal health motivates people to analyze their daily routine in detail. With the help of development in the technology, personal healthcare applications can generate a detailed analysis of the activities performed during the day. Since sleeping occupies about one-third of people's lifetime [1], it has a great potential to indicate the overall health and well-being of the people.

The analysis of the sleep requires a complicated investigation of multiple factors since the quality of sleep depends on various factors. The environmental factors like sleeping in a dark and silent room are the simplest requirements for obtaining good sleep quality. The personal factors like the lifestyle, working conditions and habits of people are more challenging and hard to change. Furthermore, people who have medical problems like cancer, depression, obesity can suffer from poor sleep quality [2,3].

While unhealthy people can have poor sleep quality, poor sleep quality can cause cognitive impairments and physical diseases like high blood pressure, heart diseases, diabetes, stroke, somnolence, arrhythmia [4]. Sleep provides refreshment from mental and physical fatigue. Sleep with good quality has a positive impact on the immunity, human's enthusiasm, learning process, creativeness and mental concentration level during the day [5]. The latter has a strong relationship with the public health and well-being. Sleep is not only important for individual health because sleep deprivation may result in motor vehicle crashes, industrial disasters and occupational errors. The National Highway Traffic Safety Administration has stated that the drowsy driving is responsible for 100,000 car accidents, 40,000 injuries and 1,550 deaths per year [6].

In order to measure the poor sleep quality, the common symptoms like sleep duration, apnea count, snoring, nightmares, leg movements during the night are needed to be investigated. Many people are suffering from different sleep disorders with different symptoms. In the literature, most common sleep disorders are stated as insomnia, sleep apnea, restless leg syndrome, narcolepsy, snoring, sleepwalking and terrors [7,8]. However, these symptoms reveal themselves in each person's sleep in different ways like not sleeping for long hours, feeling tired all the time or having apneas.

By considering all those facts, taking immediate intervention strategies to diagnose sleep-related disorders is obligatory. The main

* Corresponding author.
E-mail address: burcincamci@gmail.com (B. Camcı).

https://doi.org/10.1016/j.jbi.2019.103218
Received 4 January 2019; Received in revised form 24 April 2019; Accepted 23 May 2019
Available online 25 May 2019
1532-0464/

problem about sleep deprivation due to sleep-related disorders is that they have mostly stayed undiagnosed since people are unconscious in their sleep. In this manner, sleep monitoring systems play a significant role in determining the underlying reasons behind the poor sleep quality of subjects.

In order to diagnose sleep-related diseases, mainly two types of monitoring systems are developed. The first type which is more subjective diagnoses sleep-related diseases according to the information provided by the subject or a relative. The second type that is more objective investigates the disease by monitoring subjects during sleep. If both monitoring types are compared, the second type is much more reliable since the questions and the answers in the first type are subjective and the results can change according to the subject's current mood. The first type contains different standardized questionnaires whereas in the recent studies, researchers include new questions to the standardized ones according to their studies' objectives.

In the second type, the assessment is handled with actual sensors. Polysomnography (PSG), ActiGraph and several other smart devices are the most useful technologies. The gold standard to diagnose sleep-related problems of a subject is to use the PSG technique. However, since this technique is expensive, overwhelming and time-consuming, the subjects who afford or volunteer this examination are few compared to the potential number of subjects. Even though this system has some disadvantages, it provides the most reliable results with today's technology and therefore, it is the most preferred solution.

All in all, existing techniques need to be improved since they are time-consuming, expensive, obtrusive and subjective which is because of misleading answers of the subjects. Therefore, a new mechanism for monitoring sleep is required. The system should be unobtrusive, cheap, user-friendly and protect privacy. Also, the subjects can be monitored during the sleep in their own conditions without any intervention. By considering those conditions, the proposed system contains a smart phone and a smart watch. These smart devices are equipped with many useful built-in sensors like accelerometer, sound level, and heart rate monitor. Those sensors can be utilized to monitor human activities since they provide raw signals. With the increased Central Processing Unit (CPU) power and storage area, these smart devices can deal with complex operations easily and they are very unobtrusive compared to most of the other systems. Furthermore, most of the people already have smart devices and get used to them in their daily life. Sleep monitoring systems which require smart devices need only software development. Therefore, the additional cost of the system is very low.

The main objective of the proposed system is the *assessment of respiratory information*, since the detection of sleep disorder without consulting a doctor and with a smart device may have devastating consequences. These statistics eventually help to improve the sleep quality of the subject. The proposed system prepares some preliminary information or statistic about the subjects and this is very practical for both the subjects and the doctors. The subjects can understand the severity of their problems in respiration before consulting a doctor and the doctors can reach objective observations about their patients. Therefore, this system is actually positioned as a prescreening tool or guidance system which is a novel solution in the field by considering the system design, the data collection, the ground truth validation and the methodology utilized in this study.

This study firstly discusses the related works in the context of sleep analysis domain by considering utilized devices and study objectives in Section 2. Afterward, the background information about the respiratory events and the PSG technique is provided. In Section 3, the prescreening system design is given with its limitations and considerations. Section 4 starts with the respiratory system design. It explains the experimental setup and the data collection procedure for each device in the system and provides a detailed analysis of the subjects. It is continued with the visualization of data and explanation of the abnormal respiratory event and its arousal. The challenges through generating meaningful data are presented and the feature extraction, selection and sampling processes are introduced. Section 5 starts with the classification methodology of different algorithms and results of these algorithms. Afterward, the enhancements on the current results are discussed and different approaches are presented. In Section 6, the system performance is evaluated as a prescreening tool. Lastly, in Section 7, conclusions and directions for future research are stated.

2. Literature survey

This section provides fundamental information about sleep studies in the literature which are evaluated according to the similarity between their objective and utilized devices with this study. Furthermore, the background concepts which are implied throughout this study and the utilized devices which are the key components of the proposed system design are explained.

2.1. Related sleep studies

Sleep analysis has been a popular research field for a long time. Since sleep does not affect only the individuals' overall health and well-being but also the society's, understanding the underlying reasons for the poor sleep quality is significant. The extensive search in the field proves that the sleep can be examined in several aspects like study types, study objectives, utilized sensors, experiment design, extracted features, methodologies and achieved results. Among all those aspects, the study objective is the main measure that determines the other aspects.

By considering the study objective, the utilized sensors in the studies are altered. The sleep studies gather information from several sources which can be build-in smart phones or smart watches or some external sensors. The common smart phone or smart watch sensors are accelerometer, gyroscope, sound level, light, global positioning system (GPS), heart rate monitor, screen proximity and even clock. The general external sensors are pulse oximeter, ECG, earphone, etc. and if this is the case, then the smart devices are just a medium for transferring or interpreting the data. Furthermore, all studies should be validated with the ground truth data which is more accurate data coming from other sources. The researchers collect the ground truth data from questionnaires, diaries or applications which are filled by the subjects themselves or their relatives however it is not trustworthy because of the subjectivity. Some studies get help from a physician for the ground truth which is much more trustworthy and the most respected ground truth strategy is monitoring patients with the PSG system which is the gold standard for sleep studies. Since PSG systems are not available everywhere and only one person can use the system each night, a subset of PSG sensors or other simpler commercial products can also be used for lower quality ground truth. Furthermore, a medical database can also be a ground truth source.

In this section, the related works in the field of sleep are presented in detail and the studies are chosen according to the objective of this study, respiratory event detection and the utilized sensors in this study, the accelerometer, the heart rate monitor and the sound level sensor on smart devices.

The sleep disorder detecting studies mainly investigate respiratory events like apnea. The studies conducted by [9–12] include just a smart phone and the built-in sensors are used to collect data from subjects. In [9], the breathing and movement pattern of the subjects are learned from the accelerometer sensor and microphone. Since the smart phone collects information about the subject himself and not the environment, the smart phone is placed on arm, abdomen or near on the bed. The ground truth data is determined by the physician and the results send as a text message to the subjects. The proposed system in [10] attaches the smart phone to the anterior chest wall over the sternum in order to collect data from the microphone and utilizes the PSG system for the ground truth. The statistical method is implemented for diagnosing apnea sensitivity and specificity which are 0.70 and 0.94, respectively.

In [11], two different microphones are utilized, one placed in the smart phone is for the data collection in different contexts like in the pocket or on the desk and another external one is for the ground truth. The respiratory symptom classification accuracy 82% is achieved with the support vector machine algorithm. In [12], a smart phone is used but the placement of the smart phone and activated sensors are not given. For detecting the apnea with 96.85% accuracy, a decision tree algorithm was implemented and a medical database is used for the ground truth.

Another group of sleep disorder detecting studies activate the smart phone sensors with other external devices [13–15]. In [13], the smart phone is used to record audio, activity and body position of the subjects whereas the oxygen saturation is measured with an external pulse oximeter sensor. A physician provides the ground truth data and the apnea-healthy classification accuracy is determined as 92.2% with the support vector machine algorithm. The proposed system defined in [14] follows the same strategy for data collection with one difference. An external microphone also measures the respiratory effort. 15 patients are participated in the experiments and the results show that one healthy individual incorrectly diagnosed as having the disease which means that the accuracy for detecting apnea is evaluated as 87.5% with some heuristic methods and the ground truth is obtained from the PSG system. In [15], the proposed system records audio and accelerometer signals and also it makes use of external ECG for measuring respiratory rate. The objective is not just disorder detection but also assessing the sleep quality. A medical database is the ground truth source and the average sleep quality detection accuracy is determined as 73% with the support vector machine algorithm.

In [16] different aspects of sleep are investigated and the heart rate is estimated with accelerometer data on the smart phone. The system collects the ground truth with a diary and distinguishes the diary results lower than the computed results with the help of a heuristic approach. The respiration rate is analyzed in [17] with the accelerometer sensor in the smart watch and video recording is used for validation. Four different machine learning algorithms, Bayesian Network, Decision Tree, Random Forest and Naive Bayes are utilized for detecting the respiration rate with 95% accuracy.

In general, the conducted studies generate high-level features based on raw signals. The extracted features for each study are distinguished according to the study objective and utilized sensors. The main tendency is calculating magnitude, minimum, maximum, average and other range variables of the accelerometer signals [9,14–16]. Calculating spectral centroid and roll off is another common approach and those features are useful for noise analysis [11]. For the heart rate signals, the amplitude and peak information are widely used.

The analysis of sleep studies shows that each study has different objectives, focuses on different aspects of sleep while using different sensors and applies different methods. As a result, the obtained results of sleep studies differ from each other. Most of them focus on the accuracy of the proposed system however they measure the accuracy of different metrics.

The proposed system design actually combines several ideas which partially appear in other studies. The smart phone is combined with the smart watch like the combination of the pulse oximeter sensor in [13] and an external microphone in [14]. However, the system in [13] differs from the proposed system with respect to the ground truth source. Also, the method for measuring the system performance in [14] is different from the proposed method in this study. By observing the studies mentioned above, the general tendencies of the feature extraction, feature selection and machine learning approaches in sleep studies are identified. Furthermore, the approaches which are more useful for this study are implemented by considering the study objective and the utilized sensors.

2.2. Background concepts of the proposed system

The background concepts explain the main components and concepts of the proposed system. The information provided in this section contains the fundamentals of this study.

2.2.1. Respiratory event

The respiratory events that are mentioned in this study are apnea events. The sleep apnea can be defined as pauses in breathing or shallow breaths in sleep [18]. The breathing pauses can last from a few seconds to minutes. In an hour, thirty or more apnea events might occur. A loud snort or choking sound is the sign of the start of the normal breathing. The oxygen saturation level change and the heart rate change are the clear indicators of apnea events. Since the body gives response afterward, the body movements and loud snoring follow the apnea events. The most common symptom of obstructive sleep apnea (OSA) is loud and chronic snoring. However, it cannot be stated that all the snoring people suffer from sleep apnea disorder. Another common symptom is the daytime sleepiness which may have more devastating consequences. The treatment of sleep apnea disorder requires important lifestyle changes and/or utilizing some breathing devices while sleeping.

The sleep apnea patients can be classified according to the severity of their illness. In the literature, the severity of apnea is calculated with Eq. (1) and it is called the Apnea-Hypopnea Index (AHI) [19].

$$AHI = \frac{ApneaEventCount}{TotalSleepDuration(hours)} \tag{1}$$

According to the AHI, the subjects are divided into 3 groups.

- Mild: Subjects whose AHI is smaller than 15.
- Moderate: Subjects whose AHI is greater than 15 and smaller than 30.
- Severe: Subjects whose AHI is greater than 30.

The main problem of sleep apnea is that the disease can be left as undiagnosed because it happens in sleep and the person can be unaware. Therefore, when analyzing the sleep of a person who suffers from low quality, the diagnosis of sleep apnea disorder is crucial. While in the sleep apnea period, a person often moves out of deep sleep into a light sleep which directly affects the quality of sleep [20].

2.2.2. Polysomnography

The polysomnography is the gold standard tool for sleep analysis. It provides information about body functions during sleep and it is mainly used to diagnose sleep disorders. The PSG systems are available in hospitals or clinics and the patient comes one or two hours earlier for the introduction of the system, preparation and placement of electrodes. The sleep technician is responsible for the preparation process and after the patient falls asleep, he follows the patient all night long from video and the computer that displays all the collected data. When the patient wakes up in the morning, the monitoring ends [21].

The collected data can be automatically analyzed with the help of software by considering the previously entered parameters which are determined by the American Academy of Sleep Medicine (AASM) [22]. However, since the automatic analysis results are not precise, the specialist doctor can handle the analysis manually which provides more accurate outcomes.

Some patients who have respiratory problems are sometimes assessed for two nights if the patient can be treated with an extra mask, Continuous Positive Airway Pressure (CPAP) mask. It is useful for understanding the oxygen saturation level which is different for each person and placed on the mouth in the second night. The subjects are recommended to wear the mask with the determined oxygen saturation level for all nights in the rest of their lives [23].

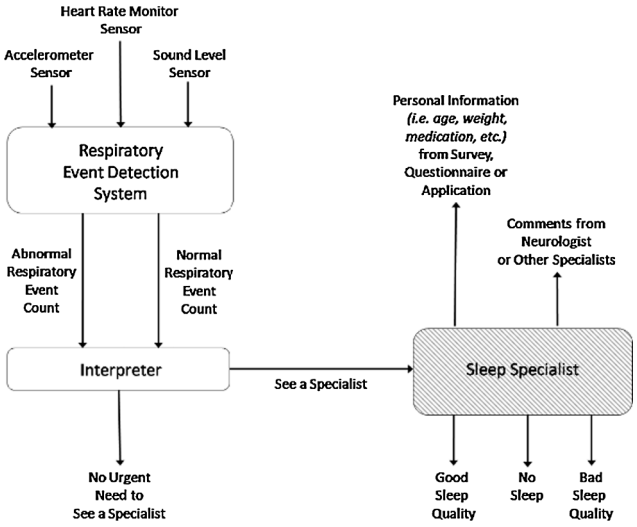


Fig. 1. Prescreening system design.

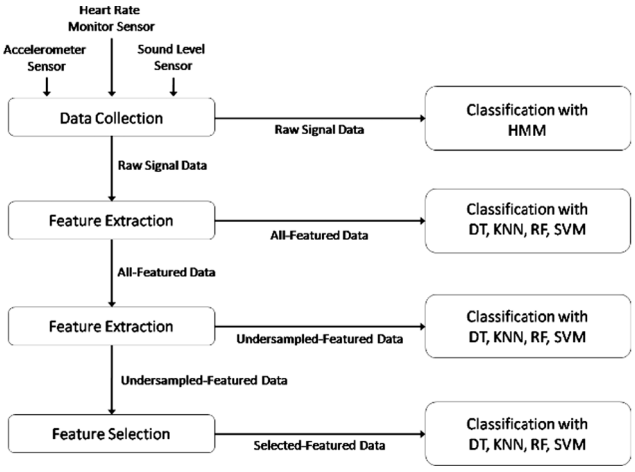


Fig. 2. Respiratory event detection system description.

3. Prescreening system architecture

The prescreening system helps to recognize people with severe respiratory problems and aims to give suggestions like consulting a doctor. The system mainly contains three components and as seen in Fig. 1. The *"Respiratory Event Detection System"* is designed to distinguish abnormal and normal respiratory events. While the *"Interpreter"* decides whether the patient needs to see a specialist urgently, the *"Sleep Specialist"* represented with a shaded box analyzes the findings of the Respiratory Event Detection System with other information about patients.

The main requirement while designing the prescreening system is collecting the most meaningful data from the subject while not ignoring the objectives, unobtrusiveness, reduced cost, objectiveness, protection of privacy and user-friendliness. These are the objectives that make the system more desirable than other monitoring devices especially the PSG technique. In order to reach these objectives, the system is preferred to contain a smart phone with a smart watch that is placed on the wrist of the subject. Since people use these devices in their daily routine, the data collection in the subject's own conditions can be handled without any help of an expert. Also, raw signals reached with the data collection eliminate the subjectiveness and protect the privacy.

The choice of these devices cause some limitations. The battery limitation for the smart watch and the limited number of activated sensors prevent the system to obtain information about all sleep-related diseases. This situation forces this study to focus on only abnormal respiratory events like apnea. Even if all aspects of sleep is not covered, determining the severity of problems in the subject's respiration during sleep is an absolute indicator of sleep quality.

4. Respiratory event detection system design

The system aims to distinguish abnormal respiratory events by interpreting raw signals coming from real subjects. The five components of this system are presented in Fig. 2. The data collection process with each utilized device is presented in the figure.

The data coming from three different sources, sound level, heart rate monitor and accelerometer sensors are merged with the true labels coming from the PSG system. This combination is referred as *"Raw Signals"*. The *"Raw Signals"* are processed and some meaningful features are extracted. This is stated as *"All-Featured Data"* that contains all generated features. In order to balance the normal and abnormal respiratory events, the data is undersampled without eliminating any features and this data is called *"Undersampled-Featured Data"*. The feature selection algorithms, Random Forest and Stepwise Feature Selection are implemented to understand the most important features and the data obtained from these algorithms are named as *"Selected-Featured Data"*. These steps are extensively explained in the related sections.

The system performance is measured with the classification accuracy. Five different machine learning algorithms namely Hidden Markov Model (HMM), Decision Tree (DT), K-Nearest Neighbors (KNN), Random Forest (RF) and Support Vector Machine (SVM) are utilized for this purpose. The derived results are generated with different subsets of the raw and featured data. All implementation details are clearly explained in the following sections and because of its computational power, MATLAB is preferred for the implementation with one exception. The classification with raw signals is implemented in Python.

4.1. Data collection with the proposed system

Subject-specific and environmental data are collected with the help of three different devices and the collected data is manually transferred to the computer afterwards. Each device is explained in detail in the following sections.

4.1.1. Smart watch

In this study, the subject-specific information is collected with the help of a smart watch, Samsung Gear S3 [24]. Besides the general advantages of smart watches, this watch is preferred since it gives permission to access the collected raw signals and alter the sampling rate of sensors. The accelerometer, gyroscope and heart rate monitor sensors are activated for this study and the users can be tracked all-day-long to the extent permitted by the battery with the help of an application developed by our research group.

The battery life of the smart watch is the main drawback. In order to obtain more precise data, the sampling rate of sensors is increased with the application which causes a decrease in the life of the battery. Since the charging unit is connected to the watch from the bottom side, it is not possible to recharge the smart watch without data interruption. Furthermore, the smart watch contains the sound level sensor. However, it is not activated since the subject can place his wrist under the blanket which may interrupt the data collection. By deactivating the sensors which do not serve this study's purposes, the battery life is tried to be extended.

The smart watch is worn on the wrist as tight as possible in order to prevent data loss as shown in Fig. 3. The sampling rate of the accelerometer signals is 20 Hz whereas the heart rate monitor sensor's sampling rate is 10 Hz. The battery lasts approximately 6 h of

Fig. 3. Smart watch on the wrist.

continuous data collection with these sampling rates.

4.1.2. Smart phone

The environmental information is collected by the smart phone. An Android Application, Smart Voice recorder [25], that is available in the market records the voice data of the subject and the environment. The sampling rate of this voice data is 16,000 Hz. Since the experiments are conducted in a controlled room in the sleep clinic, collecting information about the other variables like humidity, light, temperature is unnecessary. In order to relieve the subjects from their concerns about the electromagnetic radiation they might be exposed to, the airplane mode of the smart phone is turned on.

4.1.3. Polysomnography

The data coming from the PSG system is used as the ground truth. The PSG system in the sleep clinic was a SomnoStar Sleep System [26]. This system is capable of gathering information from all of 64 electrodes and the electrode selection can change according to the objective of monitoring which is determined by the specialist doctors. The count of electrodes utilized in this study was 18 which are determined according to the guideline provided by the AASM. The placed electrodes on the bodies of subjects with detailed information are shown in Table 1.

In this study, the sleep specialist manually generates the ground truth labels for each epoch (30 s period) with a user-friendly interface shown in Fig. 4. The PSG system allows extracting the raw signals in the European Data Format (EDF) which is a traditional method used for storing biological and physical signals [27]. Time information, sleep stages, sleep/wake times, turnovers, leg movement events and respiratory events with their types are specified by the sleep specialist which are stored in another file.

4.1.4. Experimental setup and subject information

The experiments are conducted in the sleep clinic and the procedure is approved by the Ethical Committee of Bogazici University (INAREK). The sleep technician prepares the subject for the PSG system. After all the electrodes are attached, the smart phone is positioned on the nearest place and the subject wears the smart watch. The subjects are not allowed to read or watch anything after the preparation is completed. Fig. 5 represents a subject who is ready to sleep.

The subjects are volunteer patients of the sleep specialist doctor and

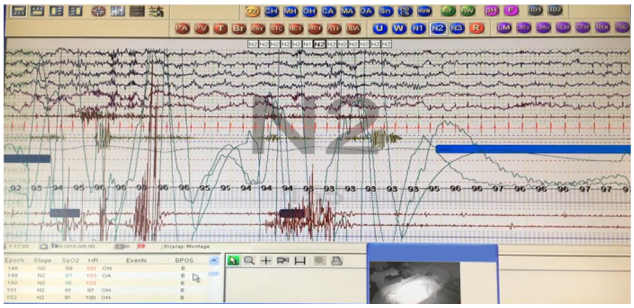


Fig. 4. Interface of the PSG system.

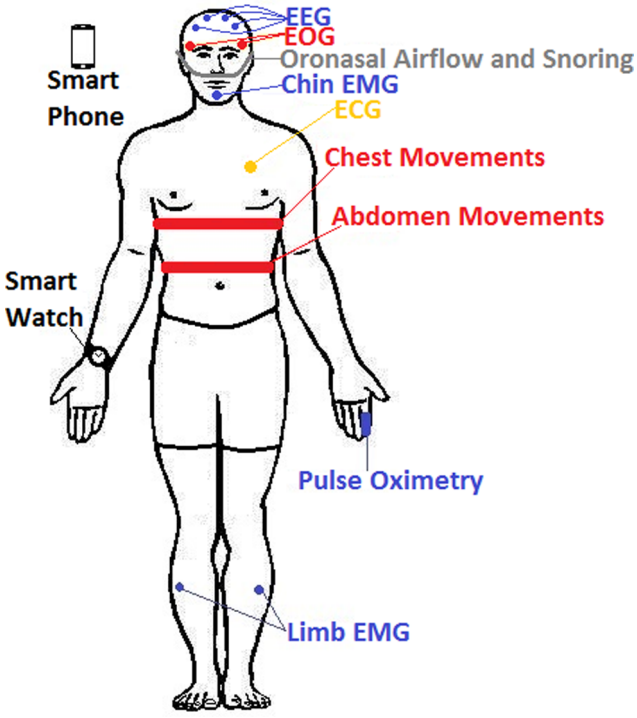


Fig. 5. A subject with the PSG system and smart devices.

chosen according to the main criteria, *"Detecting Abnormal Respiratory Events"*. In order to emphasize the difference between the subjects; the subjects who have severe apnea and the ones who sleep with a CPAP mask is preferred. This mask is a special device for apnea patients and mainly aims to prevent abnormal respiratory events and increase comfort in breathing while sleeping. As a result, no or a few abnormal respiratory events are observed while monitoring the subjects with a CPAP mask. The severe apnea patients are selected by considering their true diagnosis, namely Apnea-Hypopnea Index (AHI). This index is

Table 1
Information about electrodes used in the polysomnography system.

Parameter	Channels	Placement	Recording
Electroencephalography (EEG)	4	On face and scalp	Brain waves
Electrooculogram (EOG)	2	Adjacent to eyes	Eye movements
Chin Electromyogrpahy (EMG)	1	On chin	Mouth movements
Limb Electromyography (EMG)	2	On lower limb	Muscle tone
Electrocardiography (ECG)	1	On chest	Heart rate
Thoracoabdominal Movements	4	Around chest and abdomen	Chest/abdomen movements and breathing
Pulse Oximetry	2	On finger	Blood oxygen level
Oronasal Airflow	1	Near mouth/nose	Airflow
Snoring	1	Microphone	Snoring/breathing sounds
Video	–	Digital night camera	Whole body activities

Table 2
Information about subjects.

ID	Gender	Age	Status	Duration	True AHI
311	Male	46	Normal	5 h and 10 min	0
313	Male	51	Normal	6 h and 15 min	12
316	Male	70	Abnormal	6 h and 44.5 min	62
323	Male	30	Abnormal	8 h and 36.5 min	55
325	Male	30	Normal	7 h and 18.5 min	0
326	Female	47	Abnormal	6 h and 57 min	34
327	Male	48	Abnormal	6 h and 51.5 min	61
328	Male	47	Abnormal	7 h and 41 min	31
331	Male	45	Abnormal	6 h and 52 min	70
333	Male	48	Normal	6 h and 39.5 min	0
334	Male	45	Normal	6 h and 28 min	2
337	Male	47	Normal	5 h and 37.5 min	0
338	Male	76	Abnormal	5 h and 35 min	89
341	Male	30	Abnormal	5 h and 38.5 min	32
344	Male	27	Normal	6 h and 44.5 min	0
349	Male	43	Normal	6 h and 1 min	11
353	Male	39	Normal	6 h and 29.5 min	0

determined by the sleep specialist who examines the results of the PSG system.

The summary of all subjects who attended the experiments and met the requirements are given in Table 2. Because of data interruption and data loss caused by the battery limitation, the location and tightness of smart watch, the collected data becomes useless for some patients. Also, some patients who feel uncomfortable sleeping with the PSG system left the sleep clinic in the middle of the night. Furthermore, the subjects with moderate True AHI score were not chosen in order to distinguish the abnormality clearly. As a result, nearly 15 subjects' data are excluded and the study continued with the remaining 17 subjects, 8 of them have abnormal respiratory events while sleeping and 9 of them have only normal respiratory events with a CPAP mask. The true AHI score which is presented in Table 2 is the main criteria while grouping the subjects as normal and abnormal. The subjects who are stated as *"Normal"* can also have a few abnormal respiratory events.

As represented in Table 2, the average duration of sleep with the system is calculated as approximately 6 h and 38 min. The total duration of sleep for all *"abnormal"* subjects is 55 h, 46 min which is balanced with the duration for all *"normal"* subjects, 56 h and 43.5 min. The mean of ages is 45.24 whereas the standard deviation of ages is 12.97. All subjects are Caucasian and there is only one female subject participated in the study.

4.2. Abnormal respiratory event visualization

The arousal of an abnormal respiratory event is described as presented in Fig. 6 by the sleep specialist. This description is based on the raw signals coming from sound level, heart rate monitor and accelerometer sensors.

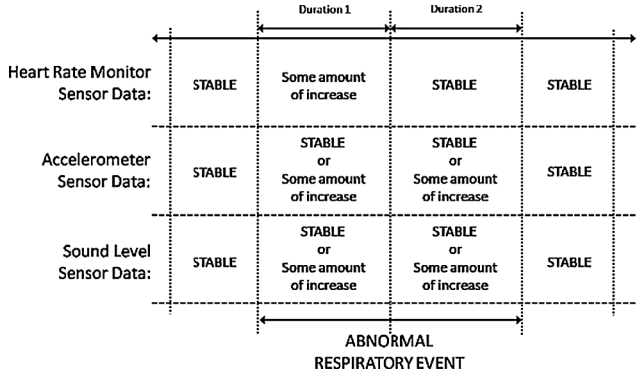


Fig. 6. Arousal of an abnormal respiratory event.

Initially, the subject has a normal sleeping period which is followed by an increase in the heart rate. No breathing or shallow breathing causes this increase and the body also gives a response in the same period (Duration 1) or in the next period (Duration 2). Because of that, the accelerometer and noise signals can vary in Duration 1 or Duration 2. The lengths of these durations can be seconds to minutes and change from subject to subject.

The abnormal respiratory events can be observed directly from the raw signals. Fig. 7(a), (b) and (c) present the heart rate, noise and accelerometer signals, respectively. *"A"* represents the epoch labeled as the abnormal respiratory event by the specialist doctor. The increase in the heart rate signal can be an indicator of the abnormality followed by the fluctuations in the noise and accelerometer signals. The fluctuations generally start in the epoch in which the increase in heart rate signal occurs and end in the same or the next epoch. The fluctuations in the noise signal represent the snoring or coughing events.

4.3. Data processing challenges

The collected data is combined for the next steps however some challenges are encountered while processing. The first challenge is data synchronization. All devices cannot be initiated at the same moment. Also, the smart watch always provides less duration of data with respect to others because of its battery limitation. In addition, the raw signals coming from the smart watch have some interruptions since the position and tightness of the smart watch are changed because of the movements. The time information is the key point while identifying the overlapping sections. In Table 2, the data collection duration for each subject is listed.

The other challenge that needs to be solved is that the data is very large. The size of data for each subject is around 1.9 GB which causes large data processing times. The data coming from different sources cannot be processed separately since the algorithms in further steps require a combination. An efficient algorithm firstly decides the range of data by considering the time information and prunes it with respect to the determined time range. Afterward, the data is relayed to the further steps.

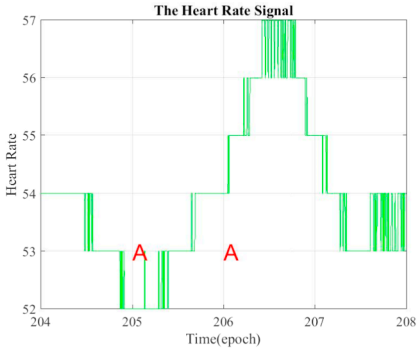
The last challenge is that the PSG system provides the ground truth data in a traditional format. Since it is a commercial product, the manufacturer company does not permit to access to all files which creates a problem in the process of extracting labels. Although the software shows the exact start time and duration of the abnormal respiratory event, the obtained file has only the duration information and does not have the exact starting moment of the abnormal respiratory event. For that reason, the label information has some inaccuracies. The ground truth labels contain an epoch as the abnormal respiratory event, even if the event starts in the last second of that specific epoch.

4.4. Feature extraction

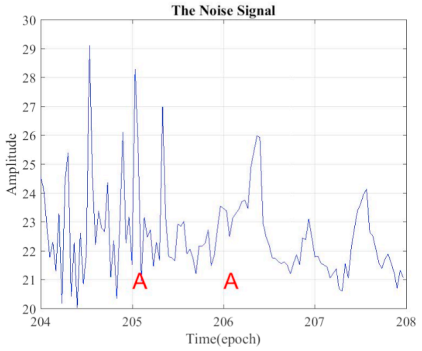
The raw signals obtained from two different sources are firstly divided into 30 s long windows since each label in the ground truth data is generated for that specific time period by the PSG system. Depending on the source type and considering the intrinsic properties and the literature survey, different features are extracted. Totally 103 features are generated for all different sensors and the summary of all extracted features is given in Table 3 which is followed by the detailed explanation of features.

The minimum, maximum, mean, variance, energy and entropy of data provide the general range of the accelerometer data. Zero Crossing Rate (ZCR) determines the count of sign change of the signal for a given period whereas the Fast Fourier Transform (FFT) is used for extracting the frequency domain features.

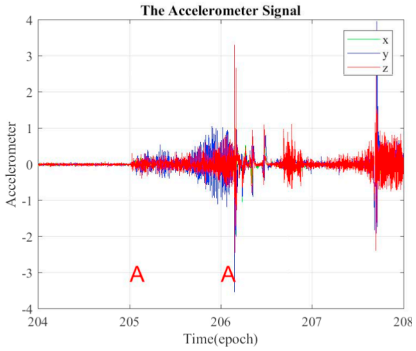
Spectral Spread, Spectral Centroid, Spectral Entropy, Spectral Flux and Spectral Rolloff measure different specifications of the acoustic spectrum of the noise signal. Mel Frequency Cepstral Coefficient



(a) On Heart Rate Signals



(b) On Noise Signals



(c) On Accelerometer Signals

Fig. 7. Abnormal respiratory event on raw signals.

Table 3
Extracted features.

Accelerometer		Noise		Heart Rate	
Magnitude	1	ZCR	1	Minimum	1
Minimum of X, Y, Z	3	Energy	1	Maximum	1
Maximum of X, Y, Z	3	Entropy of Energy	1	Mean	1
Mean of X, Y, Z	3	Spectral Centroid	1	Variance	1
Variance of X, Y, Z	3	Spectral Spread	1	ZCR	1
ZCR of X, Y, Z	3	Spectral Entropy	1	Energy	1
FFT of X, Y, Z	30	Spectral Flux	1	Hjort Mobility	1
		Spectral Rolloff	1	Hjort Complexity	1
		MFFCs	13	Skewness	1
		Harmonic Ratio	1	Kurtosis	1
		Pitch	1	RMS	1
		Chroma Vector	11	SVD	1
		Median Filter	1	FFT	10

(MFFC) converts the power spectrum into frequencies which captures important characteristic of the noise signal. While Harmonic Ratio is a representation of the signal's smoothness or dynamic stability, Pitch is defined as a major auditory attribute of musical tones. Chroma Vector represents the projection of the entire spectrum. Median Filter is used to remove the destructive noise while preserving the edges of the signal.

Hjort Mobility measures the proportion of standard deviation of the power spectrum of the heart rate signal and Hjort Complexity indicates the similarity between the signal and a pure sine wave. Skewness measures the lack of symmetry whereas Kurtosis determines whether the signal is heavy-tailed or light-tailed relative to the normal distribution.

4.5. Undersampling

The analysis of *"all-featured data"* implies that the imbalance between the counts of normal and abnormal respiratory events is obvious and this may cause poor classification performance. The elimination of some normal respiratory events is the first thing comes to mind. However, the most common problem of this strategy is defined as losing the distinctive data in the literature. In this study, this is not a problem since the number of normal respiratory events is sufficient.

To avoid prediction bias due to the uneven number of samples belonging to each class, the random undersampling strategy is utilized. The samples from the normal respiratory events class which has significantly more samples are randomly eliminated so that the two classes have roughly equal number of samples. The samples from the abnormal respiratory events do not need to be undersampled, since they are already fewer. The optimal degree of undersampling is determined via observing the classification accuracies, which tend to be highest for the case of perfectly balanced sample sizes.

The strategy is applied as follows. The count of abnormal respiratory events is very few so they are kept entirely. Among the normal respiratory events, some of them are chosen randomly and the decreased count is determined according to the classification results of different algorithms. The most accurate result is obtained with an equal number of normal and abnormal respiratory events.

4.6. Feature selection

The main objective of feature selection is understanding the importance of features when distinguishing the normal and abnormal respiratory events. Among different strategies found in the literature, Random Forest and Stepwise Feature Selection methods are preferred

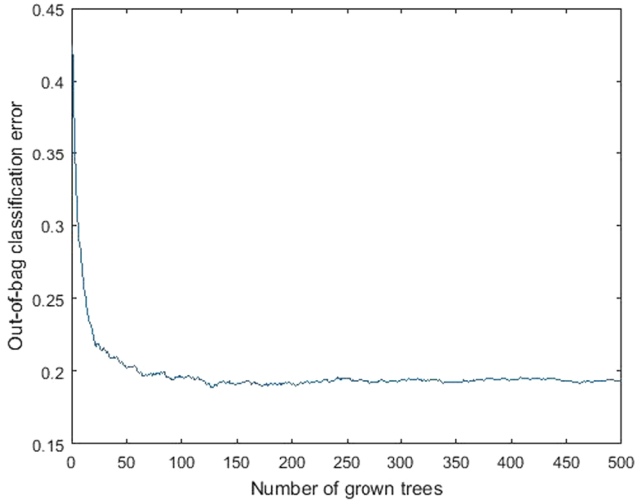


Fig. 8. Out-of-bag classification errors.

and the performance of these methods in this section is measured with the *"undersampled-featured data"*.

4.6.1. Random forest feature selection

By considering the definition of the RF, the first step is determining the number of decision trees. This determination is handled with the calculation of the out-of-bag classification errors which represents the misclassification error of the *"undersampled-featured data"*. The range of decision trees is from 1 to 500. Fig. 8 presents the convergence point of the misclassification error and 200 is chosen as the number of decision trees in the forest.

With the selected number of decision trees, the importance of features is estimated. The distance between the predicted probability for the true and wrong classes represents the importance. If the distance is larger, then the feature is more important since it differentiates the classes more precisely.

The importance estimations for all extracted features are given in Fig. 9(a). The first 46 features are derived from the accelerometer signals and the following 35 features are generated from the noise signals. The rest of the features are obtained from the heart rate signals. The analysis of the importance estimations indicates that most of the features have very low importance. Defining a threshold is a useful strategy for selecting the most important features. The threshold is chosen from the mean to the maximum of the importance estimations of features. The comparison of the misclassification errors generated with the *"selected-featured data"* proves that 0.8 as the selected threshold gives the best results.

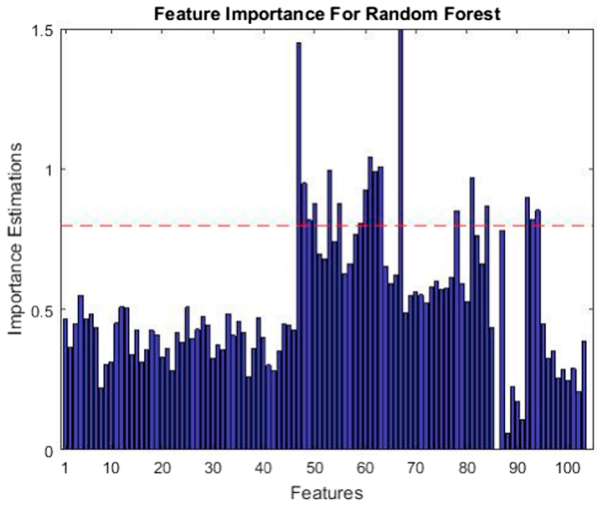
With the determined threshold, 18 features are selected and those are generated from the noise and heart rate signals. The selected features with their importance estimations can be seen in Fig. 9(b). The features extracted from the accelerometer signals are not chosen.

4.6.2. Stepwise feature selection

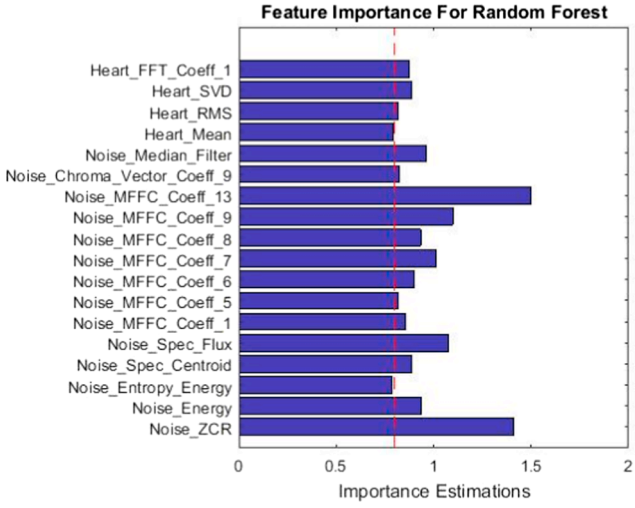
By considering the logic behind this algorithm, the forward and backward approaches are utilized together. Two different thresholds one for inclusion and another one for exclusion are determined. These thresholds mainly represent the range of p-values that measures the significance of features. The thresholds are chosen by considering the general tendency in the literature.

The algorithm is run for 10 times since each trial gives different subsets of all features. This is because the first included feature is determined randomly. The features which exist in at least 6 different subsets are chosen.

The average p-values for the selected features are given in Table 4 and the importance of features is decreased with the greater p-value.



(a) All Features



(b) Selected Features

Fig. 9. Importance estimations.

Table 4
Selected features with p-values.

Selected Features	p-Values	Selected Features	p-Values
Accel ZCR Y	0.0117	Noise MFFC Coeff 8	$6.2094 * 10^{-04}$
Accel FFT X Coeff 2	0.0103	Noise MFFC Coeff 9	$1.2864 * 10^{-10}$
Noise Energy	0.0114	Noise MFFC Coeff 10	$3.8164 * 10^{-04}$
Noise Entropy Energy	0.0029	Noise MFFC Coeff 13	$1.2363 * 10^{-06}$
Noise Spec Entropy	$3.0328 * 10^{-05}$	Noise Harmonic Ratio	0.0083
Noise Spec Flux	0.0100	Noise Chroma Vector Coeff 1	$3.2567 * 10^{-04}$
Noise MFFC Coeff 2	$6.8177 * 10^{-16}$	Noise Chroma Vector Coeff 4	0.0048
Noise MFFC Coeff 3	$7.1324 * 10^{-15}$	Noise Chroma Vector Coeff 5	$3.3861 * 10^{-07}$
Noise MFFC Coeff 4	$2.8075 * 10^{-28}$	Noise Chroma Vector Coeff 9	$1.8073 * 10^{-21}$
Noise MFFC Coeff 5	$2.3949 * 10^{-12}$	Noise Median Filter	$2.0167 * 10^{-09}$
Noise MFFC Coeff 6	$9.2606 * 10^{-19}$	Heart Hjort Complexity	0.0166

The features extracted from the noise signals have a significant majority among all selected features. This method determines that the features extracted from the heart rate signal are unimportant.

4.6.3. Discussion

The comparison of the selected features shows that some of them are common. Those features are extracted from the noise signals and the mutual selected features are *"Energy"*, *"Entropy Energy"*, *"Spectral Centroid"*, *"Spectral Flux"*, *"Median Filter"*, one coefficient of the *"Chroma Vector"* and some coefficients of the *"MFFC"*. These features mainly represent the snoring and coughing occurred after the abnormal respiratory event. The importance of features extracted from the heart rate signal and the accelerometer signal appear to be quite low. The main reason for this is that the recording of the noise signal with a smart phone is much more reliable than the recording of the heart rate and accelerometer signals with a smart watch. The position and tightness of the smart watch have a non-negligible effect on the interruptions in data collection. The wrist is an easily moving part of our body and the relative position of the sensors on the smart watch are continuously changing with respect to the skin of the subject. Since the watch uses photoplethysmogram (PPG) for measuring the heart rate, continuously moving sensor reduces the quality of the signal. Some motion compensation algorithms which could be added to the next generation smart watches may improve the signal quality in the future. Another reason can be that the body response of the abnormal respiratory event may not be clearly observable with wrist movements since the wrist is one of the most restless body parts. Also, the movements of the wrist can occur in both normal and abnormal respiratory event periods. Since the wrist is the natural place to carry the smart watch, we investigated the capabilities of the system while keeping it to be unobtrusive. In order to generate more accurate results for the importance of features, the extracted features can be altered and also the smart watch with a higher sampling rate can be chosen in the future.

5. Detection system performance and enhancements

The performance is measured with the success rate of distinguishing normal and abnormal respiratory events. The performance metrics used to measure this success rate are *"Accuracy"*, *"Precision"*, *"Sensitivity"* and *"Specificity"*. By considering the definitions of all performance metrics, the sensitivity which mainly represents the abnormality in respiratory events for this study becomes more important. Therefore, when analyzing the system performance, all performance metrics should be in balance and promising.

5.1. Detection system performance evaluation

In order to measure the system performance, different approaches are followed by considering the nature of data and also the literature. The results are generated with different machine learning algorithms with different subsets of the data coming from different sources. The generation of these subsets of data is explained in the related sections.

The classification results of raw signals are generated with the HMM. On the other hand, DT, KNN, RF and SVM algorithms generate the classification results with different subsets of the featured data. In the end, the results of binary classification algorithms with the *"undersampled-featured data"* are analyzed with Receiver Operating Characteristic (ROC) curves which are mainly designed to compare the performances of different algorithms.

5.1.1. Classification with raw signals

In this section, the main objective, distinguishing normal and abnormal respiratory events, is tried to be achieved with raw signals and the HMM approach since the abnormal respiratory events can be plainly observed from the raw signals and the HMM approach requires time series data.

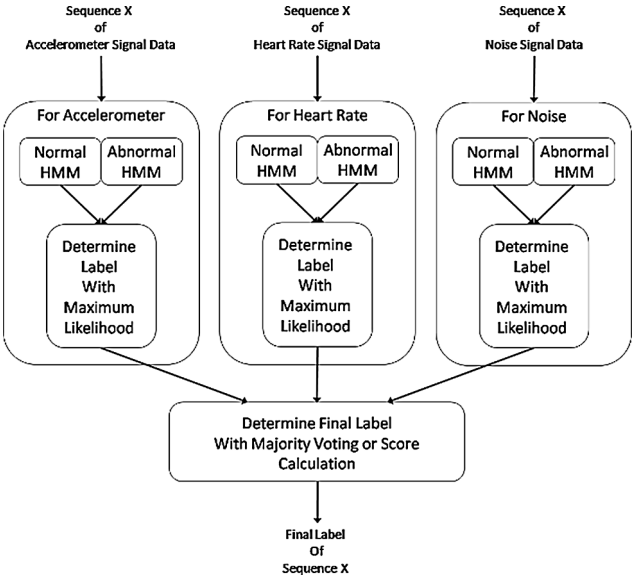


Fig. 10. Determination process of final label for the HMM.

In this approach, all subjects' data is combined and it is divided into sequences with length of 30 s. Different HMMs are generated for abnormal and normal respiratory event sequences. Furthermore, each source, sound level, heart rate monitor and accelerometer sensors have different models. As a result, totally six HMMs are obtained. The sequence of the same duration from each source is given to the related model. Fig. 10 represents this procedure clearly.

The determination of the final label of the sequence of the same time from each signal consists of two steps. In order to determine the intermediate label of a sequence, the maximum likelihood of a sequence is calculated. The intermediate labels are generated by two different approaches. The first one, the score calculation considers the score of each sequence calculated by each model and combines the scores without any generalization. The second one, the majority voting selects the alternative that has the majority. The score of each sequence is calculated for each model and the most voted one is selected. The majority voting approach is making some generalizations whereas the score calculation approach pays attention to details. The determination process of the final label is shown in Fig. 10.

Table 5 shows the example maximum likelihoods of a sequence. According to the Score Calculation approach, the sum of maximum likelihoods generates 1.3 (0.1 + 0.9 + 0.3) and 1.0 (0.3 + 0.2 + 0.5) for the Normal HMM and the Abnormal HMM, respectively. Since the Normal HMM score is greater than the Abnormal HMM, this sequence is labelled as Normal which is indicated with an asterisk in the table. By considering the Majority Voting approach, this sequence is labelled as Abnormal (0.3 > 0.1), Normal (0.9 > 0.2) and Abnormal (0.5 > 0.3) for Accelerometer, Heart Rate and Noise, respectively. This means that since Abnormal overcomes Normal, the final label is Abnormal which is indicated with an asterisk in the table. This example clearly states the underlying mechanisms of these two approaches.

Table 5
Example maximum likelihood value for final label determination. (Final label is Normal with Score Calc. and Abnormal with Majority as indicated with asterisks.)

Max. Likelihood	Accelerometer	Heart Rate	Noise	Summed Score
Normal HMM	0.1	0.9	0.3	1.3*
Abnormal HMM	0.3	0.2	0.5	1.0
	Abnormal*	Normal	Abnormal*	

5.1.2. Classification with featured data

After the feature extraction is completed, a combination of data that contains the subjects with abnormal and normal respiratory events is prepared. By using the PSG system, the true labels of this data are manually given by the specialist doctor as mentioned in Section 4.1.3.

The algorithms in this section are trained with either Percentage approach or K-Fold Cross-Validation. The first one, Percentage requires a part of data which is split into two parts and 75% of data are used for training, that contains 75% of the total abnormal and normal respiratory events. The testing is accomplished with the rest of the data.

The second one, K-Fold Cross-Validation is designed with 10 folds. Each fold preserves the ratio between abnormal and normal respiratory event counts. Therefore, each training or testing fold is a small representation of the data. Each time one fold is separated for testing and the remaining is used for training. This procedure is repeated for each fold and eventually, the predicted results are generated for the whole data.

For the results of the classification of the featured data is generated with the four different machine learning algorithms, DT, KNN, RF and SVM. For the implementation of KNN, different distances which represent the similarity of a sample to its neighbors are tried and the Cosine Distance is preferred since it generates more accurate results for a greater number of features. The neighbor count is selected as 26 for the K-Fold Cross-Validation and 33 for the Percentage approach. The different trials of neighbor counts show that the results converge around 30 neighbors which eventually yield the best results. For the RF, the determination process of the number of decision trees as 200 is explained in Section 4.6.1. For the SVM, the linear kernel is preferred since it clearly distinguishes the normal and abnormal respiratory events.

5.1.3. All results and discussion

The summary of the results for all implemented algorithms can be found in Table 6. The best result is indicated with an asterisk in the table. These results are chosen by considering the most accurate results for each algorithm. For the *"raw signals"*, the HMM results generated with Majority Voting is preferred. For the *"all-featured data"*, the results generated with the Percentage training approach is chosen for each algorithm. For the *"undersampled-featured"* and *"selected-featured data"*, the results of the K-Fold Cross-Validation approach is given for each algorithm. The general expectation is that the results of the K-Fold Cross-Validation outperform the results of the Percentage training approach for each case since the K-Fold Cross-Validation training approach considers the whole data however the imbalance between the counts of normal and abnormal respiratory events which causes the difference for *"all-featured data"*.

Table 6
Performances of all implemented algorithms.

Classification	Result (%)			
	Accuracy	Precision	Sensitivity	Specificity
HMM, Raw Signals	73.69	31.30	31.73	83.58
DT, All-Featured Data	74.82	34.33	36.11	83.87
KNN, All-Featured Data	81.42	52.50	19.78	95.82
RF, All-Featured Data	82.28	60.41	18.68	97.17
SVM, All-Featured Data	82.49	62.00	19.47	97.21
DT, Undersampled-Featured Data	72.06	72.56	70.95	73.17
KNN, Undersampled-Featured Data	73.57	69.91	82.74	69.34
RF, Undersampled-Featured Data	80.63	76.28	88.91	72.35
SVM, Undersampled-Featured Data	75.50	71.27	85.44	65.56
DT, Selected-Featured Data (RF)	73.70	73.86	73.37	74.03
KNN, Selected-Featured Data (RF)	70.81	69.89	73.14	68.49
RF, Selected-Featured Data (RF)*	80.96*	76.68*	88.99*	72.94*
SVM, Selected-Featured Data (RF)	73.33	72.50	75.17	71.50
DT, Selected-Featured Data (SW)	72.63	73.04	71.73	73.53
KNN, Selected-Featured Data (SW)	69.99	68.77	73.25	66.73
RF, Selected-Featured Data (SW)	80.30	76.01	88.56	72.04
SVM, Selected-Featured Data (SW)	75.61	71.15	86.18	65.05

The results prove that the main problem about the performance evaluation of the *"raw signals"* and *"all-featured data"* is the imbalance between the counts of normal and abnormal respiratory events. Even if the total sleeping duration time for the subjects who have abnormal respiratory events is very close to the total duration time for the subjects who have just normal events, the abnormal respiratory event durations are very minor compared to the normal respiratory event durations. The abnormal respiratory events do not occur all night long. Therefore the classification results, especially sensitivity, are affected negatively if the abnormal respiratory events are classified as normal respiratory events.

In order to solve this problem, the *"all-featured data"* is undersampled and this approach improves our system performance remarkably. More promising accuracy and sensitivity outcomes are obtained. If the system contains more abnormal respiratory events or more subjects, the undersampling rate of *"all-featured data"* is changed accordingly which may result in more improvement in the system performance.

Different feature selection algorithms are implemented to generate more accurate results. However, the features are also important for the system performance. This is because during the data collection, there are some interruptions because of the position and tightness of the smart watch. Therefore, the system needs every single feature to compensate data interruptions and to measure every single response of the body.

5.1.4. ROC curve analysis

In order to measure the strengths of the implemented algorithms, the ROC curves are generated. In each algorithm, the discrimination between the true and false classes is handled by setting a certain threshold. However, in the ROC curve, the true positive and false positive rates are generated with various thresholds and this approach makes this analysis useful and trustworthy when deciding which algorithm performs the best in every situation [28].

The ROC curves are created with the *"undersampled-featured data"* since it generates the best performance results. The results of the comparison of the performances of DT, KNN, RF and SVM algorithms are found in Fig. 11. The Area Under Curve (AUC) is calculated as 0.996, 0.850, 0.878 and 0.847 for the DT, KNN, RF and SVM algorithms, respectively. This analysis proves that even if the accuracies of the KNN, RF and SVM algorithms are more desirable, the DT algorithm provides the most reliable results for different thresholds. This is because the specificity and the sensitivity of the DT algorithm are balanced for each threshold.

5.2. Detection system performance enhancements

In order to increase the success rate of the respiratory event detection system, the problems of the design are investigated and this section explains these problems with the improvements.

5.2.1. Filtering raw signals

In the literature, the determination of a specific behavior, action or event in the raw signal is handled by using filters. The Discrete Wavelet Transform (DWT) is preferred with the HMM approach and there are

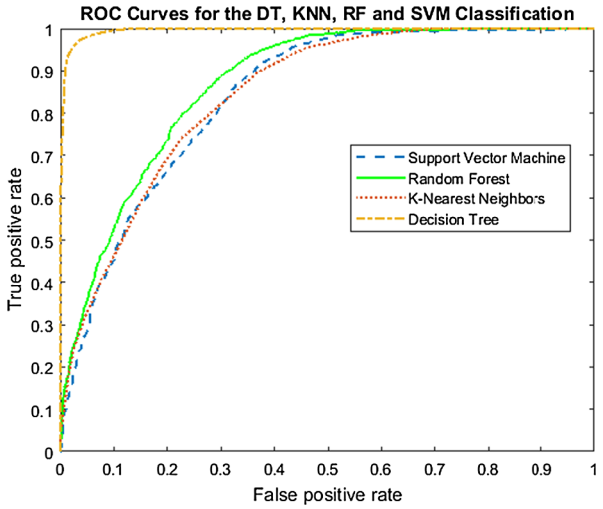
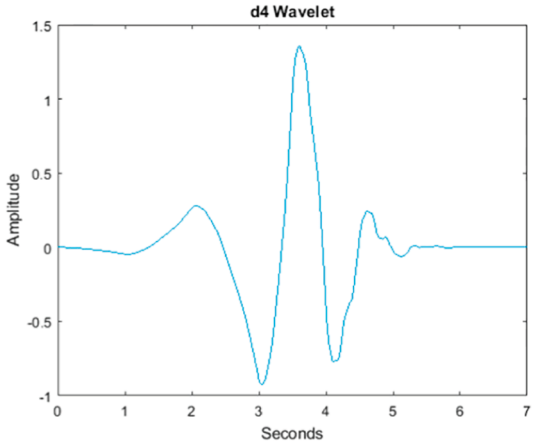
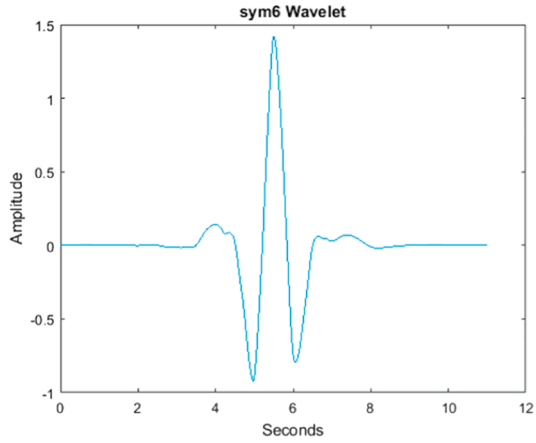


Fig. 11. ROC curves for the DT, KNN, RF and SVM.



(a) *Daubechies 4.*



(b) *Symlets 6.*

Fig. 12. Common mother wavelets.

several studies on the heart rate signal. With these two approaches, some researchers try to find the apnea and some try to do segmentation of the heart rate signal. The DWT mainly decomposes the signal into an orthogonal set of wavelets. This set of wavelets form a basis which is referred as the mother wavelet. The mother wavelets can be Gabor, Mexican Hut, Daubechies and Symlets. The researchers prefer Daubechies 4 (db4) and Symlets 6 (sym6) when extracting apnea from the heart rate signal with this method [29–35]. Fig. 12 presents those wavelets.

Table 7
Results for the HMM.

Classification	Result (%)			
	Accuracy	Precision	Sensitivity	Specificity
Raw Signals with Majority Voting	73.69	31.30	31.73*	83.58
Raw Signals with Score Calculation	73.13	31.51	25.85	86.75
Processed With db4, Score Calculation	77.17	36.58	26.78	89.05
Processed With sym6, Score Calculation	77.38*	27.70	28.48	88.91

In order to generate performance results, the HMM approach described in Section 5.1.1 is repeated for this section. The only difference is that the heart rate signals are preprocessed with *db4* and *sym6* wavelets. The results can be seen in Table 7 and the best results are indicated with an asterisk in the table. The results show that even if the accuracy is increased, the sensitivity is decreased. The filtering approach may lose the specific and most important parts of the raw signals and makes some generalizations.

5.2.2. Including pulse oximeter signals

The proposed system utilizes the sound level, heart rate monitor and accelerometer sensors. In the beginning, an external pulse oximeter is also considered however since the PSG system also contains a pulse oximeter, this idea is dropped by considering the obtrusiveness of the system design and comfort of subjects. For just evaluating the possible contribution of a pulse oximeter, the pulse oximeter data from the PSG system is extracted which eliminates the concerns about the reliability of the data. The pulse oximeter data are time series and therefore, this data is combined with the *"raw signals"* data that is used for the HMM classification.

If a pulse oximeter is included in the proposed system, the results in Table 8 could be obtained. The better result is indicated with an asterisk in the table. The results show that the accuracy is improved whereas the sensitivity is decreased. One reason for this outcome may be the difference between the data collection sources. Even if the PSG system is reliable, the smart watch data has some interruptions and this creates inconsistency between the *"raw signals"* data and the pulse oximeter data. Furthermore, because of the imbalance between the counts of normal and abnormal respiratory events, the significant improvement may not be observed.

5.2.3. Oversampling abnormal respiratory events

In the collected data, the abnormal respiratory events are very rare compared to the normal respiratory events. Since it is very difficult to collect more data, oversampling the abnormal respiratory events with Synthetic Minority Over-sampling Technique (SMOTE) library [36] is an option. This library is implemented for imbalanced datasets and it has several options like undersampling. Since one of the best system performance is obtained with the RF, the results are regenerated only for this classifier. The Percentage training approach is applied since

Table 8
Results for the HMM with majority voting.

Classification	Result (%)			
	Accuracy	Precision	Sensitivity	Specificity
Raw Signals	73.69	31.30	31.73	83.58
Included Pulse Oximeter Data	77.50*	36.64	24.61	89.96

Table 9
Results for the RF with oversampling approach.

Classification	Result (%)			
	Accuracy	Precision	Sensitivity	Specificity
All-Featured Data	82.28	60.41	18.68	97.17
Oversampled-Featured Data	81.78	53.71	32.51*	93.39

"all-featured data" yields the best result with it as shown in Table 9. The better result is indicated with an asterisk in the table. For just testing purposes, we do not equalize the counts of abnormal and normal respiratory events in the algorithm since oversampling is actually creating non-real data and the reliability of this data is very low. The table proves that even if we oversample the abnormal respiratory events slightly, the system performance is improved. Making remarkable oversampling results in more accurate classification results.

5.2.4. Generating an extended label set

The labels provided by the PSG system and the sleep specialist have limitations. This has two reasons. The first one is explained in Section 4.3. Since the decodable file includes only the duration information not the starting moment of an abnormal respiratory event, the label of an epoch (30 s period) can be incorrect. The second reason is explained in Fig. 6, the body can give a physiological response after the occurrence of an abnormal respiratory event.

In order to cover both cases, a new label set by considering the duration time of the abnormal respiratory event is created. Since the starting moment of an abnormal respiratory event is unknown, the starting moment is assumed to be uniformly distributed for each epoch (30 s period). This means that, at average, every abnormal respiratory event starts in the middle of that epoch. Therefore, if the duration time of an abnormal respiratory event is greater than 15 s, then both the current and the next epoch are labeled as an abnormal respiratory event epoch. This new label set is called as the *"Extended Label Set"*.

With this new label set, all implemented algorithms are reevaluated and the implementations that yield the best results are chosen as explained in Section 5.1.3. The classification results are shown in Table 10 and the best result is indicated with an asterisk in the table. The results show that for the *"undersampled-featured data"*, the system performance is improved significantly and the accuracy obtained with the RF algorithm as 85.95% is the best accuracy achieved in this study. If the label set is more accurate and more abnormal respiratory events can be collected, the system distinguishes the abnormal respiratory events better.

Table 10
Results for extended labels.

Classification	Result (%)			
	Accuracy	Precision	Sensitivity	Specificity
HMM, Raw Signals	71.90	52.95	55.76	78.80
DT, All-Featured Data	72.01	52.99	55.52	79.03
KNN, All-Featured Data	73.05	56.75	41.00	86.69
RF, All-Featured Data	77.77	66.58	51.34	89.03
SVM, All-Featured Data	78.10	66.79	53.03	88.77
DT, Undersampled-Featured Data	77.96	78.31	77.34	78.58
KNN, Undersampled-Featured Data	78.28	75.18	84.44	72.13
RF, Undersampled-Featured Data	85.95*	81.76*	92.55*	79.35*
SVM, Undersampled-Featured Data	79.68	74.88	89.35	70.02

6. Prescreening system performance evaluation

Until this section, the objective is distinguishing the normal and abnormal respiratory events. However, the main objective of the prescreening system is to recognize people with severe respiratory events and give suggestions like consulting a doctor. Therefore, in this section, each subject's situation is analyzed separately.

The determination process of subject classification as *"Normal"* and *"Abnormal"* is explained in Section 4.1.4. By following the same strategy, a new AHI score which is predicted by several algorithms is generated for the subjects. This predicted score is an indicator of the situation of subjects evaluated by the proposed system.

Each subject's data is analyzed with different algorithms, namely the DT, KNN, RF and SVM algorithms. The objective of each algorithm is counting the abnormal respiratory events for each subject. The count of abnormal respiratory events is divided into the total duration of collected data for that subject and the predicted AHI score is obtained as a result. The selected subject's data is preserved for testing and the rest of the collected data coming from other subjects is used for the training of the algorithm. This procedure is repeated for each subject.

The predicted labels are generated for each algorithm and a subject who is classified as *"Abnormal"* by at least one algorithm is labeled as *"Abnormal"*. Since this system is designed for recognizing people with severe respiratory problems, this cautious approach is preferred. The results of this procedure can be seen in Table 11. In the table, the subjects presented with *"1"* are classified as abnormal whereas *"0"* indicates that the subject is classified as normal.

Table 11 proves that only one subject whose id is 326 is classified wrongly. The implemented algorithms cannot detect the abnormality in this subject. One reason for this misclassification may be the gender. This subject is the one and only female subject and the body of the female subjects may not give distinguishing or sharp responses as the body of male subjects does. Other reason can be explained with the true AHI score. Since this subject's true AHI score is not as high as others, it is classified as *"Normal"*. The table also indicates the performance of the system as a prescreening tool. The DT gives the same output with the combination of all algorithms, which reveals that eliminating other algorithms is possible. If more subjects participate in the experiments and the system is trained with more abnormal respiratory events, every subject may be classified correctly. By considering the obtained results, one can say that this system which achieves the objectives, unobtrusiveness, reduced cost, objectiveness, protection of privacy and user-friendliness can be used as a prescreening tool.

7. Conclusion and future work

This study proposes a prescreening tool that recognizes the problems in respiration during sleep, analyzes the health condition of subjects and gives suggestions like consulting a doctor. This study provides some contributions to the field. The most unobtrusive, cheap and user-friendly system design is formed by utilizing smart wearables. The smart wearables solve the problems experienced because of the overwhelming design of other monitoring tools and eliminate the obligation to obey the set of rules about the sleeping place, sleeping position. Furthermore, there is no need for a professional which requires a huge amount of time and cost. This system is also designed to protect the privacy by collecting raw signals and does not permit outside intervention which provides objectiveness. The real conditions of subjects can be monitored by doctors without the concerns about the manipulation or imitation of behavior.

In general sleep studies have not a common metric for measuring their performances. However, the most similar studies conducted are in [13,14] as mentioned before and both of them utilize the smart phone with other external devices instead of a smart watch. Furthermore, both measure their performances with the same metric. In the first one, the apnea-healthy classification accuracy is 92.2%. However, its design is

Table 11
True and predicted labels for each subject.

Labels	311	313	316	323	325	326	327	328	331	333	334	337	338	341	344	349	353
True	0	0	1	1	0	1*	1	1	1	0	0	0	1	1	0	0	0
Predicted-All	0	0	1	1	0	0*	1	1	1	0	0	0	1	1	0	0	0
Predicted-DT	0	0	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0
Predicted-KNN	0	0	0	1	0	0	1	1	1	0	0	0	1	1	0	0	0
Predicted-RF	0	0	0	1	0	0	1	1	1	0	0	0	1	1	0	0	0
Predicted-SVM	0	0	0	1	0	0	1	1	1	0	0	0	1	0	0	0	0

different from this study since it utilizes a medical database and a physician generates the ground truth. In other words, the medical information has to be in the database before the analysis. Moreover, most of the time these types of databases are not refreshed very often. On the other hand, in [14], the apnea detection accuracy is obtained as 87.5% which is measured with a very similar system design. However, in order to measure the performance, this study implements some heuristic methods unlike the proposed system. Compared to those studies, the proposed system brings in novelties since it is much less obtrusive and cost efficient due to using the smart watch and due to utilizing systematic machine learning techniques.

This proposed prescreening tool is a promising alternative. The combination of all steps, the data collection with real subjects, the ground truth validation by the PSG system and also the sleep specialist, the methodology used in the data processing, feature extraction, feature selection, sampling and machine learning steps, the results obtained from this methodology and the investigation of the drawbacks with the suggested enhancements, makes this study a novel solution for both of the objectives, *"Distinguishing abnormal respiratory events from normal respiratory events"* and *"Classifying subjects according to the problems in their respiration"*. As a result, the best-achieved accuracy rate for differentiating the normal and abnormal respiratory events is 85.95%. Furthermore, the prescreening tool identifies the severity of problems in respiration with only one misclassification among 17 subjects. Both results are acceptable compared to the success rates found in the literature.

Although this prescreening tool is a promising alternative, there is still room for improvement. The future works should focus on the enhancements proposed in this study. The smart watch is the most unobtrusive option for the heart rate monitoring in today's technology and the weaknesses, the limited battery life and the data interruption caused by the position and tightness of the smart watch will be solved with the help of the developing technology. On the other hand, there is common knowledge about every system. In order to increase the reliability of a system, the system should be tested with more complex and diverse data which collected by considering the balance between the abnormal and normal respiratory events.

Declaration of Competing Interest

Authors declared that there is no conflict of interest.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] WHO, Technical Meeting on Sleep and Health, accessed at July 2017. URL http://www.who.int/en.
[2] A.A. Pona, The Impact of Sleep Quality on Energy Intake, Eating Behavior, and Physical Activity, University of Missouri-Kansas City, 2015.
[3] M. Bano, F. Chiaromanni, M. Corrias, M. Turco, M. De Rui, P. Amodio, C. Merkel, A. Gatta, G. Mazzotta, R. Costa, et al., The influence of environmental factors on sleep quality in hospitalized medical patients, Front. Neurol. 5 (2014).
[4] NHLBI, Why Is Sleep Important?, accessed at July 2017. URL http://www.nhlbi.nih.gov.
[5] H. Miwa, S.-I. Sasahara, T. Matsui, Roll-over detection and sleep quality measurement using a wearable sensor, in: Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE, 2007, pp. 1507–1510.
[6] NCSDR-NHTSA, Drowsy Driving and Automobile Crashes, accessed at July 2017. URL http://www.nhtsa.gov/people/injury/drowsy_driving1/Drowsy.html.
[7] Stanford-Medicine, Sleep Disorders, accessed at July 2017. URL http://sleep.stanford.edu/sleep-disorders.
[8] NationalSleepFoundation, Sleep Disorders, accessed at July 2017. URL https://sleepfoundation.org/sleep-disorders-problems.
[9] S. Alqassim, M. Ganesh, S. Khoja, M. Zaidi, F. Aloul, A. Sagahyroon, Sleep apnea monitoring using mobile phones, in: 2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom), 2012, pp. 443–446.
[10] H. Nakano, K. Hirayama, Y. Sadamitsu, A. Toshimitsu, H. Fujita, S. Shin, T. Tanigawa, Monitoring sound to quantify snoring and sleep apnea severity using a smartphone: proof of concept, J. Clin. Sleep Med.: JCSM: Off. Publ. Am. Acad. Sleep Med. 10 (1) (2014) 73.
[11] X. Sun, Z. Lu, W. Hu, G. Cao, Symdetector: detecting sound-related respiratory symptoms using smartphones, in, Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2015, pp. 97–108.
[12] M.-H. Tseng, H.-C. Hsu, C.-C. Chang, H. Ting, H.-C. Wu, P.-H. Tang, Development of an intelligent app for obstructive sleep apnea prediction on android smartphone using data mining approach, 2012 9th International Conference on Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC), IEEE, 2012, pp. 774–779.
[13] J. Behar, A. Roebuck, M. Shahid, J. Daly, A. Hallack, N. Palmius, J. Stradling, G.D. Clifford, Sleepap: An automated obstructive sleep apnoea screening application for smartphones, IEEE J. Biomed. Health Informat. 19 (1) (2015) 325–331.
[14] M. Al-Mardini, F. Aloul, A. Sagahyroon, L. Al-Husseini, Classifying obstructive sleep apnea using smartphones, J. Biomed. Informat. 52 (2014) 251–259.
[15] T.-Y. Han, S.-D. Min, Y. Nam, A real-time sleep monitoring system with a smartphone, 2015 9th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), IEEE, 2015, pp. 458–461.
[16] C. Bernardeschi, M.G. Cimino, A. Domenici, G. Vaglini, Using smartwatch sensors to support the acquisition of sleep quality data for supervised machine learning, International Conference on Wireless Mobile Communication and Healthcare, Springer, 2016, pp. 251–259.
[17] X. Sun, L. Qiu, Y. Wu, Y. Tang, G. Cao, Sleepmonitor: Monitoring respiratory rate and body position during sleep using smartwatch, Proc. ACM Interact., Mobile, Wearable Ubiquit. Technol. 1 (3) (2017) 104.
[18] NHLBI, What Is Sleep Apnea?, accessed at July 2017. https://www.nhlbi.nih.gov/health/health-topics/topics/sleepapnea/.
[19] Y. Kawano, A. Tamura, T. Watanabe, J. Kadota, Influence of the severity of obstructive sleep apnea on heart rate, J. Cardiol. 56 (1) (2010) 27–34.
[20] B. Camcı, A.Y. Kahveci, B. Arnrich, C. Ersoy, Sleep apnea detection via smart phones, Signal Processing and Communications Applications Conference (SIU), 2017 25th, IEEE, 2017, pp. 1–4.
[21] P.A. Deutsch, M.S. Simmons, J.M. Wallace, Cost-effectiveness of split-night polysomnography and home studies in the evaluation of obstructive sleep apnea syndrome, J. Clin. Sleep Med. 2 (2) (2006) 145–153.
[22] R.B. Berry, R. Budhiraja, D.J. Gottlieb, D. Gozal, C. Iber, V.K. Kapur, C.L. Marcus, R. Mehra, S. Parthasarathy, S.F. Quan, et al., Rules for scoring respiratory events in sleep: update of the 2007 aasm manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the american academy of sleep medicine, J. Clin. Sleep Med.: JCSM: Off. Publ. Am. Acad. Sleep Med. 8 (5) (2012) 597.
[23] N. Wolkove, M. Baltzan, H. Kamel, R. Dabrusin, M. Palayew, Long-term compliance with continuous positive airway pressure in patients with obstructive sleep apnea, Can. Respirat. J. 15 (7) (2008) 365–369.
[24] SamsungInc., Samsung Galaxy Gear S3, accessed at July 2017. http://www.samsung.com/global/galaxy/gear-s3.
[25] SmartmobDevelopment, Smart Voice Recorder, accessed at July 2017. http://recorder.smartmobdev.com/.
[26] CareFusion, SomnoStar, accessed at July 2017. http://www.carefusion.com/our-products/respiratory-care/sleep-diagnostics-and-therapy/somnostar-z4.
[27] D. Alvarez-Estevez, European Data Format, accessed at July 2017. http://www.edfplus.info/.
[28] L. Gonçalves, A. Subtil, M.R. Oliveira, P. Bermudez, Roc curve estimation: An

overview, REVSTAT–Statist. J. 12 (1) (2014) 1–20.
[29] V.P. Rachim, G. Li, W.-Y. Chung, Sleep apnea classification using ecg-signal wavelet-pca features, Bio-medical Mater. Eng. 24 (6) (2014) 2875–2882.
[30] U. Pale, F. Thürk, E. Kaniusas, Heart rate variability analysis using different wavelet transformations, 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, 2016, pp. 1649–1654.
[31] G. Kheder, A. Kachouri, R. Taleb, M. Ben Messaoud, M. Samet, Feature extraction by wavelet transforms to analyze the heart rate variability during two meditation techniques, Advances in Numerical Methods, Springer, 2009, pp. 379–387.
[32] P. Várady, T. Micsik, S. Benedek, Z. Benyó, A novel method for the detection of apnea and hypopnea events in respiration signals, IEEE Trans. Biomed. Eng. 49 (9) (2002) 936–942.
[33] O. Fukuda, Y. Nagata, K. Homma, T. Tsuji, Evaluation of heart rate variability by using wavelet transform and a recurrent neural network, Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE, vol. 2, IEEE, 2001, pp. 1769–1772.
[34] R.V. Andreao, B. Dorizzi, J. Boudy, Ecg signal analysis through hidden markov models, IEEE Trans. Biomed. Eng. 53 (8) (2006) 1541–1549.
[35] S.M. Isa, M.I. Fanany, W. Jatmiko, A.M. Arymurthy, Sleep apnea detection from ecg signal: analysis on optimal features, principal components, and nonlinearity;; (iCBBE) 2011 5th International Conference on Bioinformatics and Biomedical Engineering, IEEE, 2011, pp. 1–4.
[36] G. Lemaı˘tre, F. Nogueira, C.K. Aridas, Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Machine Learn. Res. 18 (17) (2017) 1–5.