ABNORMAL RESPIRATORY EVENT DETECTION IN SLEEP: PRESCREENING SYSTEM WITH SMART WEARABLES

by

Burçin Camcı

B.S., Computer Engineering, Boğaziçi University, 2015

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2017

ABNORMAL RESPIRATORY EVENT DETECTION IN SLEEP:
PRESCREENING SYSTEM WITH SMART WEARABLES

APPROVED BY:

Prof. Cem Ersoy .....
(Thesis Supervisor)

Assoc. Prof. Özlem Durmaz İncel .....

Assoc. Prof. Arzucan Özgür .....

DATE OF APPROVAL: 19.12.2017

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitudes to my thesis supervisor, Prof. Cem Ersoy. I always appreciate his continuous assistance, invaluable efforts and understanding in my academic life and it has been a great honour for me to work with him during this process. His wisdom and vision will lead me in my future life.

I would like to give my thanks to my jury members Özlem Durmaz İncel and Arzucan Özgür for their insightful comments and suggestions.

I sincerely thank to Prof. Dr. Hakan Kaynak for his voluntary guidance throughout my thesis. I feel truly privileged to study with him. I would like to give my deepest thanks to Gülüstü Salur for her support and assistance. I have been amazingly fortunate to meet her. I appreciate the help and valuable feedbacks of Bert Arnrich throughout my research.

I deeply thank to my friends at NETLAB and CmpE who have been generous with their help, inspiration and encouragement. Especially, I would like to thank Can Tunca, Bilgin Koşucu and Ali Yavuz Kahveci for their support and advices. I also would like to thank my colleagues from my work. I am so grateful for all we shared together.

I have been fortunate enough to have wonderful friends. I would like to thank my dearest friend Nezihe Pehlivan for encouraging talks and support. I always feel her friendship. I would like to thank Buğrahan Memiş for his help. He always provided me with motivation and support. I would like to thank Aybüke Taşdirek-Utku Simitli, Burcu Gökgöz, Burcu Güler, Ebru-Serdar Kuzucu and Fulya Başağaç-Utku Can. I enjoyed every moment I shared with them and being a member of their houses was amazing!

I was so lucky to have lovely roommates. It is not possible to mention every support they provided here, but I would like to thank Burcu Özden Yücel and Selin Erkan for letting me share all the happiness and stress of my graduate years. Sharing the same house has always been fantastic and will definitely be missed!

I would like to express my gratitude to my beloved family. I would like to thank my sister, my brother and their spouses for the support they provided through my entire life. I would like to thank my mum and dad for being next to me whenever I needed. Without their endless love, unconditional support and amazing generosity throughout my life, I could not be the person I am right now. I would like to dedicate my thesis to my mum and dad.

ABSTRACT

ABNORMAL RESPIRATORY EVENT DETECTION IN SLEEP: PRESCREENING SYSTEM WITH SMART WEARABLES

Sleeping is an important activity to monitor since it has a crucial role in the overall health and well-being of the people and the society. The problems in sleep affect the daily lives of people negatively and a great deal of diseases has a strong correlation with low sleep quality. In order to diagnose the problems in the sleep, different monitoring systems are developed in the literature. The unobtrusiveness, reduced cost and reachability are the main design considerations and in this thesis, those are accomplished with the smart wearables; smart watch and smart phone. The proposed system can be utilized as a prescreening tool which recognizes the severity of problems in respiration during sleep. For this purpose, the accelerometer and heart rate monitor sensors on smart watch and the sound level sensor on smart phone are activated. The experiments of this system are performed with 17 subjects in a sleep clinic. The subjects are examined by the specialist doctor and among those subjects; nine of them perform a few abnormal respiratory events whereas lots of abnormal respiratory events are observed for remaining eight subjects. The data collected from these subjects is merged and 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 with the various combinations of data, training and scoring strategies. The system performance is measured in two ways; discrimination success of abnormal respiratory events and classification success of subjects according to the problems in their respiration. The best achieved accuracy rate of distinguishing abnormal respiratory events is 85.95% whereas the classification success of subjects is one misclassification through 17 subjects.

ÖZET

UYKUDA ANORMAL SOLUNUM OLAYI TESPİTİ: AKILLI GİYİLEBİLİR CİHAZLARIN KULLANILDIĞI ÖNİZLEME SİSTEMİ

İnsan ve toplumun sağlık ve huzurunda elzem bir rol oynayan uyku, gözlemlenmesi gereken önemli bir aktivitedir. Uyku süresince yaşanan problemler kişilerin günlük yaşamını olumsuz etkilemekte olup düşük kalitedeki uyku pek çok hastalığa sebep olmaktadır. Uykudaki problemleri teşhis edebilmek için literatürde bir çok gözlemleme sistemi geliştirilmiştir. Kişiyi rahatsız etmeyen, düşük maliyetli ve herkes tarafından ulaşılabilir bir sistem tasarlanmasına dikkat edilmelidir ve bu tezde, bu özellikler akıllı giyilebilir saat ve akıllı telefon yardımıyla sağlanmıştır. Önerilen sistem bir önizleme sistemidir ve kişilerin uykudayken yaşadıkları solunum problemlerinin ciddiyetini ölçmektedir. Bu amaçla, akıllı saatteki ivme ve kalp atış ölçer alıcılarıyla akıllı telefondaki ses ölçüm alıcısı aktifleştirilmiştir. Tasarlanan sistemin deneyleri bir uyku kliniğinde uzman uyku doktoru tarafından muayene edilen 17 denek üzerinde yapılmıştır. Dokuz denekte az sayıda anormal solunum olayı görülürken geriye kalanlarda pek çok anormal solunum olayı saptanmıştır. Deneklerden toplanan veriler birleştirildikten sonra, bu verilerin çeşitli kombinasyonları; farklı öznitelik çıkarım, seçim ve örnekleme yaklaşımlarıyla oluşturulmuştur. Sınıflandırma sonuçları verilerin çeşitli kombinasyonları, farklı eğitim ve puanlama yaklaşımları yardımıyla beş makine öğrenimi algoritması tarafından çıkarılmıştır. Sistem performansı iki yolla ölçülmektedir; anormal solunum olaylarının ayrımsanma başarısı ve deneklerin solunum problemlerine göre sınıflandırılma başarısı. Anormal solunum olaylarının ayrımsanma başarısı için elde edilen en iyi doğruluk oranı %85,95 iken, deneklerin sınıflandırılma başarısı; 17 kişide bir yanlış sınıflandırmadır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iii
ABSTRACT v
ÖZET vi
LIST OF FIGURES ix
LIST OF TABLES xi
LIST OF SYMBOLS xiv
LIST OF ACRONYMS/ABBREVIATIONS xv
1. INTRODUCTION 1
2. LITERATURE SURVEY 9
2.1. Related Sleep Studies 9
2.2. Background Concepts of Proposed System 22
2.2.1. Respiratory Event 22
2.2.2. Smart Watch 23
2.2.3. Smart Phone 24
2.2.4. Polysomnography 24
3. PRESCREEENING SYSTEM ARCHITECTURE 27
3.1. Detection System Design 29
3.2. Data Collection with Proposed System 31
3.2.1. Smart Watch 31
3.2.2. Smart Phone 32
3.2.3. Polysomnography 33
3.2.4. Experimental Setup and Subject Information 35
3.3. Abnormal Respiratory Event Visualization 37
3.4. Data Processing Challenges 39
3.5. Feature Extraction 40
3.6. Undersampling 42
3.7. Feature Selection 43
3.7.1. Random Forest Feature Selection 43
3.7.2. Stepwise Feature Selection 46

3.7.3. Discussion 48
4. SYSTEM PERFORMANCE AND PERFORMANCE ENHANCEMENTS . 49
4.1. Respiratory Event Detection System Performance Evaluation 49
4.1.1. Classification Results with Raw Signals 50
4.1.2. Classification Results with All-Featured Data 52
4.1.2.1. Decision Tree 53
4.1.2.2. K-Nearest Neighbors 54
4.1.2.3. Random Forest 55
4.1.2.4. Support Vector Machine 56
4.1.2.5. Discussion 57
4.1.3. Classification Results with Undersampled-Featured Data 57
4.1.4. Classification with Selected-Featured Data 58
4.1.5. ROC Curve Analysis with Undersampled-Featured Data 60
4.1.6. Discussion 61
4.2. Respiratory Event Detection System Performance Enhancements 63
4.2.1. Filtering Raw Signals 64
4.2.2. Including Pulse Oximeter Signals 66
4.2.3. Oversampling Abnormal Respiratory Events 67
4.2.4. Generating Extended Label Set 68
4.2.5. Discussion 69
4.3. Prescreening System Performance Evaluation 70
5. CONCLUSION AND FUTURE WORK 72
REFERENCES 75
APPENDIX A: PERFORMANCE COMPARISONS 87

LIST OF FIGURES

Figure 2.1. Human Body with the Polysomnography Electrodes Attached [1]. 25

Figure 3.1. Prescreening System Design. 28

Figure 3.2. Respiratory Event Detection System Description. 30

Figure 3.3. Smart Watch on Wrist. 32

Figure 3.4. Interface of the PSG System. 34

Figure 3.5. A Subject Sleeping with the PSG System [2]. 35

Figure 3.6. Arousal of an Abnormal Respiratory Event. 38

Figure 3.7. Abnormal Respiratory Event on Raw Signals 39

Figure 3.8. Out-of-Bag Classification Errors with the Undersampled-Featured Data for the Random Forest Algorithm. 44

Figure 3.9. Importance Estimations of All Features for the Random Forest Algorithm. 45

Figure 3.10. Out-of-Bag Classification Errors with Selected-Featured Data for the Random Forest Algorithm. 45

Figure 3.11. Selected Features with Importance Estimations. 46

Figure 4.1. Determination Process of Final Label for the Hidden Markov Model Algorithm. 51

Figure 4.2. ROC Curves for the DT, KNN, RF and SVM Algorithms. 60

Figure 4.3. Common Mother Wavelets [3] 64

LIST OF TABLES

Table 3.1. Information about Electrodes used in the Polysomnography System 34

Table 3.2. Information about Subjects 37

Table 3.3. Extracted Features 41

Table 3.4. Selected Features with Stepwise Feature Selection and Their p-Values 48

Table 4.1. Classification Results of the Hidden Markov Model Algorithm with Score Calculation 52

Table 4.2. Classification Results of the Hidden Markov Model Algorithm with Majority Voting 52

Table 4.3. Classification Results of the Decision Tree Algorithm with K-Fold Cross Validation 53

Table 4.4. Classification Results of the Decision Tree Algorithm with Percentage 54

Table 4.5. Classification Results of the K-Nearest Neighbors Algorithm with K-Fold Cross Validation 54

Table 4.6. Classification Results of the K-Nearest Neighbors Algorithm with Percentage 55

Table 4.7. Classification Results of the Random Forest Algorithm with K-Fold Cross Validation 55

Table 4.8. Classification Results of the Random Forest Algorithm with Percentage 56

Table 4.9. Classification Results of the Support Vector Machine Algorithm with K-Fold Cross Validation 56

Table 4.10. Classification Results of the Support Vector Machine Algorithm with Percentage 57

Table 4.11. Performance Comparison between the All-Featured Data and the Undersampled-Featured Data 58

Table 4.12. Performance Comparison between the Undersampled-Featured Data, the Random Forest Algorithm Selected-Featured Data and the Stepwise Algorithm Selected-Featured Data 59

Table 4.13. Performance Comparison of All Implemented Algorithms 63

Table 4.14. Performance Comparison between Raw And Processed Heart Rate Signal for the Hidden Markov Model Algorithm with Score Calculation 65

Table 4.15. Performance Comparison between Raw And Processed Heart Rate Signal for the Hidden Markov Model Algorithm with Majority Voting 65

Table 4.16. Performance Comparison between Current System Design and Included Pulse Oximeter Signal for the Hidden Markov Model with Majority Voting 66

Table 4.17. Performance Comparison between the All-Featured Data and the Oversampled-Featured Data for the Random Forest Algorithm with Percentage 67

Table 4.18. Performance Comparison between Original Labels and Extended Labels 69

Table 4.19. True and Predicted Labels for Each Subject 71

Table A.1. Classification Results with Undersampled-Featured Data with K-Fold Cross Validation 87

Table A.2. Classification Results with Undersampled-Featured Data with Percentage 88

Table A.3. Classification Results with Selected-Featured (Random Forest) Data with K-Fold Cross Validation 89

Table A.4. Classification Results with Selected-Featured (Random Forest) Data with Percentage 90

Table A.5. Classification Results with Selected-Featured (Stepwise) Data with K-Fold Cross Validation 91

Table A.6. Classification Results with Selected-Featured (Stepwise) Data with Percentage 92

LIST OF SYMBOLS

A	Abnormal respiratory event

LIST OF ACRONYMS/ABBREVIATIONS

1R	1R Classifier
AASM	American Academy of Sleep Medicine
AHI	Apnea-Hypopnea Index
BN	Bayesian Network
CVD	Cardiovascular Disease
CSS	Cascading Style Sheets
CPU	Central Processing Unit
CRF	Conditional Random Field
CPAP	Continuous Positive Airway Pressure
db4	Wavelet of Daubechies 4
DT	Decision Tree
DWT	Discrete Wavelet Transform
ECG	Electrocardiogram
EEG	Electroencephalographic
EMG	Electromyogram
EOG	Electrooculogram
ESS	Epworth Sleepiness Scale
EDF	European Data Format
FFT	Fast Fourier Transform
GPS	Global Positioning System
HRM	Heart Rate Monitor
HMM	Hidden Markov Model
HTML	Hypertext Markup Language
KNN	K-Nearest Neighbors
LR	Logistic Regression
MFFC	Mel Frequency Cepstral Coefficient
OSA	Obstructive Sleep Apnea
PPG	Photoplethysmogram

PSQI	Pittsburgh Sleep Quality Index
PSG	Polysomnography
RF	Random Forest
RMS	Root Mean Square
ROC	Receiver Operating Characteristic
SP	Smart Phone
SW	Smart Watch
SVD	Singular Value Decomposition
SMOTE	Synthetic Minority Over-sampling Technique
sym6	Wavelet of Symlets 6
SVM	Support Vector Machines
ZCR	Zero Crossing Rate

1. INTRODUCTION

In an ordinary day, every person performs many activities like eating, walking, speaking, sleeping. Among those activities, sleeping is undoubtedly an important activity to monitor since it occupies about one-third of people's life time [4]. Beside many other factors, the quality of life depends on the quality of sleep and sleep with good quality has a crucial role in the overall health and well-being of the people and the society.

The quality of sleep is linked with various environmental and personal factors. The environmental factors are more predictable and easier to provide. The light and noise level are the most important environmental factors since falling asleep in a dark and silent room is much easier. Also, the temperature of the room and comfort of the bed have a major impact on the quality of sleep. The personal factors include individual properties like age and weight. Also, the life style, working conditions and habits of people have a non-negligible effect on the sleep quality. Most of the people who have a shift work or travel across time zones have severe sleep problems. Eating and drinking habits and also usage of medications and other substances like alcohol, cigarette can change the sleep behavior. Beside those factors, people who have medical problems like cancer, Parkinson's disease, depression, obesity or have pain, anxiety are suffering from poor sleep quality. Poor sleep quality is generally caused by not only one factor but interactions of multiple factors [5–9].

The symptoms emerging with poor sleep quality reveal themselves in each person's sleep in different ways. Some people cannot sleep for long hours, some feel tired all the time and some have apneas which can be defined as pauses in breathing or shallow breathing during sleep. Therefore, measuring the quality of sleep requires complicated investigation of multiple factors and the quality of sleep depends on various criteria. There has been a wide-scale research on the issue of which criteria has the greatest association with the quality of sleep [10]. These studies consider different aspects of the underlying reasons and analyze the interaction between the potential causes of

poor sleep quality. Sleeping time, sleep duration, snoring, apnea count, nightmares or night terrors, leg movements during night, daytime sleepiness are the common symptoms that need intervention and analyzed as sleep statistics in order to determine the quality of sleep [11].

As a result of several symptoms, lots of people are suffering from different sleep disorders and in the field, most common sleep disorders are stated as insomnia, sleep apnea, restless leg syndrome, narcolepsy, snoring, sleep walking and terrors [12, 13]. Each disorder has different causes and treatment strategies however in the end, every one of them cause a major problem, sleep deprivation. Sleep deprivation due to sleep-related disorders or other factors could introduce severe physical effects, cognitive impairments and mental health complications. Sleep provides a refreshment from mental and physical fatigue and the lack of sleep decreases mental concentration and immunity [14]. During sleep, the body restores metabolic functions and stores information collected on the previous day [15]. Furthermore, proper sleep gets people ready for the next day and is essential for healthy activity and mental functionality in all aspects of daily life.

While sleep with good quality has a positive impact on the human's enthusiasm, learning process, creativeness, concentration level during the day; poor sleep quality has a strong relationship between physical diseases like high blood pressure, obesity, heart diseases, diabetes and stroke [16]. When the mental and/or physical condition of the sleeper and the ambient factors are changed, the functionality of the subject is impaired and this situation causes problems such as somnolence and mood swings. Those problems may cause irritability and cardiovascular disease (CVD) such as hypertension, arrhythmia and others. As a result, the life quality of the subjects and the people around them are affected negatively [17, 18].

Sleep is not only important for individual health but also important for the public health. Sleep deprivation can cause severe public health problems with devastating consequences and significant financial implications. Insufficient sleep can cause decrease of reaction time and loss of coordination which may result in motor vehicle

crashes, industrial disasters and medical and other 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 [19]. The main problem about sleep deprivation due to sleep-related disorders is that they are mostly stayed undiagnosed since people are unconscious in their sleep. Altevogt *et al.* [20] estimate that 50-70 million US adults have sleep or wakefulness disorder. Furthermore, approximately 6.62% of the Americans suffer from sleep apnea and half of them stay as undiagnosed [21]. This situation forces to take immediate intervention strategies to diagnose sleep-related disorders, eliminate destructive results and decrease financial burdens for both personal and public. 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 [22]. If both monitoring types are compared, the second type is much more reliable and appropriate for the recent studies. The first type contains different standardized questionnaires which learn sleep behavior of the subject from the answers of subjects themselves or their relatives. However, since the questions and the answers are subjective, the results can change according to the subject's current mood. The most popular standardized ones are Pittsburgh Sleep Quality Index (PSQI) [23] which is a self-rated questionnaire. It monitors the quality of sleep and disturbances over one-month time interval. Subjects score themselves between 0 and 21 by answering sleep-related questions and increase in PSQI score refers to a decrease in the quality of sleep. Another popular standardized questionnaire is Epworth Sleepiness Scale (ESS) [24] in which there are eight questions. This questionnaire monitors the chance of dozing off in different situations and measures excessive day-time sleepiness of the subjects. In the recent studies, researchers include new questions to the standardized ones according to their studies' objectives. Therefore, there are lots of non-standardized questionnaires created based on the standardized ones.

In the second type, 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. This technique is available in the hospitals and subjects must spend a night in the sleep room which is observed by a technician for all night. In the PSG technique, there are many sensors attached on the body of the sleeper to extract the sleep parameters and detect whether the subject has sleep-related disorders [25]. However, since this technique is expensive, overwhelming and time consuming, the subjects who afford or volunteer this examination is few compared to the potential number of subjects. Furthermore, the sleep statistics obtained with the PSG technique are not considered as absolutely true because the available PSG systems are very obtrusive, the evaluation is handled by a technician and the subject needs to sleep somewhere else other than his/her own bed. The latter means that the first night effect has negative influence on the results [26]. Lastly, this system violates the privacy of subjects because of the video record and observation of the technician. 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.

Another state of the art sleep monitoring technique is ActiGraph [27] that is a watch-like device with embedded sensors in it. Even though the subject places the device on the wrist which provide unobtrusiveness, the accuracy of ActiGraph is not as high as the PSG technique's. The reason behind that is the lack of technician and also this system is not stationary as the PSG system since the watch can move while sleeping. Because the interface of the system is not user-friendly, the help of a technician is necessary and it is not possible to use it for all nights. It has been used as the ground truth in many studies because it can be considered as a promising and good alternative. However, the additional cost of the system makes this technique less preferable for the systems targeting low-cost.

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. The subjects who need the assessment cannot reach those facilities immediately.

Therefore, a new mechanism for monitoring sleep is required. In today's technology, the smart phones and smart watches are equipped with many useful build-in sensors like accelerometer, sound level, heart rate monitor and barometer. Those sensors can be utilized to monitor human activities. With the increased Central Processing Unit (CPU) power and storage area, those smart devices can deal with complex operations easily and they are very unobtrusive compared to most of other systems. Another important advantage of smart devices is the cost. 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.

With the help of this information, our motivation is to develop an unobtrusive, cheap and reachable system which can monitor the subjects during the sleep in their own conditions. Since most of the people are not aware of their sleep-related disorders, a system which helps them to understand the severeness of their current situation in sleep is necessary. This system should not bother the subjects in their sleep as well as it should protect their privacy. Furthermore it can be controlled by subjects that is very beneficial for ease of integration in their daily life, on the other hand, it should not depend on subjective observations of subjects or their relatives since the success rate of detection of any sleep-related disease has a high correlation with those observations. By considering those conditions, the proposed system for this thesis contains a smart phone and smart watch.

The main objective of the proposed system is determined by considering all objectives in the field. The detection of sleep disorder without consulting a doctor and with a smart device may have devastating consequences. People have a tendency to believe such diagnosis which may not be completely accurate and taking an action by considering this diagnosis does not make sense. Therefore, we define our objective as sleep statistics assessment, especially assessment of respiratory information. Those 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 learn the

reasons behind their poor sleep quality and a doctor can investigate more about the patients' sleep quality and behaviors during sleep. With the help of this proposed system, subjects can understand the severity of their problems in respiration before consulting a doctor and doctors can reach objective observations about their patients. Significant amount of time and energy saved as well as financial costs and the burden on the doctors are decreased in both cases. Therefore, this proposed system is actually positioned as a prescreening tool or guidance system.

We tested our proposed system with 17 subjects, in which 9 have no abnormal respiratory problems whereas 8 have severe respiratory problems. Those subjects are volunteer patients of the sleep specialist doctor who guides us for this thesis. Our subjects are examined by the specialist doctor before participating our experiments and the respiratory events during their sleep are labeled with the guidance of the specialist doctor by analyzing their actual sleep statistics which are derived with the help of a PSG system in the sleep clinic.

The proposed system compared to the related studies in the literature has some contributions which are as follows.

- "Unobtrusive and comfortable design" is an important aspect. In a very basic point of view, people tend to get rid of anything disturbing them. However, the proposed system does not disturb the subjects. The smart watch is not different from any other watch and most of the people have already been sleeping with some accessories. Compared to the other overwhelming systems with lots of cables and sensors, one watch is negligible. On the other hand, the proposed system does not force the subjects to sleep somewhere else which requires more time, energy and motivation, it allows to sleep in their own bed. Furthermore, the proposed system gets help from the smart phone but does not give direct commands for the placement of the smart phone. Placing the smart phone in a disturbing place may decrease the willingness of the subject which has negative impact on the results. On the other hand, the success rate of the detecting abnormal respiratory events should not be neglected to achieve unobtrusiveness since, in a way, these two

parameters are in a trade-off. The balance between the unobtrusiveness and the success rate should be considered and the smart devices are placed as near as possible.

- "Objective monitoring" is crucial for this kind of studies. If the collected data is subjective and can be easily manipulated by the subjects, the results become unreliable. The proposed system does not allow such incidences. The sensors do not permit outside intervention and the imitation of behavior is not possible for all night long.

- "Protection of privacy" is getting more important for the people since the capabilities and functionalities of the smart devices increase day to day. The collected information is about the users' personal information and sleep activity details. In this proposed system, the identities of the users stay anonymous since the collected data is in a raw signal format. Because there is no need to record a video, the privacy is not violated.

The rest of the thesis is organized as follows. In Chapter 2, the related works are discussed in the context of sleep analysis domain by considering utilized devices and study objectives. A comparative list of related research in the literature is presented. Afterwards, we provide background information about the respiratory events, smart devices in today's technology and PSG technique those are the key components of the proposed system. In Chapter 3, the system design is given with its limitations and considerations. Furthermore, this chapter explains the experimental setup and the data collection procedure for each device in the system and provides detailed analysis about the subjects. The chapter is continued with the visualization of data and explanation of the abnormal respiratory event arousal. The challenges through generating meaningful data are presented and the feature extraction, selection and sampling processes are introduced. Chapter 4 starts with the classification methodology of different algorithms and results of these algorithms. Afterwards, the enhancements on the current results are discussed and different approaches are presented. The chapter ends with

the evaluation of performance as a prescreening tool. Lastly, in Chapter 5; conclusions and directions for future research are stated.

2. LITERATURE SURVEY

This chapter provides fundamental information about sleep studies in the literature which are evaluated according to the similarity between their objective and utilized devices with this thesis. Furthermore, the background concepts which are implied throughout this thesis 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 of the poor sleep quality is significant. By considering those reasons, identifying strategies which have a positive impact on the good sleep quality at an early stage eliminate the potentially devastating consequences. In this chapter, the related works in the field of sleep is presented in detail. 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. By considering all those aspects, we restrict our related work into the sleep studies that either try to achieve our objective; sleep statistics assessment, or utilize sensors in our design; accelerometer, heart rate monitor and sound level sensors.

The common objectives for the studies are sleep quality assessment, sleep disorder detection and sleep statistics assessment. The sleep quality assessment requires the information about sleep/wake time, sleep stages like REM-NREM period or sleep-wake period, sleep duration, daily activity information including social interaction and physical activity. Since sleep quality has correlations with many factors in daily life, the studies which have objectives like assessing the sleep quality should investigate every factor at a time. Some studies give advices about daily life or interfere the daily life in order to increase the sleep quality. The sleep disorder detection studies mainly

concentrate on detecting sleep apnea. From the movement pattern, breathing pattern, oxygen level, heart rate changes and noise changes, some researchers intend to find whether a subject has apnea whereas others measure the severeness of apnea. On the other hand, some researchers focus on not apnea but just snoring or leg movements which are also accepted as symptoms of some sleep disorders. Another common objective, sleep statistics assessment, provides information about behaviors in the sleep to the subjects and doctors. These sleep statistics help to investigate underlying reasons of the complaints of subjects and give a chance to take precautions before reaching critical level.

The sleep studies gather information from several sources. Mainly they utilize some sensors which can be build-in smart phones or smart watches or some external sensors which can communicate with the smart devices. The common smart phone or smart watch sensors are accelerometer, gyroscope, sound level, light, GPS, heart rate monitor, screen proximity and even clock. The general external sensors are pulse oximeter, ECG, earphone, etc. Furthermore, the researchers collect data from questionnaires, diaries or applications which are filled by the subjects themselves or their relatives. Since the collected information is not validated, the subjects in the sleep studies are generally monitored by other sources, like PSG, diaries and Zeo [28] which is a kind of head band that collects information from brain waves. This information coming from other reliable sources is known as the ground truth and is used to measure the success rate of the proposed system by comparing the results.

The studies which aim to detect sleep disorder have many similarities with our proposed system although our objective is assessing sleep statistics. The sleep disorder detecting studies mainly investigate the respiratory events like apnea. The studies conducted by Alqassim *et al.* [29], Nakano *et al.* [30], Sun *et al.* [31] and Tseng *et al.* [32] include just a smart phone and the build-in sensors of smart phones are used to collect data from subjects. Alqassim *et al.* [29] learn breathing and movement pattern of the subjects 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. Nakano *et al.* [30] attach the smart phone

to the anterior chest wall over the sternum in order to collect data from microphone. Sun *et al.* [31] also use microphone and collect data in different contexts like placing the smart phone in the pocket or on the desk. On the other hand, Tseng *et al.* [32] state that they utilize medical server with smart phone but not stated the placement of smart phone and activated sensors.

The sleep disorder detecting studies conducted by Behar *et al.* [33], Han *et al.* [34] and Al-Mardini *et al.* [35] activate the smart phone sensors with other external devices. Behar *et al.* [33] use the smart phone to record audio, activity and body position of the subjects whereas they measure the oxygen saturation with an external pulse oximeter sensor. Al-Mardini *et al.* [35] follow the same strategy with one difference. The smart phone's microphone measures the respiratory effort. Han *et al.* [34] also record audio and accelerometer signals however they make use of external ECG for measuring respiratory rate and their objective is not just disorder detection but also assessing the sleep quality.

Other studies that try to detect sleep disorder do not get direct help from build-in sensors of smart phone. Oliver *et al.* [36] and Patil *et al.* [37] measure oxygen saturation with external pulse oximeter sensor and the smart phones are just a medium for transferring or interpreting the data. On the other hand, Nandakumar *et al.* [38] and Ren *et al.* [39] convert smart phone into another device. Nandakumar *et al.* [38] use the smart phone as a sonar radar which measures the chest movements of the subject whereas Ren *et al.* [39] convert the smart phone into special audio recorder which collect data with the help of microphone and ear buds.

All the studies which try to detect sleep disorders have one common disadvantage, unobtrusiveness. The placement of smart phone, pulse oximeter sensor, ECG or ear buds have a significant effect on the data collection process. In order to increase the success of the study, the researchers should give direct orders about the placement and also they should be careful about the cables. This situation decreases the comfort of subjects dramatically.

The sleep quality assessment studies generally utilize a smart phone stand-alone. Since quality assessment requires information about the daily life of subjects, Bai *et al.* [40] and Lane *et al.* [41] follow the subjects all day long and utilize accelerometer sensor, microphone and GPS. Bai *et al.* [40] also activate the light sensor. Both record social interaction and physical activity of subjects. On the other hand, the sleep quality depends on the sleep time, sleep duration, wake time, sleep-wake stages during night and sleep disturbances. Gautam *et al.* [42], Gu *et al.* [43], Min *et al.* [44] and Sano *et al.* [45] track the subject only during nights and try to understand sleep-wake stages of subjects. Although Gautam *et al.* [42], who activate only accelerometer sensor, and Sano *et al.* [45], who use wearable sensors for activity and skin temperature measurements, others monitor environment by activating light and screen proximity sensors and monitor behavior of the subject with the help of the phone status like application usage and personal information. Hao *et al.* [46] also monitor subjects during night but listen body movements with microphone in order to extract sleep disturbances. Furthermore, Suzuki *et al.* [47], who activate the accelerometer and heart rate monitor sensors on smart watch try to find sleep stages.

The main drawback of the sleep quality assessment studies is interruption during data collection. All day long tracking is not possible because people do not generally carry smart phone everywhere and during nights the smart phones are not attached on the body but placed nearby [42, 43, 46].

The sleep statistic assessment studies focus on different aspects of sleep. Chen *et al.* [48], Natale *et al.* [49] and Sano *et al.* [45], where all utilize a smart phone and a smart watch, compare the extracted statistics with the results obtained by other sensors or diaries. Both Chen *et al.* [48] and Natale *et al.* [49] estimate sleep duration whereas Sano *et al.* [45] try to understand the effect of sleep statistics on daily life. The placement of smart phones is not stated in any of those studies, however the activated sensors are explained in detail. Chen *et al.* [48] measure sleep duration with the help of accelerometer and light sensors, microphone and phone status. On the other hand, Natale *et al.* [49] and Sano *et al.* [45] activate the accelerometer sensor and additionally, Sano *et al.* [45] measure the skin temperature and track the phone status. Bernardeschi

et al. [50] and Hernandez *et al.* [51] estimate the heart rate with accelerometer data on the smart phone and smart watch, respectively. Furthermore, Min *et al.* [44] detect sleep stages with a smart phone and Uçar *et al.* [52] try to do the same thing with the Photoplethysmogram (PPG) sensor in smart watches. The study conducted by Sun *et al.* [53] in 2017 estimates the respiration rate with the accelerometer sensor in the smart watch.

All studies should be validated with more accurate data coming from other sources. Therefore several ground truth strategies are followed on the experiments. Filling questionnaire, diary or survey is a very common method [40, 41, 44, 45, 48, 50], however it is not trustworthy since the subjects or their relatives can manipulate the results. Some studies get help from a physician for the ground truth which is much more trustworthy but the subjectivity of the physician is not eliminated completely [29, 33]. The most trustworthy ground truth strategy is monitoring patients with the PSG system which is the gold standard for sleep studies [30, 35, 36, 38, 47, 52]. On the other hand, since the availability of a PSG system is limited, collecting the ground truth data with other sensors is a good alternative. Gautam *et al.* [42] and Gu *et al.* [43] get help from Zeo [28] and Natale *et al.* [49] use ActiWatch. Hao *et al.* [46], Sun *et al.* [31] and Ren *et al.* [39] make use of the external microphone where Ren *et al.* [39] also use an external respiration monitor sensor. Hernandez *et al.* [51] utilize extra ECG and PPG sensors. Furthermore, if the collected data is extracted from a medical database, the validity of the data is already checked and Han *et al.* [34] and Tseng *et al.* [32] make use of this advantage. The study conducted by Sun *et al.* [53] in 2017 utilize video recording for validation.

The subjects information are given in Table 2.1 with the summary of information explained above. The activated sensors are presented with their source; SP for smart phone, SW for smart watch and ext. for external resources. This table also includes information about the learning strategies. The main tendency is implementing Support Vector Machines (SVM) [54] which is a well-known method that analyzes data for classification and regression. Behar *et al.* [33], Gu *et al.* [43], Han *et al.* [34], Al-Mardini *et al.* [35], Sano *et al.* [45] and Sun *et al.* [31] follow this strategy. Gu *et al.* [43] also

propose a linear-chain Conditional Random Field (CRF) [55] that is a discriminative model in which the global probability of a sequence of random variables are estimated. CRFs are mainly the advanced version of logistic regression [56] and widely used in the speech and language processing applications.

Another common technique is the Decision Tree (DT) [57] which is a popular method in the literature in which internal nodes represent a test on an attribute, leaf nodes represent a class label and the paths from the root to leafs represent classification rules. It is commonly used for decision analysis, feature selection, classification and operations research. Chen *et al.* [48], Hao *et al.* [46], Min *et al.* [44] and Tseng *et al.* [32] utilize specialized decision trees [58]. Beside the decision tree, Min *et al.* [44] propose a Bayesian Network (BN) [59] which is a probabilistic graphical model in which a set of random variables and their conditional dependencies are presented with a directed acyclic graph and 1R Classifier (1R) [60] that builds a set of rules based on only one feature.

Mathematical methods like Fourier transform, filtering strategies and some statistical and heuristic methods are proposed by Alqassim *et al.* [29], Bai *et al.* [40], Bernardeschi *et al.* [50], Gautam *et al.* [42], Hernandez *et al.* [51], Nakano *et al.* [30], Nandakumar *et al.* [38], Natale *et al.* [49], Oliver *et al.* [36], Ren *et al.* [39], Uçar *et al.* [52] and Suzuki *et al.* [47]. Among those, Gautam *et al.* [42] also implement a Hidden Markov Model (HMM) [61], which is presented as a simpler dynamic Bayesian Network and the previous state influences the state recognition result of the next state. Furthermore, Lane *et al.* [41] apply Logistic Regression (LR) [56] whereas Patil *et al.* [37] perform a Random Forest (RF) [62] that is an advanced version of the decision tree and K-Means Algorithm [63] which is an unsupervised learning method that solves the well-known clustering problem. The study conducted by Sun *et al.* [53] in 2017 uses four different machine learning algorithms; Bayesian Network, Decision Tree, Random Forest and Naive Bayes which a probabilistic classifier utilizing the Bayes' theorem [64].

The extracted features for each study are distinguished according to their study objectives and utilized sensors. The main tendency is calculating magnitude, minimum,

maximum, average and other range variables of the accelerometer signals [29, 34, 35, 42, 48, 50]. Calculating spectral centroid and rolloff is another common approach however those features are useful for noise analysis [31, 43]. For the heart rate signals, the amplitude and peak information are widely used. However in general, studies generate their high level features based on raw signals. The general demographic data and phone status are also good sources in feature extraction.

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 analysis of sleep studies shows that there is no way to combine or group the obtained results because the studies have different objectives, focus on different aspects of sleep while using different sensors and apply different methods. Therefore, the main objective of researchers is proposing systems that do not contradict the general domain knowledge and ease the life of subjects.

Our proposed system design actually combines several studies. We combine smart phone and smart watch like the systems in [35, 48]. However, we differ according to the ground truth methodology and objective. With the help of the studies mentioned above, we understand the feature extraction, feature selection and machine learning approaches used in sleep studies and analyze which approaches are more useful for this thesis.

Table 2.1. Related Work.

Ref.	Objective	Activated Sensors	Subjects / Ground Truth	Features	Methodology	Results
[29]	Disorder detection	Accelerometer(SP), Microphone(SP)	55 samples (details not given) / Physician	Magnitude of X,Y,Z from accelerometer signal, Frequency spectrum of signal from sound, Threshold combines low level features	Mathematical	Text messages sent by the physician
[30]	Disorder detection	Microphone(SP)	50 participants (details not given) / PSG	Top 1 percentile sound pressure level (L1) and the equivalent sound pressure level (Leq)	Statistical	Diagnosing apnea sensitivity = 0.70, specificity = 0.94
[31]	Disorder detection	Microphone(SP)	16 participants 204 days of data / Microphone(ext.)	Root Mean Square, Above a-Mean Ratio, Average of Top k RMSs, Left to Right Ratio, Relative Maximum RMS, Zero Crossing Rate, Spectral Centroid, Bandwidth, Spectral Rolloff from sound signal	Support Vector Machine	Respiratory symptom classification accuracy = 82%
[32]	Disorder detection	not stated(SP), medical server	540 participants (medical database) / Medical Database	Gender, Age, Neck circumference, Body weight, Pre-overnight sleep systolic blood pressure, Post-overnight sleep systolic blood pressure	Decision Tree	Apnea detection for simple model accuracy = 89.07%, best model accuracy = 96.85%
[33]	Disorder detection	Accelerometer(SP), Microphone(SP), Oximeter(ext.)	856 participants (medical database) / Physician	Multiscale Entropy from audio and actigraphy, Average number of oxygen desaturation events, Available demographics(body mass index, weight, height, age, neck size, gender)	Support Vector Machine	Apnea-Healthy classification accuracy = 92.2%

Table 2.1. Related Work (cont.).

Ref.	Objective	Activated Sensors	Subjects / Ground Truth	Features	Methodology	Results
[34]	Disorder detection, Quality assessment	Accelerometer(SP), Microphone(SP), ECG(ext.)	3 participants 21 hours of data Medical Database	Magnitude of X,Y,Z from accelerometer signal, Amplitude of sound signal, Detection of QRS complex from ECG signal	Support Vector Machine	Average sleep quality detection accuracy is = 75%
[35]	Disorder detection	Accelerometer(SP), Microphone(SP), Oximeter(ext.)	15 participants (details not given) PSG	Magnitude of X,Y,Z from accelerometer signal, Apnea-Hypopnea Index from sound signal, Oxigen Desaturation Index from oximeter sensor	Support Vector Machine	Root mean square error for detecting the severity of apnea = 32.5
[36]	Disorder detection	Oximeter(ext.)	20 participants (details not given) PSG	% of desaturation below the baseline and in total duration of the event in minutes, the amplitude of the peak	Heuristic	Severity of the apnea condition increases, the peaks increase in the $0.015 - 0.04$ frequency range
[37]	Disorder detection	Oximeter(ext.)	(details not given) not stated	SpO2 Variability as Delta Index, Oxygen Desaturation Index	Random Forest, K-Means	Not stated
[38]	Disorder detection	Sonar system(SP)	37 participants 296 hours of data PSG	Amplitude and periodicity of the chest motion signal	Mathematical, Heuristic	Hypopnea detection accuracy = 0.9533%, central apnea detection accuracy = 0.9957%
[39]	Disorder detection	Microphone(SP), Earphone(SP)	6 participants 36 months of data Respiration Sensor(ext.), Microphone(ext.)	Trend changes as Envelope	Filtering, Mathematical	Breathing rate error = 0.5 bpm, Dleep event detection = 80%

Table 2.1. Related Work (cont.).

Ref.	Objective	Activated Sensors	Subjects / Ground Truth	Features	Methodology	Results
[40]	Quality assessment	Accelerometer(SP), Microphone(SP), Light(SP), GPS(SP)	15 participants 450 days of data Questionnaire	Extracted activity from accelerometer signal, Average sound and light signal information, Daily trajectory of subjects from GPS signal, Attribute, Social Correlation and Sleep Correlation Factors extracted	Heuristic	Sleep quality prediction accuracy = 78.26%
[41]	Quality assessment	Accelerometer(SP), Microphone(SP), GPS(SP)	5 participants 5 weeks of data Survey	Gaussian formula estimated with total sleep time and ideal sleep time, A value represents duration of performed activities, Total duration of ambient conversations	Logistic Regression	Mean absolute error for sleep duration estimation = 1.54 hours
[42]	Quality assessment	Accelerometer(SP)	4 participants 48 nights of data Zeo	Magnitude of X,Y,Z from accelerometer signal	Hidden Markov Model, Statistical	Sleep-Wake average accuracy for Hidden Markov Model = 79%, for Statistical = 68%
[43]	Quality assessment	Accelerometer(SP), Microphone(SP), Light(SP)	15 participants 30 nights of data Zeo	Micro and macro body movements from accelerometer sensor, Zero Crossing Rate, Spectral Entropy, Spectral Centroid, Spectral Flux, Bandwidth, Spectral Rolloff from microphone, Three types of illumination conditions from light sensor, Sleep duration calculated by chronological order of sleep stages, Age	Support Vector Machine, Conditional Random Field	Light sleep detection precision = 66.74%, overall accuracy = 64.55 %

Table 2.1. Related Work (cont.).

Ref.	Objective	Activated Sensors	Subjects / Ground Truth	Features	Methodology	Results
[44]	Quality assessment, Statistic assessment	Phone status, Accelerometer(SP), Microphone(SP), Light(SP), Proximity(SP)	27 participants 795 days of data Diary	(Minimum, Q1, Median, Q3, Maximum, Avg., Standard Deviation, count of peaks, Avg. width of peaks, Avg. height of peaks, interval of peaks, position of peaks) of movement, noise level and light intensity, Duration of screen-on time, Battery State, Bedtime, Waketime, Sleep duration	Decision Tree, Bayesian Network, 1R Classifier	Sleep quality detection accuracy = 81.48%, Sleep state detection accuracy = 93.06%
[45]	Statistic assessment, Quality assessment	Phone status, Accelerometer(SW), Temperature(SW)	66 participants 1980 days of data Survey	700 features which contains sleep time, bed time, sleep duration, frequency and duration of phone calls or messages, scores obtained from different questionnaires, social interaction and activity details	Support Vector Machine	Classification accuracy range 67-92%
[46]	Quality assessment	Microphone(SP)	7 participants 51 nights of data Microphone(ext.)	Loudness, Ratio of low-band to high-band energies, Variance from sound signal	Decision Tree	Sleep-related event classification accuracy = 90%
[47]	Quality assessment	Accelerometer(SW) Heart Rate Monitor(SW)	45 participants 360 hours of data PSG	The amount of activity, Pulse-to-Pulse Intervals	Statistical Mathematical	Coincidence ratio for sleep staging = 0.735

Table 2.1. Related Work (cont.).

Ref.	Objective	Activated Sensors	Subjects / Ground Truth	Features	Methodology	Results
[48]	Statistic assessment	phone status, Accelerometer(SW), Microphone(SP), Light(SP)	8 participants 500 hours of data Diary	Average, minimum, maximum, magnitude of X,Y,Z from accelerometer signal, Intensity of light signal, Duration of phone lock, Phone recharging, Phone off, Static state during sleeping hours, Silence of phone	Decision Tree	Sleep duration error for best effort sleep model = 41 min, for sleep with phone model = 23 min
[49]	Statistic assessment	Accelerometer(SP), Accelerometer(SW)	13 participants 52 nights of data ActiWatch	Sleep onset latencies; the first minute of 5-min immobility block and the first minute of 10 min of immobility with activity in no more than one epoch	Statistical	Sleep/Wake discrimination range 87.3-89.9%
[50]	Statistic assessment	Accelerometer(SW), Heart beat(SW)	3 participants 60 nights of data Diary	Standard deviation of the acceleration magnitude, Variance of the heart- beat rate, Approximate sleep staging	Heuristic	Diary results are lower than computed results
[51]	Statistic assessment	Accelerometer(SP)	12 participants (details not given) ECG, PPG	Not stated	Filtering	Mean absolute error for heart rate estimation = 0.88, for breathing rate estimation = 0.38

Table 2.1. Related Work (cont.).

Ref.	Objective	Activated Sensors	Subjects / Ground Truth	Features	Methodology	Results
[52]	Statistic assessment	PPG(ext.)	10 participants 80 hours of data PSG	Average, Standard Deviation, Average Curve Length, Average Energy, Average Taegar Energy, Hjort Activity, Hjort Mobility, Hjort Complexity, Maximum, Skewness, Kurtosis, Shape Factor, Minimum, Root Mean Square, Singular Value Decomposition from heart rate signal	Statistical	Features with p <0.05 is distinctive for apnea detection
[53]	Statistic assessment	Accelerometer(SW)	16 participants 70 nights of data Video Record	Mean, Standard Deviation, 20th percentile, Median, 80th percentile of X, Y, Z from accelerometer signal, Covariance and correlation between accelerometer data along every two different axes, Body Position	Naive Bayes, Bayesian Network Decision Tree Random Forest	Random forest respiration rate accuracy in supine position = 95%

2.2. Background Concepts of Proposed System

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

2.2.1. Respiratory Event

The respiratory events that are mentioned in this thesis are apnea events. The sleep apnea can be defined as pauses in breathing or shallow breaths in sleep [65]. 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 start of the normal breathing.

There are three types of sleep apnea: central apnea, hypopnea and obstructive apnea. The central apnea occurs when the subject holds his breath for a non-negligible period whereas the hypopnea occurs when the subject's breathing becomes shallow and the amplitude of chest motion and nasal pressure is decreased. The obstructive apnea occurs when the subject's upper airway is blocked partially or completely and this one is the most frequently diagnosed type of sleep apnea. The oxygen saturation level change and the heart rate change are the clear indicators of apnea events. Since the body gives response afterwards, the body movements and loud snoring follow the apnea events.

The sleep apnea disorder is more common among the overweight people. Furthermore, the risk increases with increasing age. 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 life-style 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 Equation 2.1 and it is called Apnea-Hypopnea Index (AHI) [66].

$$Apnea - HypopneaIndex(AHI) = \frac{ApneaEventCount}{TotalSleepDuration(hours)} \tag{2.1}$$

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. The first person diagnosing this disorder might be a family member or the bed partner. 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 since a longer deep sleep stage contributes to better sleep. Therefore a person suffering from sleep apnea would have a poor sleep quality [67].

2.2.2. Smart Watch

Smart watches are electronic devices which are commonly available in the market. There are several different models and manufacturers. In recent years, they are utilized as wearable computers which may also have mobile phone capabilities. They generally include a touch-screen for interaction with users and different communication technologies to connect with computers and smart phones. In recent years, the sensors inside the smart watches are very useful for collecting information about the user and

the environment. These sensors are mainly accelerometer, gyroscope, Global Positioning System (GPS) receivers, barometers, thermometers, heart rate monitor (HRM) sensor. Because of reduced cost, unobtrusiveness and the precision of measurements with sensors, smart watches are highly preferred in health related applications in recent years [68].

2.2.3. Smart Phone

In today's world, smart phones are the most popular electronic devices which can be used as mobile phone, camera, agenda, media player, gaming console. Even if they have lower computing capabilities than personal computers, the smart phones substitute the computers since people can carry them everywhere. Their common usage, reduced cost, CPU power, storage area, embedded sensors and communication technologies make them a good candidate for life-logging activities. They may be attached on the body for collecting personal information about the subject and also they are placed as close as possible, i.e. in the same room, to the subject for understanding the environment conditions [69].

2.2.4. Polysomnography

The polysomnography is the gold standard tool for sleep analysis. It provides information about the body functions during sleep and it is mainly used to diagnose sleep disorders. Today's PSG collects data from Electroencephalography (EEG) for monitoring brain waves, Electrooculogram (EOG) for measuring eye movements, mental-submental Electromyogram (EMG) to detect leg movements, muscle activity, Electrocardiogram (ECG) for assessing heart rate, airflow, respiratory effort, sound to record snoring, peripheral pulse oximetry, body position and continuous video recording.

The general electrode placement on the human body for PSG systems can be seen in Figure 2.1. The electrode selection can change according to the objective of monitoring which is determined by specialist doctors.

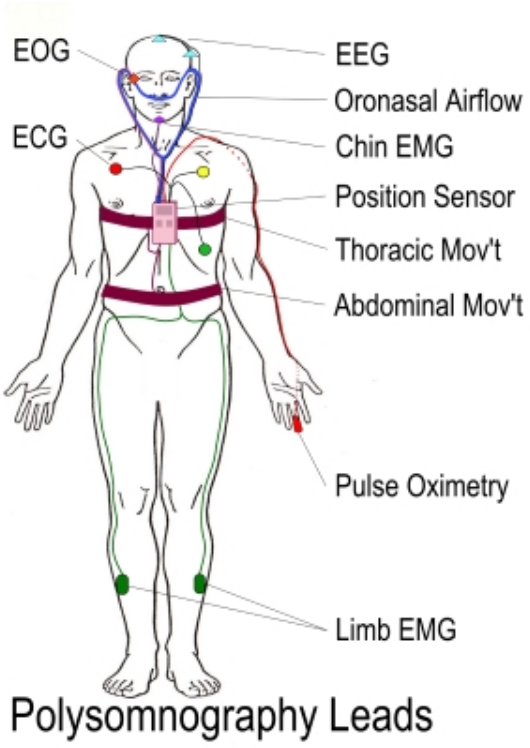


Figure 2.1. Human Body with the Polysomnography Electrodes Attached [1].

The PSG systems are available in hospitals or clinics and the patient comes one or two hour earlier for introduction about the system, preparation and placement of electrodes. The sleep technician is responsible for the preparation process and also after the patient falls asleep, he follows the patient all night long from video and the computer that displays all collected data. When the patient wakes up in the morning, the monitoring ends.

The analysis of collected data can be automatically generated with the help of software by considering previously entered parameters. Those parameters are determined by the American Academy of Sleep Medicine (AASM) [70]. However, since the automatic analysis results are not precise, the doctor can handle the analysis manually which provides more accurate outcomes [71,72]. Even though manual analysis provides more successful results, the diagnosis may vary depending on the examiner with a 20% variance [47].

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. This mask provides oxygen and is placed on the mouth at the second night. It is useful for understanding the oxygen saturation level which is different for each person. The mask with the determined oxygen saturation level is suggested to wear for all nights in the rest of the life of subjects [73].

3. PRESCREEENING SYSTEM ARCHITECTURE

This chapter explains the design of the prescreening tool which helps us to recognize people with severe respiratory problems and can give suggestions like "consult a doctor". In order to recognize people with severe respiratory problems, we design a respiratory event detection system which mainly distinguishes abnormal respiratory events from normal respiratory events.

The proposed system is described step-by-step. In order to make the system preferable with respect to other monitoring devices especially the PSG technique, it should collect the most meaningful data from the subject while it should not ignore the objectives. Therefore, the first idea is to utilize Electroencephalography (EEG) which analyzes brain waves [74] and include a pulse oximeter sensor which determines arterial blood oxygen saturation [75]. Although the PSG system collects data with EEG and pulse oximeter sensors, the attached cables on the human body are very uncomfortable and make the PSG technique obtrusive. Therefore, we eliminated these devices in our system.

The proposed system contains a smart phone with a smart watch that is placed on the wrist of the subject. These smart devices meet most of the requirements mentioned above. The data can be collected in subjects' own conditions without any help of a technician. Those devices are unobtrusive, comfortable and cheap. The privacy is not violated since those devices provide raw signals and these raw signals eliminate the subjectiveness.

The proposed system has some limitations. The first limitation is the battery limitation for the smart watch and the second problem is the limited number of activated sensors which are accelerometer, gyroscope and heart rate monitor sensor of the smart watch and the sound level sensor of the smart phone. Because of those limitations, the system cannot provide information about all sleep-related diseases since the collected data do not cover all aspects of sleep but only represents abnormal res-

piratory events like apnea. Those abnormal respiratory events provide the necessary information for determining the severity of problems in subject's respiration during sleep. This information is certainly an indicator of the sleep quality.

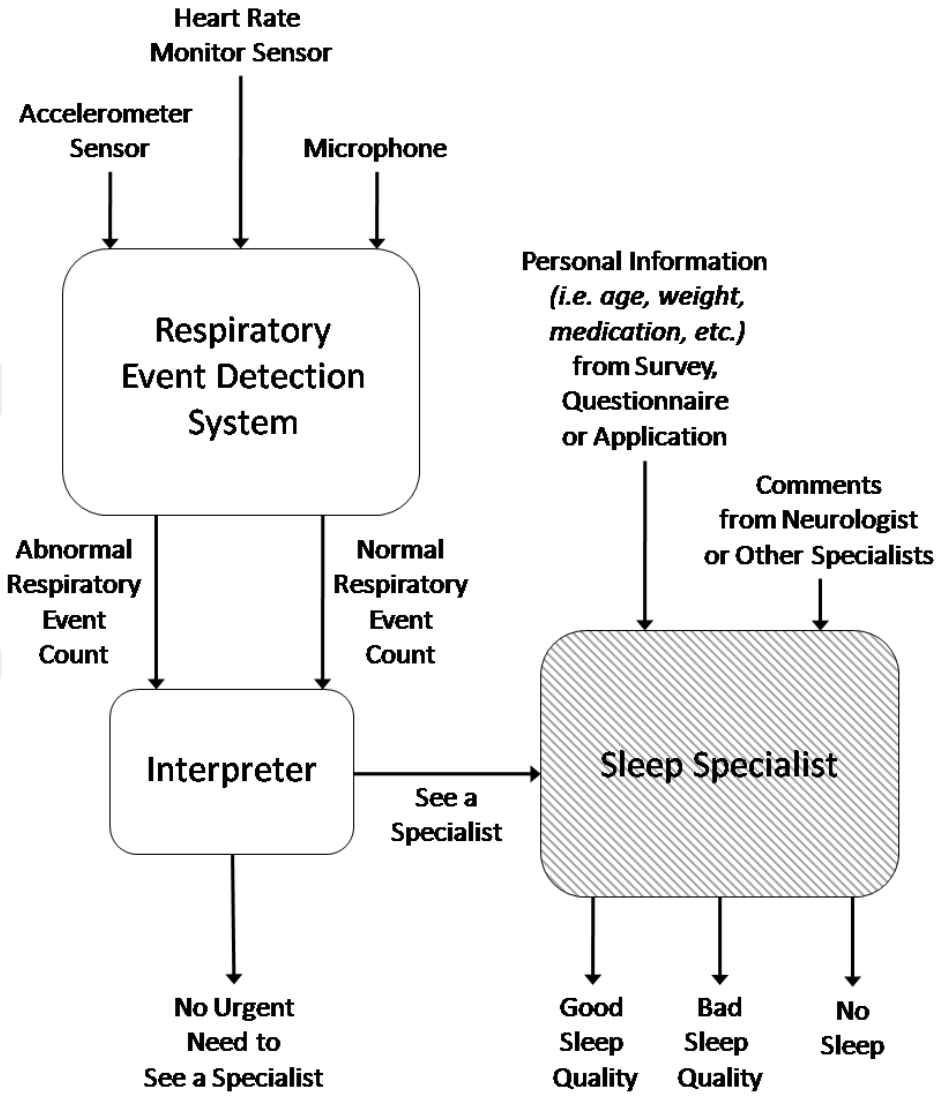


Figure 3.1. Prescreening System Design.

Utilizing the proposed system that contains smart devices, experiments are conducted and some sleep statistics are examined and determined. Sleep features are extracted out of these sleep statistics to determine the abnormal respiratory events. Extracted sleep features are eliminated with the help of feature selection algorithms and several machine learning algorithms are employed to differentiate respiratory events into two classes; normal and abnormal respiratory events. In this way, the proposed

system learns to classify events and can determine the abnormal respiratory events of future subjects. Figure 3.1 explains the prescreening system architecture. The shaded box represents the specialist medical doctor, the other boxes represent different parts of our proposed system.

3.1. Detection System Design

This section presents the determination mechanism of abnormal respiratory events which requires the interpretation of signals coming from real patients. The components of the respiratory event detection system are presented in Figure 3.2. The detection system mainly contains five components; data collection, feature extraction, undersampling, feature selection and classification. 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 combined with the true labels coming from the PSG system. The combination is stated as "Raw Signals".

Afterwards, we process the data and extract some meaningful features from raw signals. This data is called as "All-Featured Data" which contains all generated features. In order to balance the normal and abnormal respiratory events, we undersample that data without eliminating any features and name it as "Undersampled-Featured Data". This undersampled-featured data is given into some feature selection algorithms; Random Forest and Stepwise Feature Selection, which are useful for understanding the most important features. The data generated with these algorithms is called as "Selected-Featured Data". The data processing, feature extraction processes, undersampling operation and determination steps of the most important features are extensively explained in the related sections. All of these are implemented in MATLAB environment and that is preferred because of its computational power.

The classification component is mainly designed to measure the system performance. We employ different machine learning algorithms; Hidden Markov Model (HMM), Decision Tree (DT), K-Nearest Neighbors (KNN), Random Forest (RF) and Support Vector Machine (SVM), and evaluate their performance. The derived results

of classification algorithms are generated with different subsets of the raw and featured data. All implemented algorithms are clearly explained in the following sections. and they are also employed in MATLAB. However there is one exception. The classification algorithm preferred for raw signals is implemented in Python which is very useful with its various machine learning libraries.

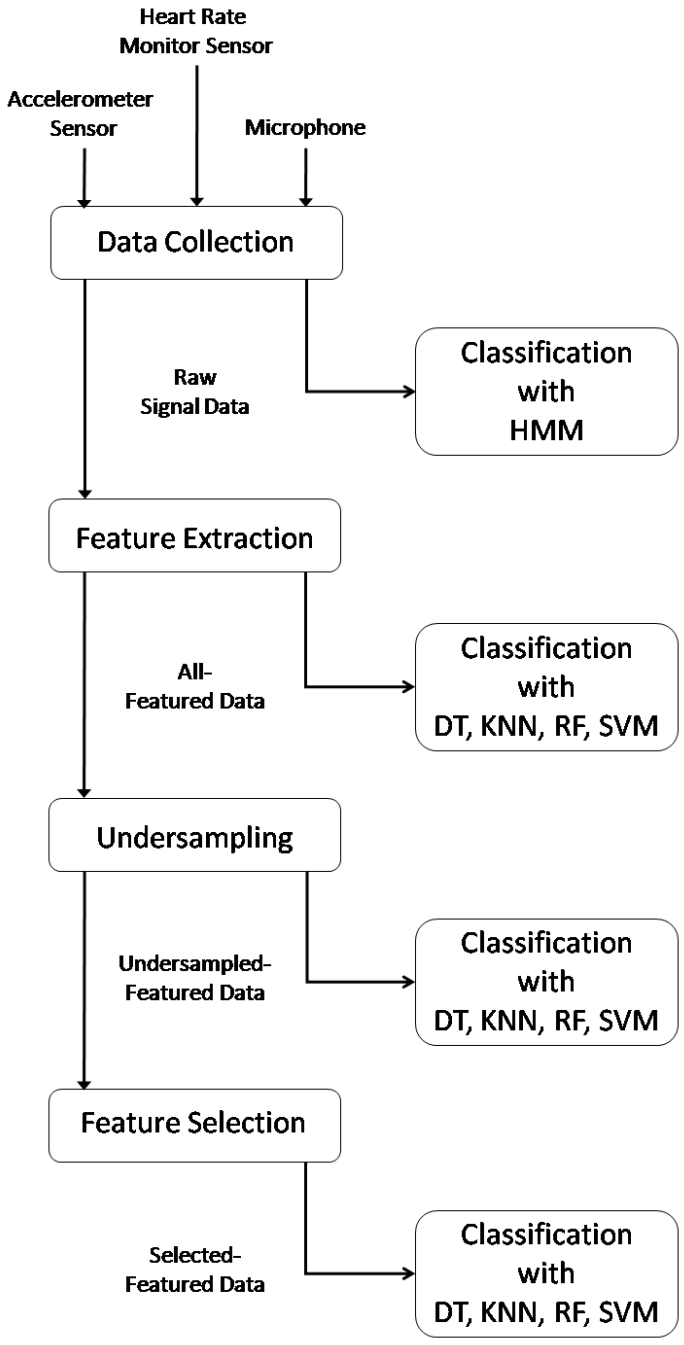


Figure 3.2. Respiratory Event Detection System Description.

3.2. Data Collection with Proposed System

The data collection is realized with the help of three different devices; smart watch, smart phone and PSG system and this section explains each data collection approach in detail.

3.2.1. Smart Watch

In this study, the personal information about the subject is collected with the help of a smart watch. We use Samsung Gear S3 [76] by considering the general advantages of smart watches which are mentioned in Section 2.2.2. Another reason of this choice is that some of other watches do not permit to access the collected raw signal and also do not give permission to change the sampling rate of sensors. Furthermore, this watch includes the accelerometer, gyroscope and heart rate monitor sensors which are obligatory for our sleep study. On the other hand, the smart watch is not equipped with any other sensors which are useful for our sleep study. This situation prevents collecting data about all aspects of sleep but just provides data that can be an indicator of abnormal respiratory events.

Our research group has developed a data collection application for Samsung Gear watches which provides the raw sensor data. Since the operating system in these watches is Tizen [77], this application is developed in the Tizen Development environment using the Hypertext Markup Language (HTML), JavaScript and Cascading Style Sheets (CSS) [78]. With this application, the users can be tracked all-day-long to the extent permitted by the battery.

The main drawback of the smart watch is its battery life. Since we need more precise data, we increase the sampling rate of sensors from the application. This decreases the life of battery. Even if it is re-chargeable, the charging unit is connected to the watch from the bottom side which interrupts the data collection. In order to increase the battery life, some sensors like the barometer or GPS, that do not serve our purposes, are deactivated. Furthermore, we can activate the sound level sensor

for determining the noise level of the room but it is not a feasible solution since the subject can place his wrist under the blanket. This situation may prevent or interrupt the data collection.

For this study, the extracted accelerometer sensor data do not include the gravity force which affects the "z" variable of the accelerometer signal and decreases the precision of data. The sampling rate of the accelerometer signals is 20 Hz whereas, the heart rate monitor sensor's sampling rate is 10 Hz. With these sampling rates, the battery lasts after approximately 6 hours of continuous data collection. The collected data is stored in the device and transferred on the computer afterwards. The smart watch is worn on the wrist as tight as possible as shown in Figure 3.3 in order to prevent data loss. The collected data is transferred to the computer with the help of Tizen Development environment and the files are in ".txt" format.



Figure 3.3. Smart Watch on Wrist.

3.2.2. Smart Phone

In this study, the environmental information is collected with the help of a smart phone. An Android application, Smart Voice recorder [79], that is available in the Google Play Store is downloaded and records the voice data of the subject and environment. At the beginning, we can use a noise recorder but we prefer to record voice since we would need to differentiate the events more precisely. However, the experiments prove that we just need the noise signals. The sampling rate of this voice data is 16000 Hz and the records are saved in ".wav" format. Since the voice data is too large, we transfer the files to computer manually.

Before the experiments start, we plan to gather the information from the light, humidity and temperature sensor with the Sensor Log application which is also available in Google Play Store [80]. However, we cannot get any significant results since the experiments are conducted in a controlled room in a sleep clinic therefore we eliminate this part.

Even if people carry the smart phones very close to their body all day long, they are not very willing to sleep in the same room with smart phones since they have some concerns about the radiation. In order to comfort the subjects, we turn the airplane mode on in the experiments which means that there is no connection and radiation risk.

3.2.3. Polysomnography

For this study, we utilize the PSG system for the ground truth. The sleep clinic that we contact with the subjects has two different fully operational PSG systems and the utilized PSG system is called as SomnoStar Sleep System [81]. This system is capable of getting information from all of 64 electrodes but the electrode set in our sleep clinic consists of 18 electrodes which are determined according to the guideline provided by AASM. The attached electrodes on the subjects' bodies with their detailed information are shown in Table 3.1.

The collected data are divided into epochs (30 seconds period) and the specialist doctor generates the labels manually. The labels that contain time information, indicate sleep stages, sleep/wake times, turn overs, leg movement events and respiratory events with their types. The PSG system provides an opportunity to extract files in the European Data Format (EDF) that is an old-fashioned method for storing multichannel biological and physical signals [82]. With the help of this format, we can reach the raw signals provided by PSG system.

Table 3.1. 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

On the other hand, the available PSG system has a user-friendly interface in which the user marks the labels as shown in Figure 3.4. This label information, which contains the true labels of each epoch, is stored in another file which has a specific extension, ".epc". The files which are not encrypted by the software can be processed with MATLAB.

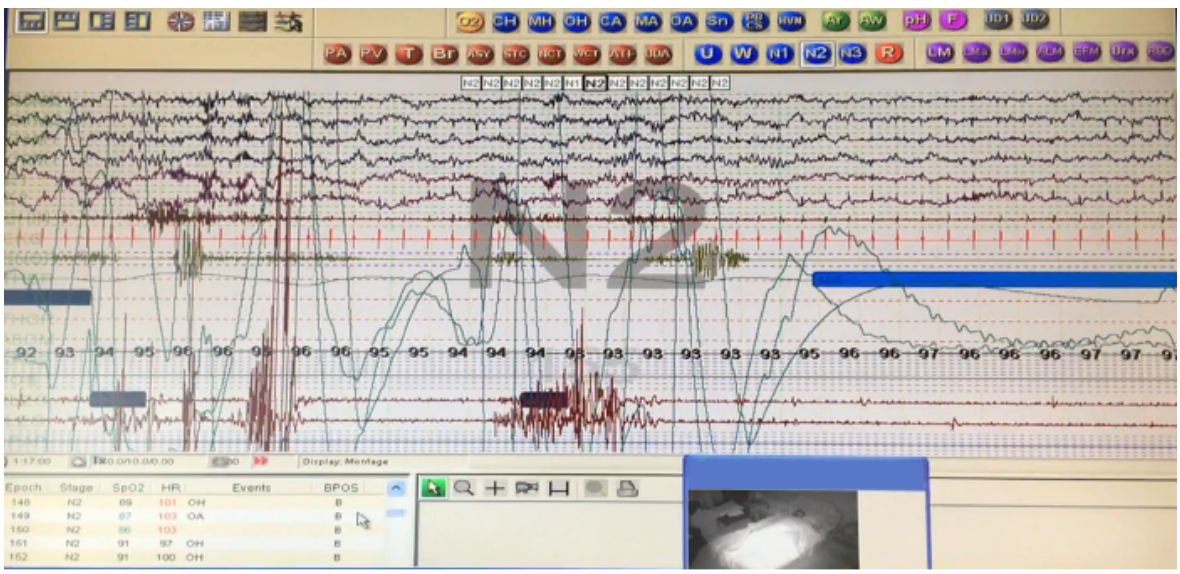


Figure 3.4. Interface of the PSG System.

3.2.4. Experimental Setup and Subject Information

The experiments are conducted in the sleep clinic and the procedure is approved by the Ethical Committee of Boğaziçi University (INAREK). The experiment design is the combination of three different devices; smart watch, smart phone and PSG system. Firstly, the sleep technician prepares the subject for the PSG system. After all the electrodes are placed, the subject wears the smart watch and the smart phone is positioned on the nearest place. Figure 3.5 shows a subject whose preparation is completed. The subjects are not allowed to read or watch anything after the preparation is completed in order to avoid attention deficit.



Figure 3.5. A Subject Sleeping with the PSG System [2].

At the beginning, the proposed system is tested without considering any criteria about the subjects. Since the capabilities of this proposed system is unknown, the criteria is determined in due course. At the end, the main criteria is determined as "Detecting Abnormal Respiratory Events". In order to accomplish this objective, the subjects who have severe apnea and the patients who sleep with a CPAP mask are preferred. The reason of this choice is emphasizing the difference between the subjects since the CPAP mask, a special mask for apnea patients, is mainly designed to prevent abnormal respiratory events and increases the comfort in breathing while

sleeping. Therefore, no or a few abnormal respiratory events occur while monitoring the subjects with CPAP mask. On the other hand, the severe apnea patients are chosen according to their true diagnosis, namely Apnea-Hypopnea Index (AHI), which is determined with the help of the PSG system. All the subjects are examined by the specialist doctor.

Table 3.2 gives the summary of all subjects who attended our experiments. We eliminated nearly 15 subjects' data by considering the severity of their sleep apnea disorder and some other problems like data interruption and data loss. Finally 17 subjects participated in this study, among those subjects; 8 of them have abnormal respiratory events while sleeping and 9 of them have only normal respiratory events with CPAP mask.

While classifying subjects as normal or abnormal, the main criteria is their true AHI score which is generated by the PSG system and the values of True AHI is also presented in Table 3.2. As presented in the table, the subjects who is stated as "Normal" can also have some abnormal respiratory events. However their True AHI score is very low compared to the subjects who is stated as "Abnormal".

As given in Table 3.2, one of the subjects is female and the average duration of sleep with our proposed system is calculated as 6 hours, 38 minutes approximately. The total duration for subjects who have abnormal respiratory events is 55 hours, 46 minutes which is balanced with the duration for subjects who have only normal respiratory events; 56 hours 43.5 minutes.

Table 3.2. Information about Subjects.

ID	**Gender**	**Status**	**Duration**	**True AHI**
311	Male	Normal	5 hours, 10 minutes	0
313	Male	Normal	6 hours, 15 minutes	12
316	Male	Abnormal	6 hours, 44.5 minutes	62
323	Male	Abnormal	8 hours, 36.5 minutes	55
325	Male	Normal	7 hours, 18.5 minutes	0
326	Female	Abnormal	6 hours, 57 minutes	34
327	Male	Abnormal	6 hours, 51.5 minutes	61
328	Male	Abnormal	7 hours, 41 minutes	31
331	Male	Abnormal	6 hours, 52 minutes	70
333	Male	Normal	6 hours, 39.5 minutes	0
334	Male	Normal	6 hours, 28 minutes	2
337	Male	Normal	5 hours, 37.5 minutes	0
338	Male	Abnormal	5 hours, 35 minutes	89
341	Male	Abnormal	5 hours, 38.5 minutes	32
344	Male	Normal	6 hours, 44.5 minutes	0
349	Male	Normal	6 hours, 1 minute	11
353	Male	Normal	6 hours, 29.5 minutes	0

3.3. Abnormal Respiratory Event Visualization

The specialist doctor defines the abnormal respiratory event as presented in Figure 3.6. Since our raw signals come from sound level, heart rate monitor and accelerometer sensors, the figure explains the arousal of abnormal respiratory event by considering these devices.

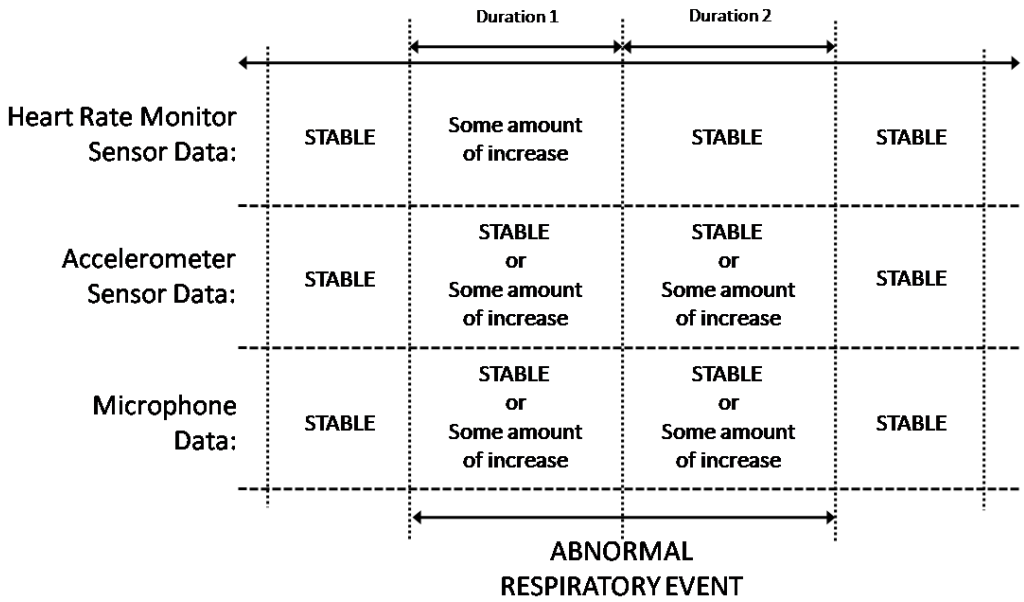
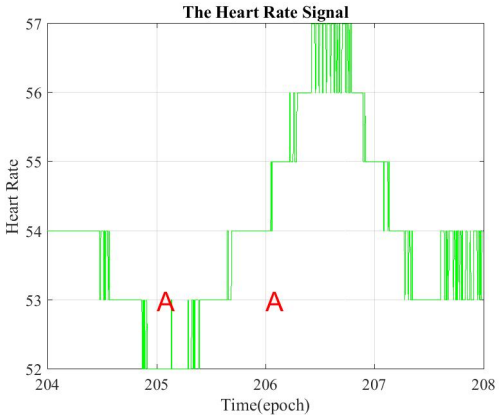


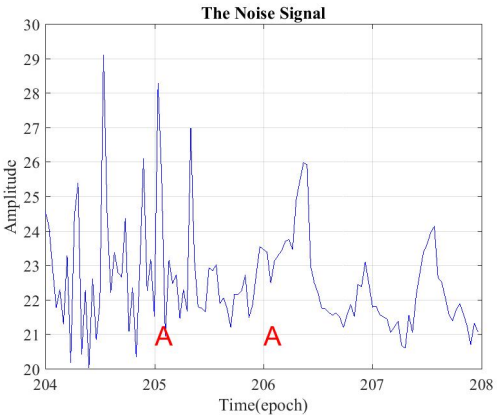
Figure 3.6. Arousal of an Abnormal Respiratory Event.

Initially, the subject has a normal sleeping period which is followed by an increase in the heart rate. This increase occurs because of no breathing or shallow breathing. The body can give response in the same period (Duration 1) or in the next period (Duration 2). Therefore, the accelerometer and noise signals can variate in Duration 1 or Duration 2. The lengths of Duration 1 and Duration 2 change subject by subject and they can be seconds to minutes. There is no proven relationship between the lengths of Duration 1 and Duration 2. By considering this information, the abnormal respiratory events can be observed with the naked eye on raw signals. Figures 3.7(a), 3.7(b), and 3.7(c) present from the heart rate, noise and accelerometer signals, respectively.

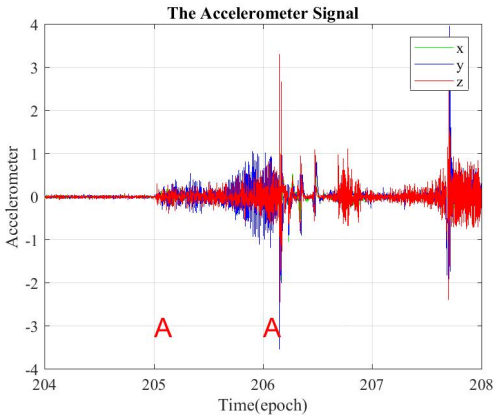
"A" represents the epoch that is labeled as the abnormal respiratory event by the specialist doctor. As described in Figure 3.6, the increase in the heart rate signal is an indicator of the abnormal respiratory event which is followed by the fluctuations in noise and accelerometer signals. These fluctuations generally start in the same epoch with the epoch in which the increase in the heart rate signal occurs and finish in the same or the next epoch. The fluctuations in the noise signal represent the snoring or coughing events.



(a) Abnormal Respiratory Event on Heart Rate Signals



(b) Abnormal Respiratory Event on Noise Signals



(c) Abnormal Respiratory Event on Accelerometer Signals

Figure 3.7. Abnormal Respiratory Event on Raw Signals

3.4. Data Processing Challenges

After the data collection completed, all the data provided by three different sources; smart phone, smart watch and PSG system, are transferred into the same computer. However, some difficulties are encountered and this section explains those difficulties and solution strategies.

The first challenge about this data is synchronization. Since all devices cannot be initialized at the same moment, there is a lack of synchronization among the data. Also, the smart watch always provides less duration of data compared to others because

of the battery limitation. In addition, the raw signals coming from the smart watch have some interruptions since the location and the tightness of the smart watch cannot be kept at the optimal position for the whole night. We combine all the data and identify the overlapping sections by considering the fact that time information comes from different sources. In Table 3.2, the duration of data for each subject is given.

The second challenge that needs to be handled is that the collected data is huge. The data size for each subject is around 1.9 GB and therefore even data preprocessing takes too much time. We cannot process the data coming from each source separately because the algorithms in further steps require their combination. Therefore, we design an efficient algorithm that firstly decides the range of data by considering all time information coming from different sources then prunes every data from each source with respect to this time range. This pruned data is combined and relayed to the further steps in the algorithm.

The last challenge is that the PSG system provides data in an old-fashioned format. Since the utilized PSG is a commercial product, the manufacturer company does not give access to all files. Some files cannot be decoded which cause a problem in the process of extracting labels. Although the software shows exact start time and duration of the abnormal respiratory events; the decodeable file, which has information about every event in each epoch (30 seconds period), has only duration information, does not have the starting moment of the abnormal respiratory event. This situation creates inaccuracies in the label information. Even if an abnormal respiratory event starts at the last second of an epoch, the PSG system marks that specific epoch as abnormal respiratory event epoch not the next one.

3.5. Feature Extraction

The raw signals coming from three different sources; smart phone, smart watch and the PSG system, are firstly divided into 30 seconds windows since the PSG system creates the true labels of respiratory events for 30 seconds long epochs. Different features are derived from each data according to the source type. The raw signals

coming from the accelerometer, heart rate monitor and sound level sensors differ from each other because of their intrinsic properties and totally 103 features are generated for all different sensors. The summary of all extracted features is presented in Table 3.3 and afterwards, the extracted features which are determined by considering the information given in the literature survey is explained in detail.

Table 3.3. 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

- Minimum, maximum, mean, variance, energy and entropy of data provide the general range of data.
- Zero Crossing Rate (ZCR) determines the count of sign change of signal for a period.
- Fast Fourier Transform (FFT), which is a mathematical method, is used to convert the data from time domain to frequency domain and vice versa. The length of FFT vectors depends on the size of window, but we take the first ten dependents of vector.
- Spectral Spread measures the bandwidth of the acoustic spectrum.

- Spectral Centroid represents the center of mass of the signal and it expresses the brightness of signal.
- Spectral Entropy determines the flatness of the acoustic spectrum.
- Spectral Flux represents the stability of the acoustic spectrum.
- Spectral Rolloff states the skewness of the acoustic spectrum.
- Mel Frequency Cepstral Coefficient (MFFC) is a filtering approach, and it converts the power spectrum into frequencies which variate in the human ear's critical bandwidth. MFCC is useful to capture important characteristic of phonetic in speech.
- Harmonic Ratio is a representation of signal's smoothness or dynamic stability.
- Pitch is defined as a major auditory attribute of musical tones.
- Chroma Vector represents the semitones of the musical octave which is the projection of the entire spectrum. We get the first 10 coefficients.
- Median Filter is used to remove destructive noise while preserving edges of signal.
- Hjort Mobility measures the proportion of standard deviation of the power spectrum.
- Hjort Complexity indicates the similarity between the signal and a pure sine wave.
- Skewness measures the lack of symmetry. If the left and right of the center point in a signal look like same, the skewness is calculated as 0.
- Kurtosis determines whether the signal is heavy-tailed or light-tailed relative to the normal distribution. Heavy-tailed signal has high kurtosis.
- Root Mean Square (RMS) measures the loudness of the signal.
- Singular Value Decomposition (SVD) is diagonalization in the basic point of view. If the covariance matrix is diagonal, then it represents a new form for the original data.

3.6. Undersampling

When we analyze all-featured data, the imbalance between the counts of normal and abnormal respiratory events is obvious and this situation may cause poor classification performance. In order to overcome this problem, we eliminate some normal

respiratory events and the decrease in the count of normal respiratory events creates the balance between the counts of normal and abnormal respiratory events. In the literature, the common problem of undersampling strategy is losing the distinctive data. However, this is not a problem for this case since all-featured data contains sufficient number of normal respiratory events.

The undersampling strategy is designed as follows. Since we have less abnormal respiratory events, we do not make any changes on the count of them. We mainly focus on normal respiratory events and among the normal respiratory events, some of them are randomly chosen. The classification results of different algorithms are measured with different number of normal respiratory events. The most accurate result is obtained when the count of normal respiratory events equals to the abnormal respiratory event count.

3.7. Feature Selection

The general objective of the feature selection section is understanding the importance of features when distinguishing the abnormal respiratory events from the normal respiratory events. This importance is determined by considering the role of feature extraction in the classification algorithm. There are several different strategies found in the field. However, we use two common methods; Random Forest and Stepwise Feature Selection. The performance of these methods in this section is measured with the undersampled-featured data.

3.7.1. Random Forest Feature Selection

The Random Forest algorithm will be explained in Section 4.1.2.3. In order to determine the number of decision trees in the forest, we calculate the out-of-bag classification errors for different number of decision trees. The range is started with 1 and ended with 500. This out-of-bag classification error represents the misclassification error of the undersampled-featured data. Figure 3.8 presents the convergence point of the misclassification error and we choose 200 as the number of decision trees in forest.

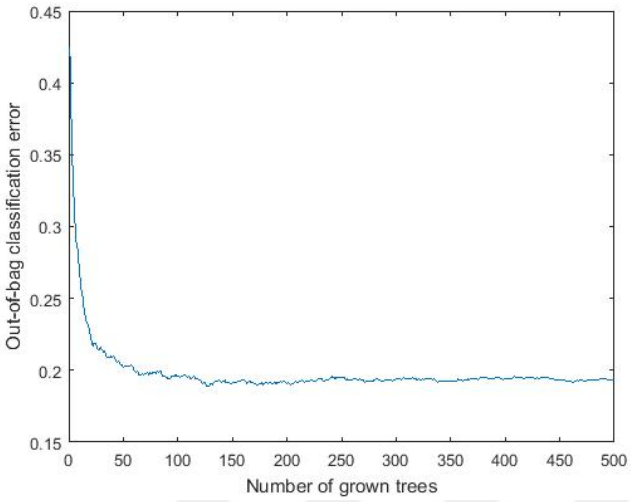


Figure 3.8. Out-of-Bag Classification Errors with the Undersampled-Featured Data for the Random Forest Algorithm.

After the number of decision trees in the forest is determined, the importance of features are estimated. This is handled by calculating the distance between the predicted probability for the true and wrong classes. If the distance is larger, then the feature is more important since it differentiate the classes more precisely. The importance estimations for all extracted features are given in Figure 3.9. The first 46 features are derived from the accelerometer signals and the following 35 features are generated from the noise signals. The rest of features are obtained from the heart rate signals.

When we analyze the importance estimations, we understand that most of the extracted features have very low importance. Figure 3.9 proves that the features extracted from the accelerometer signals, the first 46 features, are not as important as the features extracted from the noise signals, from 47 to 81. Among the features extracted from the heart rate signal, the last 22 features, some can be directly eliminated since the importance is very close to zero and some of them are as important as the features generated from the noise signals.

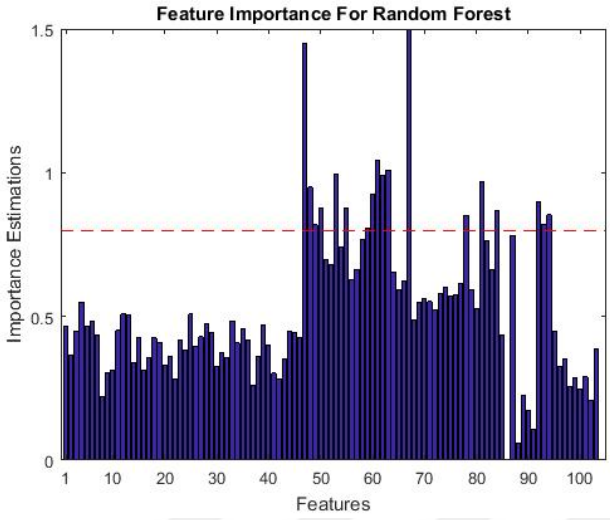


Figure 3.9. Importance Estimations of All Features for the Random Forest Algorithm.

In order to select the most important features, we try to find a threshold and the possible values of the threshold is chosen between the mean of the importance estimations of features to the maximum of the importance estimations of features. However, when we compare the misclassification errors generated with selected-featured data, 0.8 gives the best outcome.

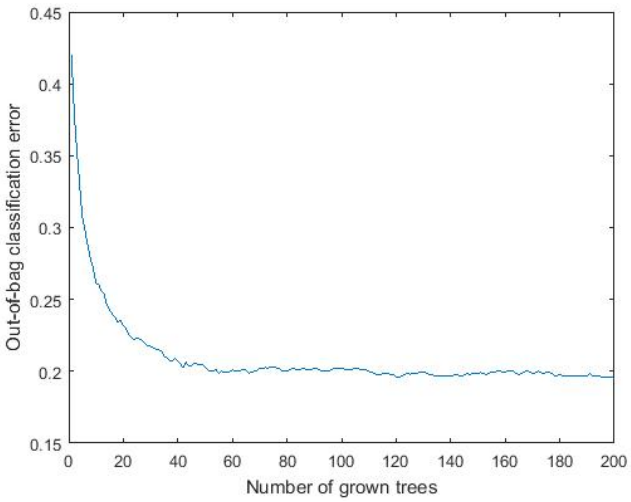


Figure 3.10. Out-of-Bag Classification Errors with Selected-Featured Data for the Random Forest Algorithm.

The comparison of Figure 3.8 and Figure 3.10 proves that the misclassification error for the undersampled-featured data and the selected-featured data with the determined threshold is very similar to each other. With the determined threshold, 18 features are selected and those are generated from noise and heart rate signals. The selected features with their importance estimations can be seen in Figure 3.11. The features extracted from the accelerometer signals are not chosen. Since we activate the accelerometer sensor on the smart watch and the wrist is one of the most restless body parts, the movements of wrist can occur in both abnormal and normal respiratory event periods. Also, the body response of the abnormal respiratory event may not be clearly observable with wrist movements. Because of those reasons, the features extracted from the accelerometer signals are not good candidates for distinguishing the normal and abnormal respiratory events.

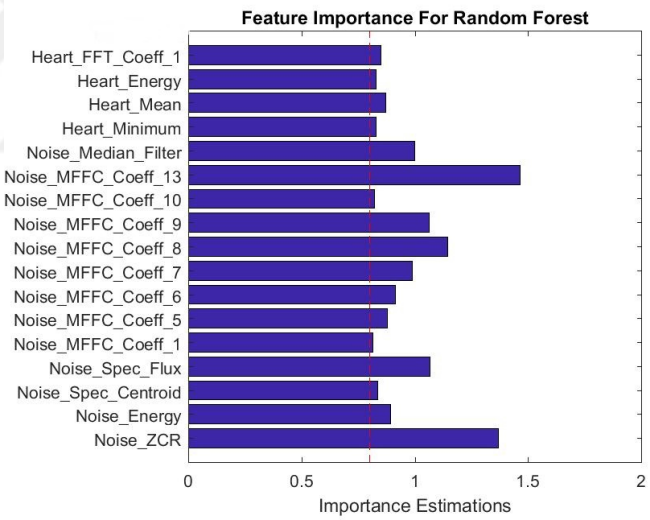


Figure 3.11. Selected Features with Importance Estimations.

3.7.2. Stepwise Feature Selection

The stepwise feature selection is a well-known algorithm in which all extracted features in the dataset are evaluated one-by-one. This evaluation can be handled in two directions which are forward and backward. In forward stepwise feature selection, the algorithm starts with a null subset of features and includes the features one-by-one

in each step by evaluating the importance of each feature in the dataset. On the other hand, in the backward stepwise feature selection; the initial subset contains all features in the dataset and it excludes the features one-by-one in each step by again evaluating the importance of each feature in the dataset. Both algorithms stop when two steps contain the same subset of features. The inclusion and exclusion operation depend on a threshold which is determined by the user [83].

We use the forward and backward approaches together. At the beginning, we define two different thresholds; one of them is determined for inclusion and other one is determined for exclusion. 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 field. The algorithm starts with the null subset of the all extracted features and includes a feature according to the inclusion threshold. When including the next feature, the algorithm computes the importance of already included features and if the exclusion threshold is reached for any feature, this feature will be excluded. At the end, the features which hold both; the inclusion and exclusion, thresholds exist in the subset.

In our implementation, each trial gives different subsets of all features since the first included feature is determined randomly. Therefore, we run this algorithm for 10 times and choose the features exist in at least 6 different subsets.

The average p-values for the selected features are given in Table 3.4 and the importance of features is decreased with the greater p-value. When we analyze the selected features with the stepwise feature selection algorithm, the features extracted from the noise signals have a significant majority among all selected features. The body response of an abnormal respiratory event can be observed with the features extracted from the accelerometer signal. This method ignores the fact that the features extracted from the heart rate signal are important.

Table 3.4. Selected Features with Stepwise Feature Selection and Their 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

3.7.3. Discussion

When we compare the selected features, we observe that some of the selected features are common for both methods. 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 occured after the abnormal respiratory event. The importance of features extracted from the heart rate signal and the accelerometer signal is quite low. This has several reasons. The first one is explained in Section 3.7.1. The body response of the abnormal respiratory event may not be observed from the accelerometer signals recorded on the wrist. The second reason is that the recording of noise signals is much more reliable than the recording of heart rate and accelerometer signals. The location and tightness of smart watch have a non-negligible effect on the interruptions in data collection.

4. SYSTEM PERFORMANCE AND PERFORMANCE ENHANCEMENTS

The system performance can be measured in two different ways. The first way evaluates the success rate of distinguishing abnormal respiratory events from normal respiratory events and the second way classifies subjects according to the problems in their respiration and analyzes the results. Therefore, this chapter firstly evaluates the performance of respiratory event detection afterwards some improvements are implemented for generating more promising results. Lastly, the predictions about the subjects according to their respiration problems are analyzed.

The performance metrics used to measure the success rate of the system is Accuracy, Precision, Sensitivity and Specificity. The accuracy represents the true amount of predictions in the sample dataset whereas the precision measures the reliability of system in different trials. The sensitivity and specificity represent whether the samples have specific property. If a person is tested about a disease, the test is highly sensitive when he has the disease and the test is highly specific when he does not have the disease [84]. Those metrics are given with the Confusion Matrix that represents the information about the actual and the predicted numbers of samples obtained with classification algorithms. By considering the definitions of all performance metrics; the sensitivity, which mainly represents the abnormality in respiratory events for this thesis, becomes more important. Therefore, when analyzing the system performance, all performance metrics should be in balance and promising.

4.1. Respiratory Event Detection System Performance Evaluation

In order to evaluate the respiratory event detection system performance, we follow different approaches by considering the nature of data and also the literature survey. The main objective of these different approaches is distinguishing normal respiratory events from abnormal respiratory events.

While measuring the system performance, we compare the results of different machine learning algorithms with different subsets of the data coming from three different sources; smart phone, smart watch and PSG system. These different subsets are generated from both raw signals and featured data and each subset is explained in each section in detail.

The classification results of raw signals are generated with Hidden Markov Model (HMM). On the other hand; Decision Tree (DT), K-Nearest Neighbors (KNN), Random Forest (RF) and Support Vector Machine (SVM) algorithms generate the classification results with different subsets of featured data. In the end, the results of binary classification algorithms with the undersampled-featured data are analyzed with Receiver Operating Characteristic (ROC) curve which is mainly designed to compare the performances of different algorithms.

4.1.1. Classification Results with Raw Signals

In this section, we are trying to differentiate the abnormal respiratory event from the normal respiratory event on raw signals coming from three different sources with the help of the Hidden Markov Model approach since the abnormal respiratory event can be observed with the naked eye on raw signals and the Hidden Markov Model approach requires time series of data.

Hidden Markov Model is an unsupervised machine learning technique and it is a statistical sequence model in which each unit in the sequence has a label. The best label sequence is chosen by considering the probability distribution over possible sequences of labels. Therefore each unit affects the next unit's label [61].

As stated before, we implemented the Hidden Markov Model approach in Python. For this approach, we combine all subjects' data and the combined raw data are divided into sequences with length 30 seconds. We generate different Hidden Markov Models for abnormal and also normal respiratory event sequences. Furthermore, each source; sound level, heart rate monitor and accelerometer sensors, have different mod-

els. Therefore, we have totally six HMMs. The sequence of the same duration from each signal is given to the related models. Figure 4.1 represents explained HMMs clearly. We can also develop another system that contains only two HMMs; abnormal and normal. However, since we already combine all subjects' data, combining more data coming from diffent sources makes the problem more complex.

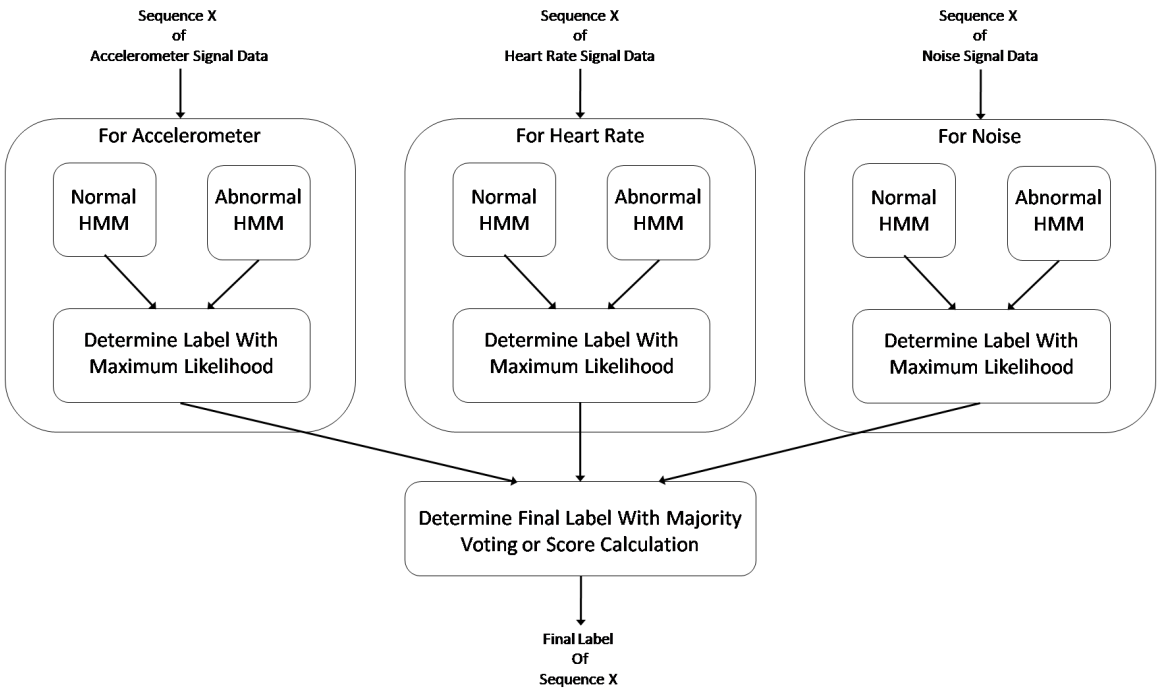


Figure 4.1. Determination Process of Final Label for the Hidden Markov Model Algorithm.

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 score calculation approach pays attention to details. The determination process of the final label is shown in Figure 4.1.

The results of Hidden Markov Model for two different methods can be seen in Tables 4.1 and 4.2. The results prove that the imbalance between the counts of normal and abnormal respiratory events decreases the sensitivity. Even if the accuracy is acceptable, the classification performance of the abnormal respiratory events is not satisfying. On the other hand, the system classifies the abnormal respiratory events as normal respiratory events with this algorithm and therefore, the raw signals are processed and the features are extracted for generating more promising results.

Table 4.1. Classification Results of the Hidden Markov Model Algorithm with Score Calculation.

	Predicted Class		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	2377	363	75.13	31.51	25.85	86.75
Abnormal	479	167				

Table 4.2. Classification Results of the Hidden Markov Model Algorithm with Majority Voting.

	Predicted Class		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	2290	450	73.69	31.30	31.73	83.58
Abnormal	441	205				

4.1.2. Classification Results with All-Featured Data

After the feature extraction is completed, a combination of data that contains the subjects with abnormal and normal respiratory events is prepared. All the algorithms in this section generate results with this combination of data. The true labels of this data are extracted from the PSG system which are manually given by the specialist

doctor. Each algorithm in this section is trained with two different approaches. The first one, Percentage, requires a part of all-featured data. We split all-featured data 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 data.

The second one, K-Fold Cross Validation, requires all-featured data and 10 Folds are generated for this thesis. We have enough data for 10-Fold Cross Validation since we combine all subjects' data. Each fold preserves the ratio between abnormal and normal respiratory event counts. Therefore, each training or testing fold is a small representation of all-featured 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.

4.1.2.1. Decision Tree. The Decision Tree algorithm is a supervised machine learning method that is simply trying to classify the data by creating a tree structure. In this structure the leaf nodes contain the samples of data and the internal nodes represent the decision mechanism. From the root node to leaf nodes, the classification rules can be tracked. Since decision trees are easy to understand and interpret, they are widely used in the field [57]. The results of the Decision Tree algorithm for two different training strategies can be seen in Tables 4.3 and 4.4.

Table 4.3. Classification Results of the Decision Tree Algorithm with K-Fold Cross Validation.

	Predicted Class		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	8080	2854	66.63	24.22	35.61	73.90
Abnormal	1649	912				

Table 4.4. Classification Results of the Decision Tree Algorithm with Percentage.

	Predicted Class		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	2287	440	74.82	34.33	36.11	83.87
Abnormal	407	230				

4.1.2.2. K-Nearest Neighbors. The K-Nearest Neighbors algorithm is mainly measuring the distance of a sample to its neighbors. The distance can be calculated with Euclidean Distance, Hamming Distance or other formulas and it represents the similarity of that sample to each neighbor [85]. We try different distances and in the end, we prefer to use the Cosine Distance which generates more accurate results with great number of features. The neighbor count is selected as 26 for the K-Fold Cross Validation and 33 for the Percentage approach. The reason of these choices is that we try different neighbor counts and the results are converged around 30 neighbors. The chosen neighbor numbers yield the best results. The results of the K-Nearest Neighbors algorithm for two different training strategies can be seen in Tables 4.5 and 4.6.

Table 4.5. Classification Results of the K-Nearest Neighbors Algorithm with K-Fold Cross Validation.

	Predicted Class		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	9392	1542	75.55	25.72	20.85	85.90
Abnormal	2027	534				

Table 4.6. Classification Results of the K-Nearest Neighbors Algorithm with Percentage.

	Predicted Class		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	2613	114	81.42	52.50	19.78	95.82
Abnormal	511	126				

4.1.2.3. Random Forest. The Random Forest algorithm is again a supervised machine learning method and it is an advanced version of the decision tree. This algorithm contains more than one decision tree. Each tree in the forest makes its own classification and the final decision is determined by calculating to the most voted class [62]. We choose 200 decision trees for this algorithm and the determination process of the count of decision trees is explained in Section 3.7.1. The results of the Random Forest algorithm for two different training strategies can be seen in Tables 4.7 and 4.8.

Table 4.7. Classification Results of the Random Forest Algorithm with K-Fold Cross Validation.

	Predicted Class		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	9046	1888	70.59	20.27	18.74	82.73
Abnormal	2081	480				

Table 4.8. Classification Results of the Random Forest Algorithm with Percentage.

	Predicted Class		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	2649	78	82.28	60.41	18.68	97.14
Abnormal	518	119				

4.1.2.4. Support Vector Machine. The Support Vector Machine algorithm is a supervised machine learning technique and it mainly focuses on finding hyperplanes or decision boundaries that best divide the data by considering features. The critical data points affect the position of hyperplane and the main objective is to place the hyperplane with the greatest margin to every data point [54]. When placing the boundaries, Kernel functions are utilized and we use the linear kernel since we need to find only one boundary that clearly distinguishes the abnormal respiratory events from the normal respiratory events. The results of the Support Vector Machine algorithm for two different training strategies can be seen in Tables 4.9 and 4.10.

Table 4.9. Classification Results of the Support Vector Machine Algorithm with K-Fold Cross Validation.

	Predicted Class		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	9905	1029	75.77	23.72	12.50	90.59
Abnormal	2241	320				

Table 4.10. Classification Results of the Support Vector Machine Algorithm with Percentage.

	Predicted Class		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	2651	76	82.49	62.00	19.47	97.21
Abnormal	513	124				

4.1.2.5. Discussion. When we analyze the results of all implemented algorithms with all-featured data, the imbalance between the counts of normal and abnormal respiratory events decreases the sensitivity. This situation is very similar with the Hidden Markov Model results. Although the accuracy is desirable, the system gets confused when detecting abnormal respiratory events. For each algorithm, the Percentage training approach gives more accurate results because the abnormal respiratory event count is greater for the K-Fold Cross Validation and the greater number of falsely classified abnormal respiratory event decreases the accuracy more dramatically.

4.1.3. Classification Results with Undersampled-Featured Data

The classification results of raw signals and all-featured data with different classification algorithms prove that the imbalance between the counts of normal and abnormal respiratory events causes poor classification performance. In order to overcome this problem, we try to balance the counts of normal and abnormal respiratory events by undersampling.

The classification results of the DT, KNN, RF and SVM algorithms for two different training strategies can be seen in Table A.1 and Table A.2 in Appendix A. In Table 4.11, the performance comparison for the DT, KNN, RF and SVM algorithms with two different datasets is given. The classification results with the all-featured data are obtained with the Percentage training approach whereas the classification results

with the undersampled-featured data are generated with the K-Fold Cross Validation training approach since these results are more accurate compared to the other training strategy. The expectation is that the results of the K-Fold Cross Validation outperforms 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 normal and abnormal respiratory events for all-featured data causes this difference. The best result is indicated with bold in the table.

Table 4.11. Performance Comparison between the All-Featured Data and the Undersampled-Featured Data.

	All-Featured Data		**Undersampled-Featured Data**	
Classification	**Accuracy**	**Sensitivity**	**Accuracy**	**Sensitivity**
DT	74.82	36.11	72.06	70.95
KNN	73.57	35.01	71.22	70.21
RF	82.28	18.68	80.63*	88.91*
SVM	82.49	19.47	75.50	85.44

The performance comparison proves that even if the accuracy decreases slightly, the sensitivity increases significantly for each algorithm. This indicates that the abnormal respiratory events are no longer classified as the normal respiratory event and the performance of the system is improved with the undersampling approach. Therefore, the following sections in this thesis utilize the undersampled-featured data in training and testing phases.

4.1.4. Classification with Selected-Featured Data

The undersampled-featured data is pruned according to selected features from the Random Forest algorithm and the Stepwise algorithm and we obtain two different datasets. With these different datasets, the algorithms are rerun.

The classification results of the DT, KNN, RF and SVM algorithms for two different training strategies and the selected-featured dataset generated with the Random Forest algorithm can be seen in Table A.3 and Table A.4 in Appendix A. Furthermore, the classification results of the DT, KNN, RF and SVM algorithms for two different training strategies and the selected-featured dataset generated with the Stepwise algorithm can be seen in Table A.5 and Table A.6 in Appendix A.

In Table 4.12, the performance comparison for the DT, KNN, RF and SVM algorithms with three different datasets is given. The classification results with the undersampled-featured data and both selected-featured data are obtained with the K-Fold Cross Validation training approach since the results are more accurate compared to the percentage training approach since it considers whole data. The best result is indicated with bold in the table.

Table 4.12. Performance Comparison between the Undersampled-Featured Data, the Random Forest Algorithm Selected-Featured Data and the Stepwise Algorithm Selected-Featured Data.

	Undersampled-Featured		**Random Forest-Selected**		**Stepwise-Selected**	
Classification	**Accuracy**	**Sensitivity**	**Accuracy**	**Sensitivity**	**Accuracy**	**Sensitivity**
DT	72.06	70.95	73.70	73.37	72.63	71.73
KNN	71.22	70.21	70.81	73.14	69.99	73.25
RF	80.63	88.91	80.96*	88.99*	80.30	88.56
SVM	75.50	85.44	73.33	75.17	75.61	86.18

The performance comparison proves that each algorithm behaves differently. For the DT algorithm, both the accuracy and sensitivity increase slightly, whereas for the KNN algorithm, even if the sensitivity increases, the accuracy decreases. The feature selection does not make any significant difference in the classification results of the RF algorithm and the features selected with the stepwise algorithm outperforms the results of the features selected with random forest for the case of SVM algorithm. These outcomes indicate that the features not selected with the random forest and stepwise algorithms are also important for the system performance. This is because

the proposed system design needs every single response of the body. In the light of this information, the ROC curve analysis utilizes the undersampled-featured data.

4.1.5. ROC Curve Analysis with Undersampled-Featured Data

Receiver Operating Characteristic (ROC) curve is mainly used to measure the performances of different algorithms. It compares the strengths of the algorithms by plotting true positive rate that is (1-Sensitivity) against the false positive rate which is (1 − Specificity). 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 best in every situation [86].

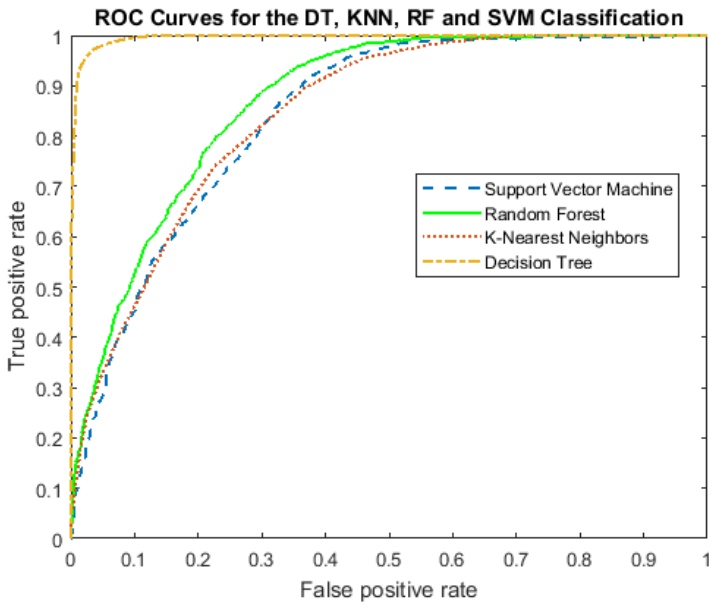


Figure 4.2. ROC Curves for the DT, KNN, RF and SVM Algorithms.

We prepared the ROC curves in the MATLAB environment which generates different thresholds for each model. We prefer to utilize the undersampled-featured data and the ROC curves created for comparing the performances of DT, KNN, RF and SVM algorithms are found in Figure 4.2.

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.

4.1.6. Discussion

All in all, the results prove that the main problem about the classification algorithms with raw signals and all-featured data is the imbalance between 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 this imbalance affects the classification results, especially sensitivity, negatively and the abnormal respiratory events are classified as normal respiratory events.

In order to solve this problem, we undersample with the all-featured data and this approach improves our system performance remarkably. We get promising accuracy and sensitivity outcomes. We also implement undersampling with the raw signals however we obtain very poor classification results. Since the sequence count of raw signals is not enough for undersampling, we eliminate some distinctive sequences from the raw signals. If the system contains more abnormal respiratory events or more subjects, we may observe more improvement in the system performance. The same situation is hold for all-featured data since the undersampling rate of all-featured data is changed accordingly if the system contains more abnormal respiratory events or more subjects.

For generating more accurate results, we employed different feature selection algorithms however, our feature selection algorithms do not make any significant difference in the system performance. This is because during the data collection there are some

interruptions because of the location and tightness of the smart watch and therefore the system needs every single feature to compensate data interruptions and to measure every single response of the body.

On the other hand, the algorithms are trained with all collected data without discrimination between the subjects. We may analyze the classification results of each subject data which is more preferable since the abnormal respiratory events may occur in a different way for each person. However we have to ignore this part since anyone among the subjects is participated in the experiments for two nights and one night data is not enough for training and testing the algorithm. If the subject just performs normal respiratory events, making classification does not make sense. Furthermore, if the subject is labeled as abnormal, this means that the subject performs abnormal respiratory events, the imbalance between the normal and abnormal respiratory events again causes poor classification results.

The summary of all implemented algorithms can be found in Table 4.13. These results are chosen by considering the most accurate results for each algorithm. For the raw signals, the Hidden Markov Model 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 and for the undersampled-featured and selected-featured data, we take the results of the K-Fold Cross Validation approach for each algorithm. The best result is indicated with bold in the table.

Table 4.13. Performance Comparison of All Implemented Algorithms.

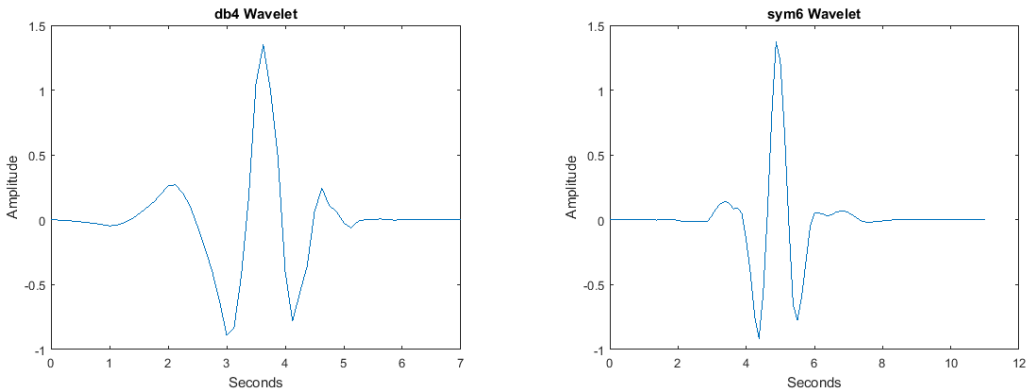
	Result (%)			
Classification	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
HMM, Raw Signals	73.69	31.30	31.73	83.58
DT, All-Featured Data	74.82	34.33	36.11	83.87
KNN, All-Featured Data	81.42	52.50	19.78	95.82
RF, All-Featured Data	82.28	60.41	18.68	97.17
SVM, All-Featured Data	82.49	62.00	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

4.2. Respiratory Event Detection System Performance Enhancements

In order to increase the success rate of the proposed system, the problems of the design are investigated. In this chapter, those problems are explained in detail and the improved results are presented.

4.2.1. Filtering Raw Signals

In our proposed system, the data coming from the sound level, heart rate monitor and accelerometer sensors are time series and this data is used to generate results with the Hidden Markov Model. However, in the literature, the researchers who try to find specific behavior, action or event in the raw signal prefer to use filtering. Since the heart rate signal is the main source of detecting abnormal respiratory events in our proposed system, we research about specific filtering techniques. In the field, the Discrete Wavelet Transform (DWT) is preferred with the Hidden Markov Model approach and there are several studies on the heart rate signal. With these two algorithms 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 and Symlets 6 when extracting apnea from the heart rate signal with this method [87–93]. Figure 4.3 presents those wavelets.



(a) Daubechies 4. (b) Symlets 6.

Figure 4.3. Common Mother Wavelets [3]

.

The Hidden Markov Model described in Section 4.1.1 is followed for this section also. The only difference is that the heart rate signals are preprocessed with Daubechies 4 (db4) and Symlets 6 (sym6) wavelets. The results can be seen in Table 4.14 and 4.15 and the best results are indicated with bold in the tables.

Table 4.14. Performance Comparison between Raw And Processed Heart Rate Signal for the Hidden Markov Model Algorithm with Score Calculation.

	Result (%)			
Classification	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Raw Signals	73.13	31.51	25.85	86.75
Processed With db4	77.17	36.58	26.78	89.05
Processed With sym6	77.38*	27.70	28.48*	88.91

Table 4.15. Performance Comparison Between Raw and Processed Heart Rate Signal for the Hidden Markov Model Algorithm with Majority Voting.

	Result (%)			
Classification	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Raw Signals	73.69	31.30	31.73	83.58
Processed With db4	74.63	28.31	21.52	87.15
Processed With sym6	74.84*	29.07	22.14	87.26

When we analyze the results, we can see that the results become more accurate as expected. 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 generalization. Also, our mother wavelet choice may not be appropriate. However, the most important reason is the imbalance between the normal and abnormal respiratory events. In order to overcome this problem, we implement undersampling with the raw signals and analyze the system performance. However, the improvement in the system performance is not significant. While undersampling, we may eliminate some distinctive sequences from the raw signals. If raw signals contain more abnormal

respiratory event data and the counts of normal and abnormal respiratory events are more balanced, the improvement may be more obvious.

4.2.2. Including Pulse Oximeter Signals

The abnormal respiratory events are detected with the help of several sensors and this is explained in Chapter 2. In our proposed system, we utilize the sound level, heart rate monitor and accelerometer sensors. At the beginning, we consider to add an external pulse oximeter however since the PSG system also contains a pulse oximeter, we do not want to decrease the obtrusiveness of the system and comfort of subjects. This section investigates an improved system design. For just testing purposes, we extract the pulse oximeter data from the PSG results and this eliminates the concerns about the reliability of data. The pulse oximeter data are time series and therefore, we prefer to combine this data with the raw signals that are used for the Hidden Markov Model classification. If we include a pulse oximeter in our proposed system, we obtain the results in Table 4.16. The better result is indicated with bold in the table.

Table 4.16. Performance Comparison between Current System Design and Included Pulse Oximeter Signal for the Hidden Markov Model Algorithm with Majority Voting.

	Result (%)			
Classification	**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

The results are generated for just majority voting and we see that the accuracy is improved whereas the sensitivity is decreased. One reason of 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 and pulse oximeter data. Other reason is that we cannot observe the significant improvement because of the imbalance between the counts of normal and

abnormal respiratory events. For dealing with this problem, we undersample the raw signals and measure the system performance one more time. Since we eliminate some distinctive sequences from the raw signals, there is no remarkable improvement in the system performance. If more subjects participate the experiments and the data is balanced, we may generate better system performance.

4.2.3. Oversampling Abnormal Respiratory Events

In the collected data, the abnormal respiratory events are very rare compared to normal respiratory events. Since we do not have a chance to collect more data, we try to oversample the abnormal respiratory events with Synthetic Minority Over-sampling Technique (SMOTE) library [94]. This library is implemented for imbalanced datasets and it has several different options like undersampling. We apply this to only the Random Forest algorithm since one of the best system performance is obtained with this algorithm. The Percentage training approach is applied and the results are shown in Table 4.17. The better result is indicated with bold in the table.

Table 4.17. Performance Comparison between the All-Featured Data and the Oversampled-Featured Data for the Random Forest Algorithm with Percentage.

	Result (%)			
Classification	**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

For just testing purposes we do not equalize the counts of abnormal and normal respiratory events in the algorithm. There are two reasons behind this choice. The first one is that oversampling is actually creating non-real data and the reliability of this data is very low. The other reason is that if we equalize the counts, we obtain a dataset that multiplies the count of normal respiratory events with two and processing this much of data takes too much time.

Table 4.17 proves that even if we oversample the abnormal respiratory events slightly, the system performance is improved. Making remarkable oversampling generates more accurate classification results.

4.2.4. Generating Extended Label Set

The labels provided by the PSG system are restricted. This has two reasons. The first one is explained in Section 3.4. Since the decodeable file includes only the duration information not the starting moment of abnormal respiratory event, the label of an epoch (30 seconds period) can be incorrect. The second reason is explained in Figure 3.6, the body can give response after the occurrence of abnormal respiratory event.

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

With this new label set, the Hidden Markov Model results with the raw signals are calculated with Majority Voting. The rest of the classification algorithms; the DT, KNN, RF and SVM algorithms, are repeated for featured data. The Percentage training approach is prefered for all-featured data whereas the K-Fold Cross Validation training approach is used for undersampled-featured data. The reason behind these choices is that they yield the best results. The classification results are shown in Table 4.18 and the best result is indicated with bold in the table.

Table 4.18. Performance Comparison between Original Labels and Extended Labels.

	Original Labels				Extended Labels			
Classification	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
HMM, Raw Signals	73.69	31.30	31.73	83.58	71.90	52.95	55.76	78.80
DT, All-Featured Data	74.82	34.33	36.11	83.87	72.01	52.99	55.52	79.03
KNN, All-Featured Data	81.42	52.50	19.78	95.82	73.05	56.75	41.00	86.69
RF, All-Featured Data	82.28	60.41	18.68	97.17	77.77	66.58	51.34	89.03
SVM, All-Featured Data	82.49	62.00	19.47	97.21	78.10	66.79	53.03	88.77
DT, Undersampled-Featured Data	72.06	72.56	70.95	73.17	77.96	78.31	77.34	78.58
KNN, Undersampled-Featured Data	73.57	69.91	82.74	69.34	78.28	75.18	84.44	72.13
RF, Undersampled-Featured Data	80.63	76.28	88.91	72.35	85.95*	81.76*	92.55*	79.35*
SVM, Undersampled-Featured Data	75.50	71.27	85.44	65.56	79.68	74.88	89.35	70.02

When we analyze the results, we observe that, for raw signals and all-featured data the slight decrease in the accuracy is compensated with remarkable increase in the sensitivity. For the undersampled-featured data, the system performance is improved significantly and the accuracy obtained with the RF algorithm, 85.95%, is the best accuracy we achieve. These results prove that if the label set is more accurate and we collect more abnormal respiratory events, the system distinguishes the abnormal respiratory events better.

4.2.5. Discussion

The enhancements on the proposed system design show that the current system design is open to improvements. The main problem about the improved results is the imbalance between the normal and abnormal respiratory events. If the system can be trained with more abnormal respiratory events and the label set is more accurate, the difference between the results of the current system design and improved system design is more noticeable.

4.3. Prescreening System Performance Evaluation

The main objective of the proposed design is to recognize people with severe respiratory events and give suggestions like "'consult a doctor". Until this section, we try to distinguish the abnormal respiratory events from the normal respiratory events. In order to make this distinction, we combine all subjects' data. However, in this section, we analyze each subject's situation in itself.

The subjects participated in this study are examined by the specialist doctor and the subjects are separated according to their true AHI score which is generated by the PSG system. The "Normal" subjects have no or a few abnormal respiratory events whereas the "Abnormal" subjects have lots of abnormal respiratory events during their sleep. The subjects are divided into 3 groups according to their true AHI score in the literature and we use this decision mechanism when classifying subjects as "Normal" and "Abnormal".

By considering this information, we try to generate a new AHI score which is predicted by several algorithms. This predicted score is an indicator of the situation of subjects measured by our proposed system design and the predicted "Normal" and "Abnormal" labels of the subjects are determined with the same mechanism mentioned above.

We analyze each subject's data 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 algorithm training. This procedure is repeated for each subject.

The predicted labels are generated for each algorithm and a subject who is classified as "Abnormal" in at least one algorithm is labeled as "Abnormal". Since this

system is designed for recognizing people with severe respiratory problems, we prefer this cautious approach. The results of this procedure can be seen in Table 4.19. In the table, the subjects presented with "1" are classified as abnormal whereas "0" indicates that the subject is classified as normal.

Table 4.19. 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

Table 4.19 proves that only one subject, 326, is classified wrongly. The implemented algorithms cannot detect the abnormality in this subject. One reason of 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".

Table 4.19 indicates the performance of system as a prescreening tool. According to the classification, the Decision Tree algorithm gives the same output with the combination of all algorithms. Therefore, we can eliminate other algorithms and continue with just DT algorithm since it is sufficient. The accuracy of the system is 94.12% and the system contains just 17 subjects' data. If more subjects participate in the experiments and the system is trained with more abnormal respiratory events, we may classify every subject correctly. All in all, by considering the obtained results one can say that this unobtrusive, cheap and reachable system is preferable with respect to other monitoring tools.

5. CONCLUSION AND FUTURE WORK

In this thesis, we proposed a prescreening tool design that recognizes the problems in respiration during sleep. With the help of this information, the system analyzes the health condition of subjects and gives suggestions like "consult a doctor". The system is utilizing smart wearables. Several advantages and disadvantages of smart wearables are discussed throughout this thesis and the most unobtrusive, cheap and reachable system design is formed.

The two objectives of this study can be defined as "Distinguishing abnormal respiratory events from normal respiratory events" and "Classifying subjects according to the problems in their respiration". For this reason, we generated different combinations of collected data. These combinations of data are obtained with the help of different feature extraction, feature selection and sampling approaches. Later, we employed five different machine learning algorithms in order to differentiate abnormal respiratory events. Two different training methodologies and two different score generation strategies are utilized when achieving the results of the classification algorithms. The performance of the algorithms is evaluated by calculating accuracy, precision, sensitivity and specificity. Each generated result is discussed and the reasons behind the success and failure rates of the algorithms are explained. The drawbacks of the proposed system are investigated and some enhancements are suggested. The results of realized enhancements are presented. In the end, the performance of the system is measured as a prescreening tool and the severity of problems in each subject's respiration is analyzed.

This thesis provides some contributions to the field. These contributions can be analyzed in four main topics. Firstly, this thesis has an unobtrusive and comfortable design with respect to other systems in the literature. The smart wearables solve the problems experienced because of overwhelming design of other monitoring tools. Also, the comfort of subjects is looked out and the monitoring does not force subjects to obey any set of rules about the sleeping place, sleeping position. This design also eliminates

the need of a professional which requires huge amount of time and cost. Secondly, the system does not permit outside intervention which provides objectiveness. The manipulation or imitation of behavior does not possible for all night long. Therefore, the system provides an opportunity to analyze the real conditions of subjects. The third contribution is about privacy. The system does not permit to violate privacy by collecting raw signals. While sleeping, subjects are unconscious and the concerns about this situation are resolved with this prescreening tool. The last contribution is that our system is a beneficial tool which differentiates the abnormal respiratory events from the normal respiratory events with the best achieved accuracy rate 85.95%. Furthermore, it identifies the severity of problems in respiration for each subject. Only one misclassification among 17 subjects is a promising result when we consider the success rates of studies mentioned in Section 2.1.

This thesis proves that an unobtrusive, cheap and reachable prescreening tool can be designed to ease the life subjects who have respiratory problems. Although this prescreening tool is a promising alternative, it is currently not practical and easy to integrate in real life and there is still room for improvement. The future works should focus on the weaknesses of the proposed system design in this thesis. Since the location and tightness of the smart watch has a non-negligible impact on the interruptions in data collection, the feature selection approaches in this thesis generally choose noise features and therefore, one can say that the sound level sensor is enough for data collection. However, the smart watch is the most unobtrusive option for heart rate monitoring in today's technology and therefore eliminating smart watch is not a good alternative. In order to catch every response of body and increasing battery life of smart watch, the accelerometer sensor is deactivated on the smart watch and an external accelerometer sensor can be placed at somewhere else on the body, i.e. abdomen. With the help of developing technology, in a short while, the interruptions in data coming from smart watches will be solved and also the battery life of smart watches will be increased. In a future design, an external pulse oximeter data can be added which provides more information about the arousal of an abnormal respiratory event. Furthermore, the label set that we extract from the PSG system has some missing information and the future design should eliminate this problem. 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. However, the collected data for testing this system is imbalanced. As a future work, more data should be collected by considering the balance between the abnormal and normal respiratory events. On the other hand, since sleeping is a long activity, the data collection process can focus on the two hours from the middle of sleep which represents the sleeping activity of whole night. This decreases the data size and also speeds up the result generation process. Furthermore, some subjects should participate in the experiments more than one night. Our system is tested with only real patients and therefore the healthy subjects should be encouraged to participate in the experiments which increases the diversity of data.

REFERENCES

1. Boswell, L., *Polysomnography - Sleep Apnoea*, `http://bozwell.co.uk/poly.html`, accessed at July 2017.

2. Sleep-Apnea-Guide, *The Polysomnogram Test*, `http://www.sleep-apnea-guide.com/polysomnogram.html`, accessed at July 2017.

3. Wasilewski, F., *Wavelet Properties*, `http://wavelets.pybytes.com/`, accessed at July 2017.

4. WHO, *Technical Meeting on Sleep and Health*, `http://www.who.int/en`, accessed at July 2017.

5. Pona, A. A., *The impact of sleep quality on energy intake, eating behavior, and physical activity*, University of Missouri-Kansas City, 2015.

6. Carskadon, M. A., "Patterns of sleep and sleepiness in adolescents", *Pediatrician*, Vol. 17, No. 1, pp. 5–12, 1990.

7. Bano, M., 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", *Frontiers in neurology*, Vol. 5, 2014.

8. Beccuti, G. and S. Pannain, "Sleep and obesity", *Current opinion in clinical nutrition and metabolic care*, Vol. 14, No. 4, p. 402, 2011.

9. Menza, M. A. and R. C. Rosen, "Sleep in Parkinson's disease: the role of depression and anxiety", *Psychosomatics*, Vol. 36, No. 3, pp. 262–266, 1995.

10. Havermans, J. R., *The Limits of a Reductionist Approach to Improving Quality of*

Sleep through Smartphone Applications, Master's Thesis, University of Amsterdam, the Netherlands, 2014.

11. Haakma, R. and R. J. Beun, "Unobtrusive Sleep Monitoring", *Measuring Behavior*, Vol. 2012, p. 122, 2012.

12. Stanford-Medicine, *Sleep Disorders*, `http://sleep.stanford.edu/sleep-disorders`, accessed at July 2017.

13. NationalSleepFoundation, *Sleep Disorders*, `https://sleepfoundation.org/sleep-disorders-problems`, accessed at July 2017.

14. Miwa, H., S.-I. Sasahara and T. Matsui, "Roll-over Detection and Sleep Quality Measurement using a Wearable Sensor", *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pp. 1507–1510, Aug 2007.

15. Drucker-Colin, R., "The Function of Sleep is to Regulate Brain Excitability in order to Satisfy The Requirements Imposed by Waking", *Behavioural Brain Research*, Vol. 69, No. 1–2, pp. 117 – 124, 1995, the Function of Sleep.

16. NHLBI, *Why Is Sleep Important?*, `http://www.nhlbi.nih.gov`, accessed at July 2017.

17. Terán-Santos, J., A. Jimenez-Gomez and J. Cordero-Guevara, "The Association between Sleep Apnea and the Risk of Traffic Accidents", *New England Journal of Medicine*, Vol. 340, No. 11, pp. 847–851, 1999.

18. Lattimore, J.-D. L., D. S. Celermajer and I. Wilcox, "Obstructive Sleep Apnea and Cardiovascular Disease", *Journal of the American College of Cardiology*, Vol. 41, No. 9, pp. 1429 – 1437, 2003.

19. NCSDR-NHTSA, *Drowsy Driving and Automobile Crashes*, `http://www.nhtsa.`

gov/people/injury/drowsy_driving1/Drowsy.html, accessed at July 2017.

20. Altevogt, B. M., H. R. Colten *et al.*, *Sleep disorders and sleep deprivation: an unmet public health problem*, National Academies Press, 2006.

21. SleepDisordersGuide, *Sleep Apnea Statistics*, http://www.sleepdisordersguide.com/article/sleep-disorders/sleep-apnea-statistics-the-statistics-of-sleep-apnea, accessed at July 2017.

22. Krystal, A. D. and J. D. Edinger, "Measuring sleep quality", *Sleep medicine*, Vol. 9, pp. S10–S17, 2008.

23. Buysse, D. J., C. F. Reynolds, T. H. Monk, S. R. Berman and D. J. Kupfer, "The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research", *Psychiatry research*, Vol. 28, No. 2, pp. 193–213, 1989.

24. Johns, M. W., "What is excessive daytime sleepiness", *Sleep Deprivation: Causes, Effects and Treatment. Nova Science Publishers*, pp. 59–94, 2009.

25. Deutsch, P. A., M. S. Simmons and J. M. Wallace, "Cost-effectiveness of split-night polysomnography and home studies in the evaluation of obstructive sleep apnea syndrome", *Journal of Clinical Sleep Medicine*, Vol. 2, No. 2, pp. 145–53, 2006.

26. Suetsugi, M., Y. Mizuki, K. Yamamoto, S. Uchida and Y. Watanabe, "The effect of placebo administration on the first-night effect in healthy young volunteers", *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, Vol. 31, No. 4, pp. 839–847, 2007.

27. ActiGraphCorp, *Accurate and Reliable In-home Sleep Assessment*, http://www.actigraphcorp.com, accessed at July 2017.

28. ComprendoCorp, *Zeo Sleep Manager Pro*, 2012, http://www.digifit.com/Zeo/,

accessed at June 2016.

29. Alqassim, S., M. Ganesh, S. Khoja, M. Zaidi, F. Aloul and A. Sagahyroon, "Sleep Apnea Monitoring using mobile phones", *e-Health Networking, Applications and Services (Healthcom), 2012 IEEE 14th International Conference on*, pp. 443–446, Oct 2012.

30. Nakano, H., K. Hirayama, Y. Sadamitsu, A. Toshimitsu, H. Fujita, S. Shin and T. Tanigawa, "Monitoring sound to quantify snoring and sleep apnea severity using a smartphone: proof of concept", *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*, Vol. 10, No. 1, p. 73, 2014.

31. Sun, X., Z. Lu, W. Hu and G. Cao, "SymDetector: detecting sound-related respiratory symptoms using smartphones", *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 97–108, ACM, 2015.

32. Tseng, M.-H., H.-C. Hsu, C.-C. Chang, H. Ting, H.-C. Wu and P.-H. Tang, "Development of an intelligent app for obstructive sleep apnea prediction on Android smartphone using data mining approach", *Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC), 2012 9th International Conference on*, pp. 774–779, IEEE, 2012.

33. Behar, J., A. Roebuck, M. Shahid, J. Daly, A. Hallack, N. Palmius, J. Stradling and G. D. Clifford, "SleepAp: An automated obstructive sleep apnoea screening application for smartphones", *IEEE journal of biomedical and health informatics*, Vol. 19, No. 1, pp. 325–331, 2015.

34. Han, T.-Y., S.-D. Min and Y. Nam, "A real-time sleep monitoring system with a smartphone", *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2015 9th International Conference on*, pp. 458–461, IEEE, 2015.

35. Al-Mardini, M., F. Aloul, A. Sagahyroon and L. Al-Husseini, "On the use of smartphones for detecting obstructive sleep apnea", *Bioinformatics and Bioengineering*

(BIBE), 2013 IEEE 13th International Conference on, pp. 1–4, IEEE, 2013.

36. Oliver, N. and F. Flores-Mangas, "Healthgear: Automatic sleep apnea detection and monitoring with a mobile phone.", *JCM*, Vol. 2, No. 2, pp. 1–9, 2007.

37. Patil, D., V. Wadhai, S. Gujar, K. Surana, P. Devkate and S. Waghmare, "APNEA detection on smart phone", *International Journal of Computer Applications*, Vol. 59, No. 7, 2012.

38. Nandakumar, R., S. Gollakota and N. Watson, "Contactless sleep apnea detection on smartphones", *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 45–57, ACM, 2015.

39. Ren, Y., C. Wang, J. Yang and Y. Chen, "Fine-grained sleep monitoring: Hearing your breathing with smartphones", *Computer Communications (INFOCOM), 2015 IEEE Conference on*, pp. 1194–1202, IEEE, 2015.

40. Bai, Y., B. Xu, Y. Ma, G. Sun and Y. Zhao, "Will you have a good sleep tonight?: sleep quality prediction with mobile phone", *Proceedings of the 7th International Conference on Body Area Networks*, pp. 124–130, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2012.

41. Lane, N. D., M. Mohammod, M. Lin, X. Yang, H. Lu, S. Ali, A. Doryab, E. Berke, T. Choudhury and A. Campbell, "Bewell: A smartphone application to monitor, model and promote wellbeing", *5th international ICST conference on pervasive computing technologies for healthcare*, pp. 23–26, 2011.

42. Gautam, A., V. S. Naik, A. Gupta, S. Sharma and K. Sriram, "An smartphone-based algorithm to measure and model quantity of sleep", *Communication Systems and Networks (COMSNETS), 2015 7th International Conference on*, pp. 1–6, IEEE, 2015.

43. Gu, W., L. Shangguan, Z. Yang and Y. Liu, "Sleep hunter: Towards fine grained

sleep stage tracking with smartphones", *IEEE Transactions on Mobile Computing*, Vol. 15, No. 6, pp. 1514–1527, 2016.

44. Min, J.-K., A. Doryab, J. Wiese, S. Amini, J. Zimmerman and J. I. Hong, "Toss'n'turn: smartphone as sleep and sleep quality detector", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 477–486, ACM, 2014.

45. Sano, A., A. J. Phillips, Z. Y. Amy, A. W. McHill, S. Taylor, N. Jaques, C. A. Czeisler, E. B. Klerman and R. W. Picard, "Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones", *Wearable and Implantable Body Sensor Networks (BSN), 2015 IEEE 12th International Conference on*, pp. 1–6, IEEE, 2015.

46. Hao, T., G. Xing and G. Zhou, "iSleep: unobtrusive sleep quality monitoring using smartphones", *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, p. 4, ACM, 2013.

47. Suzuki, T., K. Ouchi, K.-i. Kameyama and M. Takahashi, "Development of a Sleep Monitoring System with Wearable Vital Sensor for Home Use.", *BIODEVICES*, pp. 326–331, 2009.

48. Chen, Z., M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury and A. T. Campbell, "Unobtrusive sleep monitoring using smartphones", *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*, pp. 145–152, IEEE, 2013.

49. Natale, V., M. Drejak, A. Erbacci, L. Tonetti, M. Fabbri and M. Martoni, "Monitoring sleep with a smartphone accelerometer", *Sleep and Biological Rhythms*, Vol. 10, No. 4, pp. 287–292, 2012.

50. Bernardeschi, C., M. G. Cimino, A. Domenici and 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*, pp. 251–259, Springer, 2016.

51. Hernandez, J., D. McDuff and R. W. Picard, "Biowatch: estimation of heart and breathing rates from wrist motions", *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*, pp. 169–176, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2015.

52. Uçar, M. K., M. R. Bozkurt and C. Bilgin, "Statistical investigation of heart rate variable derived from photoplethysmography signal in sleep staging process", *Signal Processing and Communications Applications Conference (SIU), 2017 25th*, pp. 1–4, IEEE, 2017.

53. Sun, X., L. Qiu, Y. Wu, Y. Tang and G. Cao, "SleepMonitor: Monitoring Respiratory Rate and Body Position During Sleep Using Smartwatch", *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 1, No. 3, p. 104, 2017.

54. Joachims, T., *Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, 2002.

55. Lafferty, J. D., A. McCallum and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 282–289, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001.

56. Friedman, J., T. Hastie and R. Tibshirani, "Additive Logistic Regression: a Statistical View of Boosting", *Annals of Statistics*, Vol. 28, p. 2000, 1998.

57. Mitchell, T. M., *Machine Learning*, McGraw-Hill, Inc., New York, NY, USA, 1 edn., 1997.

58. Quinlan, J. R., "Induction of Decision Trees", *Machine Learning*, Vol. 1, No. 1, pp. 81–106, Mar. 1986.

59. Friedman, N., D. Geiger and M. Goldszmidt, "Bayesian Network Classifiers", *Mach. Learn.*, Vol. 29, No. 2-3, pp. 131–163, Nov. 1997.

60. Nevill-Manning, C. G., G. Holmes and I. H. Witten, "The development of Holte's 1R classifier", *Artificial Neural Networks and Expert Systems, 1995. Proceedings., Second New Zealand International Two-Stream Conference on*, pp. 239–242, Nov 1995.

61. Blunsom, P., "Hidden markov models", *Lecture notes, August*, Vol. 15, pp. 18–19, 2004.

62. Breiman, L., "Random Forests", *Machine Learning*, Vol. 45, No. 1, pp. 5–32, Oct. 2001.

63. Hartigan, J. A. and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm", *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 28, No. 1, pp. 100–108, 1979.

64. Ng, A. Y. and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes", *Advances in neural information processing systems*, pp. 841–848, 2002.

65. NHLBI, *What Is Sleep Apnea?*, `https://www.nhlbi.nih.gov/health/health-topics/topics/sleepapnea/`, accessed at July 2017.

66. Kawano, Y., A. Tamura, T. Watanabe and J. Kadota, "Influence of the severity of obstructive sleep apnea on heart rate", *Journal of cardiology*, Vol. 56, No. 1, pp. 27–34, 2010.

67. Camcı, B., A. Y. Kahveci, B. Arnrich and C. Ersoy, "Sleep apnea detection via

smart phones", *Signal Processing and Communications Applications Conference (SIU), 2017 25th*, pp. 1–4, IEEE, 2017.

68. Rughani, P. H. and M. Dahiya, "Analysis of Android Smart Watch Artifacts", *International Journal of Scientific & Engineering Research*, Vol. 6, No. 8, 2015.

69. Rawassizadeh, R., M. Tomitsch, K. Wac and A. M. Tjoa, "UbiqLog: a generic mobile phone-based life-log framework", *Personal and ubiquitous computing*, Vol. 17, No. 4, pp. 621–637, 2013.

70. Berry, R. B., 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", *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*, Vol. 8, No. 5, p. 597, 2012.

71. Haba-Rubio, J. and J. Krieger, "Evaluation instruments for sleep disorders: A brief history of polysomnography and sleep medicine", *Introduction to modern sleep technology*, pp. 19–31, Springer, 2012.

72. Barreiro, B., G. Badosa, S. Quintana, L. Esteban and J. Heredia, "Comparison between automatic and manual analysis in the diagnosis of obstructive sleep apnea-hypopnea syndrome", *Archivos de Bronconeumología (English Edition)*, Vol. 39, No. 12, pp. 544–548, 2003.

73. Wolkove, N., M. Baltzan, H. Kamel, R. Dabrusin and M. Palayew, "Long-term compliance with continuous positive airway pressure in patients with obstructive sleep apnea", *Canadian respiratory journal*, Vol. 15, No. 7, pp. 365–369, 2008.

74. Teplan, M. *et al.*, "Fundamentals of EEG measurement", *Measurement science review*, Vol. 2, No. 2, pp. 1–11, 2002.

75. Jubran, A., "Pulse oximetry", *Intensive care medicine*, Vol. 30, No. 11, pp. 2017–2020, 2004.

76. SamsungInc., *Samsung Galaxy Gear S3*, `http://www.samsung.com/global/galaxy/gear-s3`, accessed at July 2017.

77. TizenProject, *Tizen Operating System*, `https://www.tizen.org`, accessed at July 2017.

78. Data, R., *Learn HTML, JavaScript, CSS*, `https://www.w3schools.com/`, accessed at July 2017.

79. SmartmobDevelopment, *Smart Voice Recorder*, `http://recorder.smartmobdev.com/`, accessed at July 2017.

80. Alan, H. F., *Sensor Log*, `play.google.com/store/`, accessed at July 2017.

81. CareFusion, *SomnoStar*, `http://www.carefusion.com/our-products/respiratory-care/sleep-diagnostics-and-therapy/somnostar-z4`, accessed at July 2017.

82. Alvarez-Estevez, D., *European Data Format*, `http://www.edfplus.info/`, accessed at July 2017.

83. Zhang, Z., "Variable selection with stepwise and best subset approaches", *Annals of translational medicine*, Vol. 4, No. 7, 2016.

84. Baratloo, A., M. Hosseini, A. Negida and G. El Ashal, "Part 1: simple definition and calculation of accuracy, sensitivity and specificity", *Emergency*, Vol. 3, No. 2, pp. 48–49, 2015.

85. Peterson, L. E., "K-nearest neighbor", *Scholarpedia*, Vol. 4, No. 2, p. 1883, 2009.

86. Gonçalves, L., A. Subtil, M. R. Oliveira and P. Bermudez, "ROC curve estimation:

An overview", *REVSTAT–Statistical Journal*, Vol. 12, No. 1, pp. 1–20, 2014.

87. Rachim, V. P., G. Li and W.-Y. Chung, "Sleep apnea classification using ECG-signal wavelet-PCA features", *Bio-medical materials and engineering*, Vol. 24, No. 6, pp. 2875–2882, 2014.

88. Pale, U., F. Thürk and E. Kaniusas, "Heart rate variability analysis using different wavelet transformations", *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2016 39th International Convention on*, pp. 1649–1654, IEEE, 2016.

89. Kheder, G., A. Kachouri, R. Taleb, M. Ben Messaoud and M. Samet, "Feature extraction by wavelet transforms to analyze the heart rate variability during two meditation techniques", *Advances in Numerical Methods*, pp. 379–387, Springer, 2009.

90. Várady, P., T. Micsik, S. Benedek and Z. Benyó, "A novel method for the detection of apnea and hypopnea events in respiration signals", *IEEE Transactions on Biomedical Engineering*, Vol. 49, No. 9, pp. 936–942, 2002.

91. Fukuda, O., Y. Nagata, K. Homma and 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, pp. 1769–1772, IEEE, 2001.

92. Andreao, R. V., B. Dorizzi and J. Boudy, "ECG signal analysis through hidden Markov models", *IEEE Transactions on Biomedical engineering*, Vol. 53, No. 8, pp. 1541–1549, 2006.

93. Isa, S. M., M. I. Fanany, W. Jatmiko and A. M. Arymurthy, "Sleep Apnea Detection from ECG signal: analysis on optimal features, principal components, and nonlinearity", *Bioinformatics and Biomedical Engineering,(iCBBE) 2011 5th International Conference on*, pp. 1–4, IEEE, 2011.

94. Lemaître, G., F. Nogueira and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning", *Journal of Machine Learning Research*, Vol. 18, No. 17, pp. 1–5, 2017.

APPENDIX A: PERFORMANCE COMPARISONS

Table A.1. Classification Results with Undersampled-Featured Data with K-Fold Cross Validation.

DT	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	1874	687	72.06	72.56	70.95	73.17
Abnormal	744	1817				
KNN	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	1649	912	73.57	69.91	82.74	64.39
Abnormal	442	2119				
RF	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	1853	708	80.63	76.28	88.91	72.35
Abnormal	284	2277				
SVM	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	1679	882	75.50	71.27	85.44	65.56
Abnormal	373	2188				

Table A.2. Classification Results with Undersampled-Featured Data with Percentage.

DT	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	450	187	69.86	70.18	69.07	70.64
Abnormal	197	440				
KNN	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	355	282	65.54	62.99	75.35	55.73
Abnormal	157	480				
RF	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	448	189	74.33	72.53	78.34	70.33
Abnormal	138	499				
SVM	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	424	213	71.82	69.74	77.08	66.56
Abnormal	146	491				

Table A.3. Classification Results with Selected-Featured (Random Forest) Data with K-Fold Cross Validation.

DT	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	1896	665	73.70	73.86	73.37	74.03
Abnormal	682	1879				
KNN	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	1754	807	70.81	69.89	73.14	68.49
Abnormal	688	1873				
RF	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	1868	693	80.96	76.68	88.99	72.94
Abnormal	282	2279				
SVM	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	1831	730	73.33	72.50	75.17	71.50
Abnormal	636	1925				

Table A.4. Classification Results with Selected-Featured (Random Forest) Data with Percentage.

DT	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	463	174	68.92	70.46	65.15	72.68
Abnormal	222	415				
KNN	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	401	236	67.82	66.24	72.68	62.95
Abnormal	174	463				
RF	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	459	178	74.49	73.35	76.92	72.06
Abnormal	147	490				
SVM	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	402	235	71.59	68.46	80.06	63.11
Abnormal	127	510				

Table A.5. Classification Results with Selected-Featured (Stepwise) Data with K-Fold Cross Validation.

DT	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	1883	678	72.63	73.04	71.73	73.53
Abnormal	724	1837				
KNN	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	1709	852	69.99	68.77	73.25	66.73
Abnormal	685	1876				
RF	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	1845	716	80.30	76.01	88.56	72.04
Abnormal	293	2268				
SVM	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	1666	895	75.61	71.15	86.18	65.05
Abnormal	354	2207				

Table A.6. Classification Results with Selected-Featured (Stepwise) Data with Percentage.

DT	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	409	228	62.87	63.23	61.54	64.21
Abnormal	245	392				
KNN	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	382	255	65.07	63.68	70.17	59.97
Abnormal	190	447				
RF	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	452	185	75.20	73.23	79.43	70.96
Abnormal	131	506				
SVM	**Predicted Class**		**Results (%)**			
True Class	Normal	Abnormal	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Normal	413	224	74.57	70.57	84.30	64.84
Abnormal	100	537				