DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

PERFORMANCE ANALYSIS OF FEATURES OBTAINED BY PCA (PRINCIPAL COMPONENT ANALYSIS) DIMENSIONALITY REDUCTION METHOD FOR DIAGNOSING PAF (PAROXYSMAL ATRIAL FIBRILLATION) PATIENTS

by

Safa SADAGHIYANFAM

December, 2018
İZMİR

PERFORMANCE ANALYSIS OF FEATURES OBTAINED BY PCA (PRINCIPAL COMPONENT ANALYSIS) DIMENSIONALITY REDUCTION METHOD FOR DIAGNOSING PAF (PAROXYSMAL ATRIAL FIBRILLATION) PATIENTS

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Science
in Biomedical Technologies Program

by
Safa SADAGHIYANFAM

December, 2018
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **“PERFORMANCE ANALYSIS OF FEATURES OBTAINED BY PCA (PRINCIPAL COMPONENT ANALYSIS) DIMENSIONALITY REDUCTION METHOD FOR DIAGNOSING PAF (PAROXYSMAL ATRIAL FIBRILLATION) PATIENTS”** completed by **SAFA SADAGHIYANFAM** under supervision of **PROF. DR. MEHMET KUNTALP** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Mehmet KUNTALP

Supervisor

Dr. M. Makinacı

(Jury Member)

Dr. Nalan ÖZKURT

(Jury Member)

Prof. Dr. Kadriye ERTEKİN
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to express my special thanks of gratitude to my supervisor Prof. Dr. Mehmet Kuntalp who gave me the golden opportunity to do this wonderful project, which also helped me in doing a lot of research and coming to know about many things. I am really thankful to him. I am greatly indebted to him for accepting me as his student and giving me an opportunity to work towards my M.Sc. degree under his patient guidance.

Living abroad far from family and hometown for the first time was a nice challenge. But I found new friends. Başak Ilgım was one who accepted me as a friend and helped me like a sister. Also special thanks to her mother Serap and her father Prof, Dr, Ali Necati Gökmen for their emotional supports. They always were with me like a family and didn’t allow me to feel alone. My special thank goes to my dear friends, Omid and Nesa and Ata, who were there anytime I needed.

I am especially thankful to Aran Gül who encouraged me to come to Izmir for educating and helped me a lot in finalizing this project within the limited time frame.

Finally, and most importantly, I would like to extend special thanks to my parents, Gholamhossein and Nahid for their sacrifices, inspiration and patience throughout my studies. Without their full support and love, this thesis would not have been possible to achieve. I am also thankful to my brother and his wife and their lovely child, Amin, Sahar and Amirali, who loved me and trusted me.

Safa SADAGHIYANFAM

PERFORMANCE ANALYSIS OF FEATURES OBTAINED BY PCA (PRINCIPAL COMPONENT ANALYSIS) DIMENSIONALITY REDUCTION METHOD FOR DIAGNOSING PAF (PAROXYSMAL ATRIAL FIBRILLATION) PATIENTS

ABSTRACT

Principal Component Analysis (PCA) is an offered scheme for feature extraction and dimension reduction. It has been used extensively in many applications involving high-dimensional data. In this study, we compared the effectivity of PCA features extracted from 33 short-term Heart Rate Variability (HRV) features obtained from normal sinus rhythm (NSR) ECG records for the diagnosis of Paroxysmal Atrial Fibrillation (PAF) disease. Within this framework, different data sets consisting of 33 to 1 features obtained from PCA were used as input to the classification algorithm, which is chosen as the K-Nearest Neighbor (kNN) algorithm. Different values for K and difference distance metrics were utilized to find the best performance. Then the same procedure is applied to another HRV dataset. This set consists of 8 best HRV indices chosen from among the 33 HRV indices by a Genetic Algorithm. The obtained results from both studies elicit that it is possible to further reduce the number of input dimension of a classification system by using PCA algorithm without a reduction in the performance of the system.

Keywords: Principal component analysis, Paroxysmal atrial fibrillation, Normal sinus rhythm, HRV, ECG, K-nearest neighbor

PAF (PAROKSİSMAL ATRİYAL FİBRİLASYON) HASTALARININ TEŞHİSİ İÇİN TBA (TEMEL BİLEŞENLER ANALİZİ) BOYUT AZALTMA METODUYLA ELDE EDİLEN ÖZNİTELİKLERİN PERFORMANS ANALİZİ

ÖZ

İlke Bileşen Analizi (PCA), özellik çıkarma ve boyut küçültme için sunulan bir şemadır. Yüksek boyutlu verileri içeren birçok uygulamada yaygın olarak kullanılmıştır. Bu çalışmada, normal sinüs ritm (NSR) EKG kayıtlarından Paroksismal Atriyal Fibrilasyon (PAF) tanısı koyabilmek için bu kayıtlardan elde edilen 33 tane kısa süreli Kalp Hızı Değişkenliği (KHD) indisinden PCA tarafından çıkartılan özniteliklerin etkinliğini karşılaştırdık. Bu çerçevede, K-En Yakın Komşu (kNN) algoritması olarak seçilen sınıflandırma algoritmasına girdi olarak PCA'dan elde edilen 1'den 33'e kadar öznitelik içeren farklı veri setleri kullanılmıştır. En iyi performansı bulmak için farklı değerlerde K ve farklı mesafe metrikleri kullanılmıştır. Daha sonra aynı yaklaşım başka bir KHD veri setine uygulanmıştır. Bu veri seti önceki 33 KHD indisi arasından bir Genetik Algoritma tarafından belirlenen en iyi 8 KHD indisinden oluşmuştur. Elde edilen sonuçlar, bir sınıflandırıcı sistemin girdi boyutlarının PCA algoritması kullanılarak performansda bir azalma olmadan daha da düşürülebileceğine işaret etmektrdir.

Anahtar kelimeler: Temel bileşenler analizi, Paroksismal atriyal fibrilasyon, Normal sinüs ritim, KHD, EKG, K-en yakın komşu

CONTENTS

Page

M.Sc THESIS EXAMINATION RESULT FORM ii

ACKNOWLEDGMENTS iii

ABSTRACT iv

ÖZ v

LIST OF FIGURES ix

LIST OF TABLES xii

CHAPTER ONE - INTRODUCTION 1

CHAPTER TWO - PHYSIOLOGICAL BACKGROUND 5

2.1 Circulatory System 5
2.2 Anatomy of Heart 6
  2.2.1 Electroconduction System of the Heart 8
  2.2.2 Heart Rate Regulation 10
2.3 Electrocardiography 12
2.4 Arrhythmias 16
  2.4.1 Ventricular Based Arrhythmias 17
  2.4.2 Supraventricular Based Arrhythmias 20
  2.4.3 Atrial Fibrillation 25

CHAPTER THREE - HEART RATE VARIABILITY ANALYSIS 27

3.1 Introduction 27
3.2 ECG Recording and Generation of RR interval 27
  3.2.1 RR Interval Time Series 27
  3.2.2 Time and Amplitude Resolution of ECG Recording 28
  3.2.3 Duration of the Recording 28

3.2.4 Stationarity of the Recording 28
3.2.5 Removing Trends 29
3.2.6 Ectopic Beats, Arrhythmias and Noise 29
3.3 Time Domain Analysis 30
3.4 Frequency Domain Analysis 31
3.4.1 Fourier Transform 31
3.4.2 Autoregressive Modeling 31
3.4.3 Resampling 32
3.4.4 Spectral Power 33
3.4.5 Effects of Respiration 34
3.4.6 Time-Frequency Analysis 35
3.5 Non-linear Analysis 35
3.5.1 Approximate Entropy and Sample Entropy 36
3.5.2 Detrended Fluctuation Analysis 38
3.5.3 Correlation Dimension 40
3.5.4 Return Map 41

CHAPTER FOUR - FEATURE EXCTRACTION WITH PCA 43

4.1 Principal Components Analysis (PCA) 44
4.1.1 Eigenstructure of Principal Components Analysis 46
4.1.2 Basic Data Representation 50
4.1.3 Dimensionality Reduction 51
4.1.4 The curse of Dimensionality 53

CHAPTER FIVE - METHODS 54

5.1 Data Acquisition 54
5.2 Classification 55
5.2.1 K-Nearest Neighbor Algorithm 56
5.3 Model Evaluation 58

CHAPTER SIX - RESULTS AND CONCLUSION 60

6.1 Results 60
6.2 Using the PCA algorithm on another dataset 93
6.3 Conclusion 103
6.4 Future Work 105

REFERENCES 106

YÖK

LIST OF FIGURES

Page

Figure 2.1 Principle representation of the human circulatory system.......................... 7
Figure 2.2 Principle anatomic structure of the heart and the direction of the blood.... 8
Figure 2.3 Conductive system of the heart.. 9
Figure 2.4 Regulation of heart rate by ANS system 12
Figure 2.5 Vectors for the standard ECG lead configuration 13
Figure 2.6 Conventional agreement positions for 12-lead ECG configuration 14
Figure 2.7 Einthoven’s triangle.. 14
Figure 2.8 Conventional arrangement of electrodes and Einthoven’s triangle.......... 15
Figure 2.9 One 12 lead ECG sheet... 16
Figure 2.10 Normal electrocardiogram.. 17
Figure 2.11 An example ECG strip of PVC: PVC every three beats. Inverted peaks can be observed between QRS complexes.. 18
Figure 2.12 An example ECG strip with ventricular flutter. Sinus wave pattern by oscillations.. 18
Figure 2.13 An example of tachycardia ECG strip.. 19
Figure 2.14 An example ECG strip with ventricular fibrillation. Rapid, wide, irregular ventricular complexes... 19
Figure 2.15 An example ECG stripe with ventricular escape beats. The QRS complexes are broad. 20
Figure 2.16 An example ECG stripe of right bundle branch block 20
Figure 2.17 An example ECG stripe with sinus bradycardia.................................... 20
Figure 2.18 An example ECG strip with sinus tachycardia...................................... 21
Figure 2.19 An example ECG strip with 2nd degree Type II SA block.................... 21
Figure 2.20 An example ECG strip with sinus arrest.. 22
Figure 2.21 An example ECG strip with sick sinus syndrome 22
Figure 2.22 An example ECG strip with 1st degree AV block with long PR interval. PR interval is increased 23
Figure 2.23 An example ECG strip with second degree AV block type 1 23
Figure 2.24 An example ECG strip with second degree AV block type two 23
Figure 2.25 An example ECG stripe with complete A-V block................................ 24

Figure 2.26 An example ECG strip with PAC. The second and the 7th are PACs..... 24
Figure 2.27 An example ECG stripe with atrial tachycardia. Abnormality of P wave 24
Figure 2.28 An example ECG stripe with atrial flutter. Sawtooth pattern is noticeable 25
Figure 2.29 An example ECG with AF.. 25
Figure 3.1 An example of RR interval series and corresponding FFT and AR spectrum...34
Figure 3.2 RR interval FFT as three different breathing frequencies of 0.25, 0.10, 0.05, and white noise breathing 35
Figure 3.3 The integrated time series and the local trends.. 39
Figure 3.4 A return map from an RR interval time series with SD along the diagonal SD2=5 ms and SD perpendicular to the diagonal SD1=18 ms 42
Figure 4.1 Relationship between vector X, its recreated version X' and vector e…..53
Figure 4.2 The curse of dimensionality... 53
Figure 5.1 The summary of data acquisition. The records which are represented by bold characters were utilized in this thesis ……………………………56
Figure 5.2 A simple diagram of KNN classifier. The test sample that is shown in green to be classified. The two classes train samples elicit with blue and red. If k is selected 3, the test samples will have belonged to blue class 58
Figure 6.1 Flowchart of the study ..……………………………………………60
Figure 6.2 The results of dimension reduction in 1NN classifier with 3 different distance functions 76
Figure 6.3 The results of dimension reduction in 3NN classifier with 3 different distance functions 76
Figure 6.4 The results of dimension reduction in 5NN classifier with 3 different distance functions 77
Figure 6.5 The performance evaluation of KNN with different K number in Euclidean distance 77
Figure 6.6 The performance evaluation of KNN with different K number in City Block distance 78

Figure 6.7 The performance evaluation of KNN with different K number in Minkowski…………………………………………………………...78
Figure 6.8 The results of PCA using GAs new dimensions as a new input in 1NN classifier with 3 different distance functions... 94
Figure 6.9 The results of PCA using GAs new dimensions as a new input in 3NN classifier with 3 different distance functions... 94
Figure 6.10 The results of PCA using GAs new dimensions as a new input in 5NN classifier with 3 different distance functions... 95
Figure 6.11 The performance evaluation of KNN with different K number in Euclidean distance by using GAs new features as a new input for PCA 96
Figure 6.12 The performance evaluation of KNN with different K number in City Block distance by using GAs new features as a new input for PCA...... 96
Figure 6.13 The performance evaluation of KNN with different K number in Minkowski distance by using GAs new features as a new input for PCA 97

LIST OF TABLES

Page

Table 2.1 Consequences of the autonomic nervous system on the heart 11

Table 6.1 Illustrating the features which were extracted by HRV analyzing……… .61

Table 6.2 Evaluating the performances of KNN after reducing dimension with K=1 and Euclidean distance 65

Table 6.3 Evaluating the performances of KNN after reducing dimension with K=1 and City Block 66

Table 6.4 Evaluating the performances of KNN after reducing dimension with K=1 and Minkowski 67

Table 6.5 Evaluating the performances of KNN after reducing dimension with K=3 and Euclidean distance 68

Table 6.6 Evaluating the performances of KNN after reducing dimension with K=3 and City Block 69

Table 6.7 Evaluating the performances of KNN after reducing dimension with K=3 and Minkowski 71

Table 6.8 Evaluating the performances of KNN after reducing dimension with K=5 and Euclidean 72

Table 6.9 Evaluating the performances of KNN after reducing dimension with K=5 and City Block 73

Table 6.10 Evaluating the performances of KNN after reducing dimension with K=5 and Minkowski distance 74

Table 6.11 Performance result of PCA with K=1 and Euclidean distance 79

Table 6.12 Performance result of PCA with K=1 and City Block distance............... 80

Table 6.13 Performance result of PCA with K=1 and Minkowski distance 82

Table 6.14 Performance result of PCA with K=3 and Euclidean distance 83

Table 6.15 Performance result of PCA with K=3 and City Block distance............... 85

Table 6.16 Performance result of PCA with K=3 and Minkowski distance 86

Table 6.17 Performance result of PCA with K=5 and Euclidean distance 88

Table 6.18 Performance result of PCA with K=5 and City Block distance............... 89

Table 6.19 Performance result of PCA with K=5 and Minkowski distance 91

Table 6.20 Performance result of GAs+PCA with K=1 and Euclidean distance 97

Table 6.21 Performance result of GAs+PCA with K=1 and City Block distance 98
Table 6.22 Performance result of GAs+PCA with K=1 and Minkowski distance 98
Table 6.23 Performance result of GAs+PCA with K=3 and Euclidean distance 99
Table 6.24 Performance result of GAs+PCA with K=3 and City Block distance 99
Table 6.25 Performance result of GAs+PCA with K=3 and Minkowski distance .. 100
Table 6.26 Performance result of GAs+PCA with K=5 and Euclidean distance 100
Table 6. 27 Performance result of GAs+PCA with K=5 and City Block distance .. 101
Table 6.28 Performance result of GAs+PCA with K=5 and Minkowski distance .. 101

CHAPTER ONE
INTRODUCTION

The body is a social order of about 100 trillion cells organized into different function structures, some of which are called organs. Each function structure contributes its share to the maintenance of homeostatic conditions in the extracellular fluid, which is called the internal environment. As long as normal conditions are maintained in this interval environment, the cells of the body continue to live and function properly. Each cell benefits from homeostasis, and in turn, each cell contributes its share toward the maintenance of homeostasis (Hall, 1946).

There are 11 body systems working in coordination. One of these body systems is the Cardiovascular system which is responsible for the circulation of the blood throughout the body. The main function of circulation is to serve the needs of the body tissues by: transporting nutrients to the body and waste product away, transporting hormones from one part of the body to another, and maintaining an appropriate environment in all the tissue fluids of the body for optimal survival and function of the cells. The rate of the blood flow through many tissues is controlled mainly in response to tissue requirements (Hall, 1946).

Arrhythmia refers to an irregular cardiac activity wherein the heartbeat is too fast or too slow which causes the heart to pump blood insufficiently. Atrial fibrillation (AF) is a type of arrhythmia which is caused by quivering heartbeat that can lead to the disordered electrical impulses in the atria. These disordered impulses conquer the normal sinus rhythm and the atria begin to contact in an unsuitable way. By this way, the upper chambers (atria) beat irregularly without coordinating with the lower chambers (ventricles) and atria cannot empty the blood to ventricles completely. It can increase the risk of having a stroke or forming blood thrombosis in the long run. If a clot enters the bloodstream, it can lead to stroke. Approximately 15% of all stroke cases are because of AF (Go, Hylek, Phillips, & Chang, 2001). AF is the most common arrhythmia (Camm et al., 2010; Miyasaka et al., 2006) and identified as a growing health-care burden because of the old population and survival from cardiac disorders.

In general population, currency of PAF is estimated 2 % (Camm et al., 2010; Stewart, Hart, Hole, & McMurray, 2001) while this ratio increases 0.5% with age at 40-50 years to 5% more than 65 years, and also this ratio increases by 14% in subjects more than 85 years (Majeed, Moser, & Carroll, 2001; Naccarelli, Varker, Lin, & Schulman, 2009). On the other hand, studies elicit that the real commonness is much higher (Gladstone et al., 2014; Sanna et al., 2014). The real numbers are debatable due to the approximately one-third of PAF cases is asymptomatic (Furberg et al., 1994; Savelieva & Camm, 2000) and the symptoms are assigned to other illnesses in symptomatic cases.

AF can be classified into self-terminating or non-self-terminating. If the episodes of AF terminate spontaneously or self-terminate within 48 hours, it is called paroxysmal atrial fibrillation (PAF). However, PAF episodes terminate within minutes generally (Hoshino, Ishizuka, Nagao, Shimizu, & Uchiyama, 2013; Page, Wilkinson, Clair, McCarthy, & Pritchett, 1994). If the episodes of AF last more than 7 days and do not self-terminate, it is called persistent AF. In this condition, electrical or pharmacological cardioversion must be used to recover the normal rhythm. AF is called permanent if AF does not terminate and exists for some time and cardioversion failed or not attempted (Levy et al., 2003; Lip & Hee, 2001).

Recent studies illustrate that a large number of PAF patients would switch to permanent AF by time (de Vos et al., 2010; Kato, Yamashita, Sagara, Iinuma, & Fu, 2004; Kerr et al., 2005; Van Gelder & Hemels, 2006). However, the risk of stroke in PAF is as same as the risk of permanent AF patients (Friberg, Hammar, & Rosenqvist, 2010; Hart et al., 2000). So the true diagnosing of PAF is crucial and necessary. If PAF patients are identified at the beginning stage of illness, the process of developing to more sustained AF can be avoided and also the risk of stroke can be decreased with suitable antithrombotic treatment in PAF patients.

On the other hand, detection of PAF is difficult due to it is asymptomatic behavior that there could be no specific symptoms assigned to PAF and also there could be no clear complaint of the patient. The symptoms of AF can include lightheadedness, weakness, dyspnea, palpitation, dizziness, chest pain and stroke ischemic attack some people may not have any symptoms at all. The first step in PAF diagnosis is to examine of the rhythm to find that if there is any irregular pulse. After detecting an irregular pulse, an electrocardiogram is performed. William Evens describes the complexity of diagnosing paroxysmal atrial fibrillation as “A fugitive illness that only visits a patient periodically may be more difficult to bear than one whose effects are durable and persistent. Paroxysmal atrial fibrillation is such an illness.” (Evans, 1959). In suspected PAF cases, a 24 hours ECG monitoring is used in patients with believed asymptomatic episodes or symptomatic episodes less than 24 hours apart. In patients with episodes, ECG recorder is used more than 24 hours apart (Cowan et al., 2014).

If the patient has PAF, it is very easy to identify this arrhythmia with a standard ECG. However, in reality, it is very difficult to take an ECG record during a PAF episode because these episodes could occur in a random manner and end in a few minutes. For this reason, there is a need for an efficient method that can diagnose PAF patients by using ECG recording which is taken during normal sinus rhythm periods. Many studies have been conducted in last decades none of them gained definitive results, so the problem stays open. Most of them are the number of materials passing through a system of Computers in Cardiology Challenge 2001: Predicting the onset of Paroxysmal Atrial Fibrillation which had two events. The first event was PAF screening, which is to clustering the PAF and non-PAF clusters. The second was PAF prediction, which is detecting the record that precedes PAF, by using the records of detected patients in the first event (Moody, Goldberger, McClennen, & Swiryn, 2001). The subject of this thesis is similar to the first event and a section of the database of this competition was used. On the other hand, there are rudimentary differences between the competition and this work, which will be mentioned in the next chapters.

In the literature, morphological features that are obtained from ECG records are the basis of the most successful works. Martinez et al. used morphological features of P-waves to discriminate ECG segments of healthy subjects and patients suffering from PAF (Martínez, Alcaraz, & Rieta, 2012, 2014). Maier et al. used different features obtained from heart rate variability analysis describing the magnitude as well as the regularity of heart rate fluctuations and the number of supraventricular and ventricular premature beats (Maier, Bauch, & Dickhaus, 2001). Chazal and Heneghan examined features from the interval based power spectral density of RR intervals, time domain measures, P-wave amplitude features and frequency representation of the P-wave. The effect of the length of the signal was also controlled by using 30-minute, 10-minute and 5-minute windows of the ECG signals. Their best performance was obtained with power spectral density (de Chazal & Heneghan, 2001).

The main goal of this thesis is to develop a system to detect PAF patients by comparing the effectivity of features obtained from Principal Component Analysis (PCA) from their normal sinus rhythm (NSR) ECG records. Thirty-three features, which were derived from heart rate variability (HRV) analysis (Hilavin, 2016) were used to create new features by using PCA algorithm. Within this framework, a set of features obtained from PCA were used as an input to the classification algorithm, which is chosen as the K-Nearest Neighbor (KNN) algorithm due to its simplicity and effectiveness. The results represent reducing dimensions' until approximately 10 have acceptable performances at first while reducing further affects negatively.

Chapter 2 of this thesis presents the physiological background of circulatory system and arrhythmias. Feature selection with PCA algorithm is explained in Chapter 3. The classification and evaluation methods are explained in Chapter 4. Results and discussion are presented in Chapter 5. Finally, a conclusion is given in Chapter 7.

CHAPTER TWO
PHYSIOLOGICAL BACKGROUND

The very fact that we remain alive is the result of complicated control systems, for hunger makes us seek food and fear makes us seek systems, the sensation of cold makes us look for warm places. Other forces cause us to look sociability and reproduce. Therefore, a human being is, in many ways, like an automaton. The fact that we are sensing, the feeling is part of this automatic sequence of life. So these special attributes permit us to exist under widely varying conditions.

The basic living unit of the body is the cell. Each organ is the aggregation of many different cells joined together by intercellular supporting structures. Each type of cells is performing one or more than one particular functions. Although the red blood cells are the most plentiful of any single type of cell in the body, there are the most 75 trillion additional cells of other types that do functions. Cells which ones have similar functions and also structures are organized into tissues. Two or more types of primary tissues construct organs who are their function is to do a specific function and different organs functions are cooperated each other to achieve a common activity and form body systems. The human body consists of digestive, respiratory, urinary, circulatory, skeletal muscular, immune system, integumentary, nerves, endocrine and reproductive systems which purpose in close collaboration in order to keep the body survival in different conditions.

The circulatory system maintains the appropriate environment in all the tissue fluids of the body. In this section, basic components of the circulatory system, the anatomy and, regulation system of the heart, measuring of cardiac signals and arrhythmias are described with related literature review.

2.1 Circulatory System

The function of the circulation is to service the needs of the body tissues. It is necessary to transport nutrients and minerals to the body tissues and also transport waste products away from the tissues. Transportation of hormones from one part of the body to another is another function of the circulation system. In general, it is vital to maintaining an appropriate environment in all the tissue fluids of the body for optimal survival. The rate of blood flow through many tissues is controlled in response to tissue requisite for nutrients. The principle representation of the human circulatory system is shown in Figure 2.1. The function of the arteries is to transport blood under high pressure to the tissues. Because of this, arteries have strong vascular walls and blood flow have a high velocity in the arteries. The arterioles are the small branches of the arterial system which have strong muscular walls that can close the arterioles or dilate the vessels. The functions of capillaries are to exchange fluids, nutrients, electrolytes, hormones, and other substances. At the result of this, the capillary walls are very thin and have many pores that permit to water and other small molecular substances to cross. The heart is actually two separate pumps which are a right heart that pumps blood through the lungs and left the heart that pumps blood through the peripheral organs. The heart provides the necessary pressure to the system for the circulation of blood through blood vessels.

2.2 Anatomy of Heart

The heart consists of four chambers; two upper chambers called atria which they receive blood and two lower chambers called ventricles which discharge blood. The right and the left sight of the heart work simultaneously. The right side of the heart receives carbon dioxide-rich blood from the body and sends it to the lungs; the left sight of the heart receives oxygen-rich blood from the lungs and send it to the rest of the body. The principle structure of the heart is given in Figure 2.2.

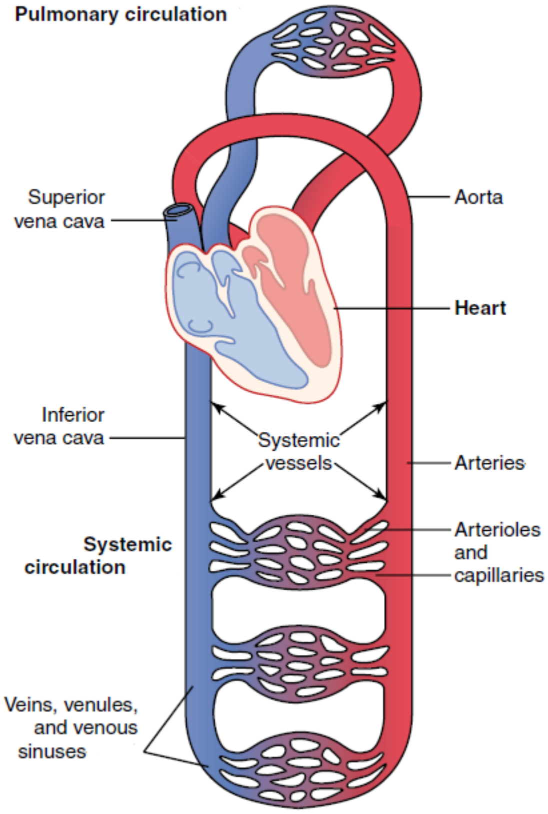

Figure 2.1 Principle representation of the human circulatory system (Hall, 1946)

The heart has four valves to ensure that the blood only flows in one direction:

- Aortic valve: Between the left ventricle and the aorta
- Mitral valve: Between the left atrium and the left ventricle
- Pulmonary valve: Between the right ventricle and the pulmonary artery
- Tricuspid valve: Between the right atrium and the right ventricle

The A-V valves that are the tricuspid and mitral valve, prevent backflow of blood from the ventricles to the atria. The semilunar valves that are the aortic valve and pulmonary valve, prevent backflow from the aorta and pulmonary arteries into the ventricles. These valves, close when a backward pressure gradient pushes blood backward. They open when a forward pressure gradient forces blood in the forward direction. Due to anatomical reasons, the thin and filmy A-V valves need almost no backflow to cause closure, while the much heavier semilunar valves need rather a rapid backflow for a few milliseconds. Papillary muscles that attach to the vanes of the A-V valves by the chordae tendineae. When ventricular walls contract the papillary muscles contract, but they do not help the valves to close. Instead, “they pull the vanes of the valves inward toward the ventricles to prevent their bulging too far back toward the atria during ventricular contraction”. The aortic and pulmonary valves function is different from the A-V valves. Because of the high pressures in the arteries at the end of systole, the semilunar valves snap to the closed position while to the much softer closure of A-V valves. Due to the smaller opening, the velocity of blood ejection through the aortic and pulmonary valves is far greater than that through the much larger A-V valves and also because of the rapid closure and ejection, the edges of the aortic and pulmonary valves are subjected to much greater mechanical abrasion than A-V valves (Hall, 1946).

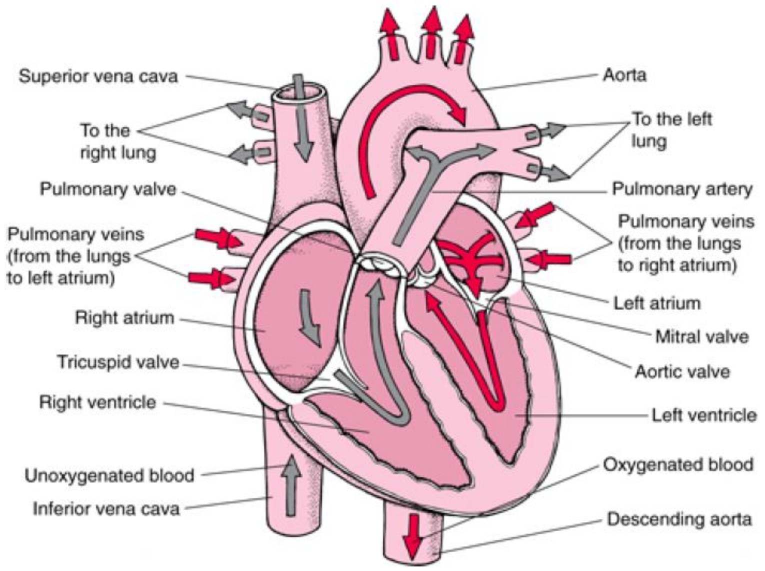

Figure 2.2 Principle anatomic structure of the heart and the direction of the blood (Webster, 1998)

2.2.1 Electroconduction System of the Heart

The heart beats as a result of the generation and conduction of electrical impulses. These impulses cause to the rhythmical contraction of the heart muscle and relax. The stable cycle of heart muscle contraction followed by relaxation causes blood to be pumped throughout the body. Figure 2.3 shows the specialized excitatory and conductive of the heart that controls cardiac contractions. The sinus node also called sinoatrial or S-A node where the normal rhythmical impulses are generated. There is an intermodal pathway that conducts impulses from sinus node to atrioventricular (A-V) node. The A-V node where impulses from atria delayed before passing into ventricular. A-V bundle conducts impulses into the ventricles and the left and right bundle branches of Purkinje fibers conduct the cardiac impulses to all parts of the ventricles.

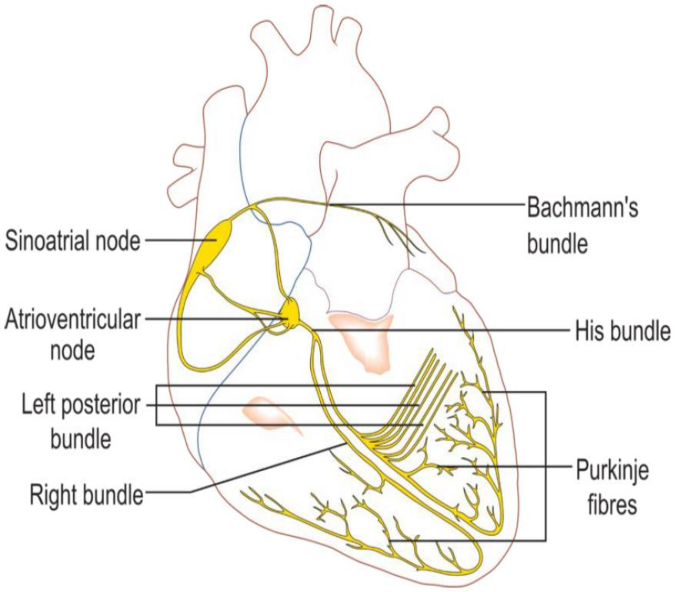

Figure 2.3 Conductive system of the heart (Todd, 2013)

The S-A node is a small and flattened cardiac muscle. It is located in the superior posterolateral wall of the right atrium. The sinus nodal fibers connect directly the atrial muscle fibers so that any action potential that begins in the sinus node separate into the atrial muscle wall. It can produce 70-80 action potentials per minute and has the rate of action potential origination. S-A node also referred to as the pacemaker of the heart and dominates all other cells. In this way, the action potential spreads through the entire atrial muscle to the A-V node.

The A-V node is located in the posterior wall of the right atrium behind the tricuspid valve. Pulse connects with the entering atrial intermodal pathway fibers and exciting A-V bundle. It can produce 40-60 action potentials per minute. It is noted that the A-V node has an important duty to delay electrical signals to pass to ventricles before the atria are fully empty. The action potentials continue into the bundle of His and the Purkinje network after the delay. The bundle of His originate at A-V node and enter the septum between the ventricle and divides into right and left bundle branches that go down the septum. Each branch spreads downward toward the apex of the ventricle dividing into smaller branches. From the time the impulse enters the bundle branches and reaches the terminations of the Purkinje fibers.

Special Purkinje fibers lead from the A-V node through the A-V bundle into the ventricles. Therefore, once the cardiac impulse enters the ventricular Purkinje conductive system, it spreads to the whole ventricular muscle mass. Not only the bundle of His, but also Purkinje fibers can fabricate 20-40 action potentials per minute.

2.2.2 Heart Rate Regulation

The cardiovascular system should be able to modify to changing circumstances to provide essential blood to tissues. This can be possible by the automatic nervous system (ANS) and also circulate hormones (Klabunde, 2011). The role of ANS on the cardiac system is changing the cardiac cycle length and also heart rate. Another consequence of ANS is the speed of conduction of electrical activity through the heart. Self-governing regulation of cardiovascular function is controlled by two nervous systems which are parasympathetic and sympathetic (Berne & Levy, 1997).

The heart is innervated by sympathetic and parasympathetic fibers. These sympathetic and parasympathetic nerves are regulating by the medulla that is the primary site in the brain. The hypothalamus and higher centers change the activity of medullary centers and also particularly prominent in regulating cardiovascular responses to adrenergic neurotransmitters. The amount of blood pumped can be increased by sympathetic while the output can be decreased by vagal (parasympathetic) stimulation.

Table 2.1 Consequences of the autonomic nervous system on the heart (Sherwood, 2015)

Affected area	**Effect of Parasympathetic Nervous system**	**Effect of Sympathetic Nervous system**
SA node	Decrease rate of depolarization	Increase rate of depolarization
AV node	Increase AV nodal delay	Decrease AV nodal delay
Ventricular conduction pathway	No effect	Increase the speed of conduction through the bundle of His and Purkinje
Atrial Muscle	Decrease contractility	Increase contractility
Ventricular Muscle	No effect	Increase contractility
Adrenal Medulla	No effect	Promotes epinephrine secretion which augments the sympathetic nervous system’s action the heart
Veins	No effect	Increase venous return, which strengthens the cardiac contraction

Strong sympathetic stimulation can increase the heart rate, also can increase the force of heart contraction, thereby increasing the volume of blood pumped and increasing the ejection pressure. Sympathetic stimulation often can increase the maximum cardiac output. On the other hand, sympathetic stimulation controls heart action in exercise situations or in an emergency when there is a need for increased blood flow. The vagal fibers are distributed into atria and not much to ventricles, where the power contraction of the heart occurs. The vagal stimulation decreases the heart rate combined with a slight decrease in heart contraction (Sherwood, 2015). The summary of effects of the autonomic nervous system is shown in Table 2.1.

The sympathetic nervous system can work for minutes to days to regulate blood pressure, depending on stress level. In addition to the sympathetic nervous system, there is another short-term regulator of blood pressure that named as arterial baroreceptors. These receptors located in the carotid sinus and the aortic arch. They sense the mechanical deformation and send action potential to the cardioacceleratory centers or cardioinhibitory centers of the brain and both of them connect to vasomotor center in the brain.

The baroreceptors are functional during very short-term changes in blood pressure that we can experience many times during the day. They are prominent for regulating the blood pressure changing in a very short term. They do not have any effect on long-term blood pressure regulation. Using variety of hormones change the blood pressure relies on changes in blood volume in long-term (Purves et al., 2012).

The regulation of heart rate is summarized in Figure 2.4. The control of heart rate is made with the direct intervention of sympathetic and parasympathetic nerves.

2.3 Electrocardiography

In a normal heart, the cardiac impulse originates in the sinus node and it appears at each respective point in the heart. The impulse spreads at moderate velocity through the atria but it delayed much more than 0.1 seconds in the A-V nodal region before reaching in the ventricular septal A-V bundle. When it has entered this bundle, it spreads very fast through the Purkinje fibers to the whole endocardial surfaces of the ventricles and then the impulse again spreads less rapidly through the ventricular muscle to the epicardial surfaces.

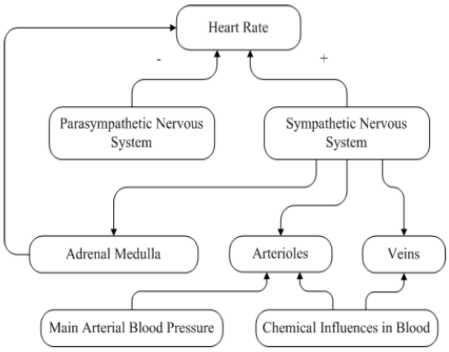
Heart Rate
-
+
Parasympathetic Nervous System
Sympathetic Nervous System
Adrenal Medulla
Arterioles
Veins
Main Arterial Blood Pressure
Chemical Influences in Blood

Figure 2.4 Regulation of heart rate by ANS system (Hilavin, 2016)

The simple and the most common method to monitor the electrical activity of the heart is to place electrodes on the surface of the body and record the electrical signal versus time. This noninvasive tool that gives a clinician valuable information about the electrical activity of the heart is called electrocardiogram (ECG).

The choice of the electrode configuration on the chest to record the ECG is prescribed by the type of clinical information required due to the voltage difference between a pair of electrodes that are called leads is the only representative of variation along a single axis from the heart (see Figure 2.5), there is not any three-dimensional activity information in single lead configuration.

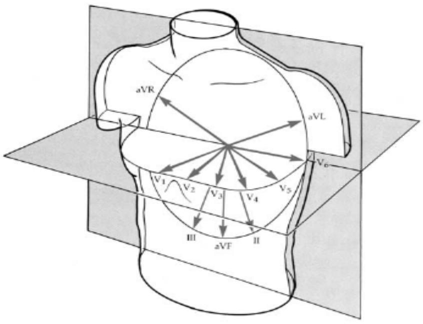

Figure 2.5 Vectors for the standard ECG lead configuration (Hilavin, 2016)

In order to record the ECG of patients, there is a need for 12 leads which are placed on the patient chest according to the standard positions. Figure 2.6 shows the standard positions of 12 ECG leads and Figure 2.7 shows the conventional arrangement vector of electrodes for recording the standard ECG.

Figure 2.8 illustrates electrical connections between the patient’s limb and the ECG for recording from the standard bipolar limb leads. The term bipolar means that ECG is recorded from the two electrodes placed on different sides of the heart. Therefore, one lead is not a single wire connecting from the body while the combination of pair wires and their electrodes to make an entire circuit between the body and the ECG.

Lead I. To record limb lead I, the positive terminal of electrocardiograph is connected to the left arm and the negative terminal of ECG is connected to the right arm. So when the point where the right arm connects to the chest is electronegative in regard to the point where the left arm connects, the ECG record positively which is above the zero voltage line in ECG. When the opposite is true then the ECG records below the line.

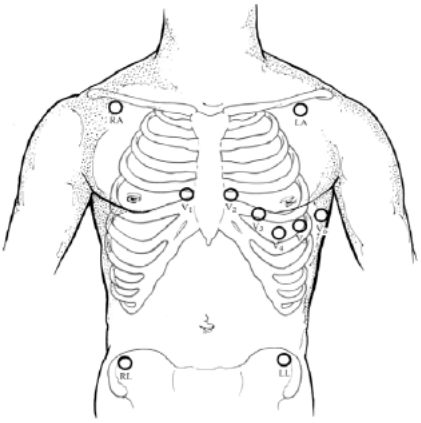

Figure 2.6 Conventional agreement positions for 12-lead ECG configuration (Hilavin, 2016)

Lead II. In recording limb Lead II, the positive terminal of the ECG is connected to the left leg and the negative terminal to the right arm. So when the right arm is negative in regret to the left leg the ECG records positively.

Lead III. In recording limb Lead III, the positive terminal of the ECG is connected to the left leg and the negative terminal to the left arm. It means that the ECG records positively when the left arm is negative in regret to the left leg.

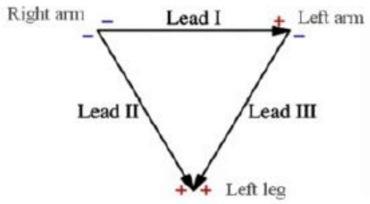

Figure 2.7 Einthoven’s triangle (Hilavin, 2016)

In Figure 2.7 the Einthoven’s triangle is drawn around the area of the heart. This elicits the two arms and the left leg from apices of a triangle surrounding the heart. The two apices at the upper part of the triangle illustrate the points at which he two arms connect electrically with the fluids surrounding the heart, and the lower apex is the point at which the left leg connects with the fluids.

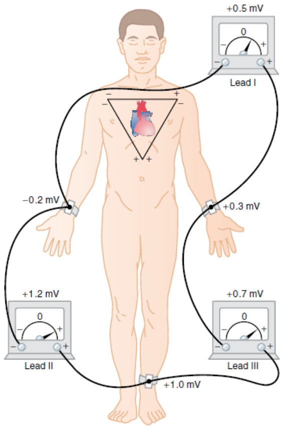

Figure 2.8 Conventional arrangement of electrodes and Einthoven’s triangle (Hall, 1946)

Einthoven’s Law said that “if the electrical potentials of any two of the three bipolar limb electrocardiographic leads are known at any given instant, the third one can be determined mathematically by simply summing the first two. Note, however, that the positive and negative signs of the different leads must be observed when making this summation” (Hall, 1946, p.125).

An example of normal ECG sheet is represented in Figure 2.9. ECG sheet is a graph paper that each square has 1 mm width and height. For every 5 small 1 mm squares, there is a heavier line which is forming a larger 5 mm square. The vertical axis measures the amplitude of the electrical current of heart and the horizontal axis measures time. It is standard that 10 mm in height equals 1mV and each 1mm square on horizontal equals 0.04 second and each large square equal to 0.20 second. By standard, short segments of every lead are presented with labels for the unchanged measurement time then after the cycle, the new measurement cycle results are represented.

The normal ECG record is composed of P wave and a QRS complex and a T wave which is shown in Figure 2.10. These waves are described as follows:

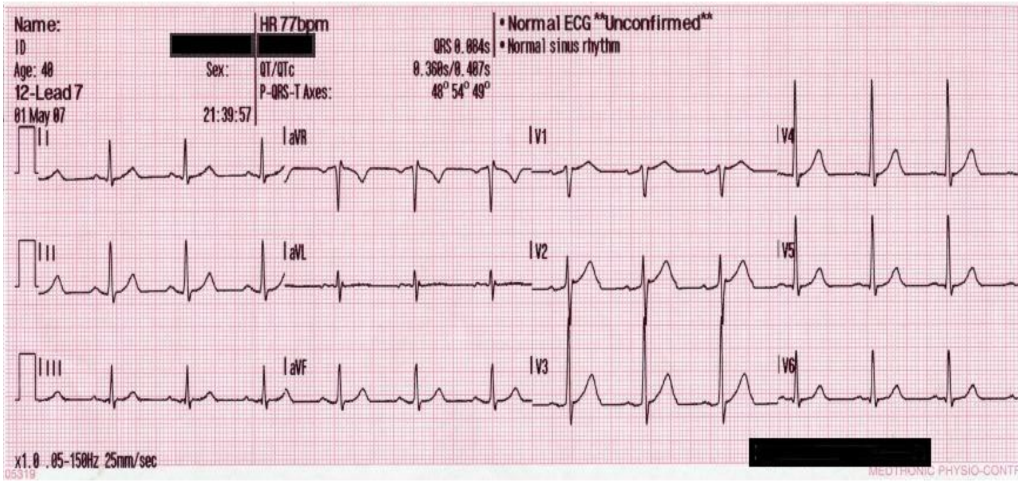

Figure 2.9 One 12 lead ECG sheet

- P-wave: It is caused by electrical potentials generated by the atria contraction begins after the atrial depolarization. Its duration is less than 0.12 s with the amplitude of less than 0.25 mV.
- QRS complex: It is caused by electrical potentials generated when the ventricles contracts after ventricle depolarization. At the same time, the atrial repolarization occurs but due to its low amplitude, it is not seen in the QRs complex. Therefore, both P wave and the components of the QRS are depolarization waves. The duration of QRS complex is less than 0.1 second and amplitude can be changed in different lead configuration while the maximum limit is 2.5-3 mV.
- T-wave: It is caused by electrical potentials generated when the ventricles recover from the state of depolarization and this wave refers to a repolarization wave.

Therefore, the ECG is composed of a depolarization and repolarization waves (Yanowitz, 2012).

2.4 Arrhythmias

The arrhythmia can be defined as any change of heart’s rhythm. During an arrhythmia, the electrical impulses may happen too fast or too slowly. If the heart does not beat properly, it will not pump blood effectively. By this way, the lungs and brain and all other organs will not work properly and may be damaged. The cause of the cardiac arrhythmias is usually one or a combination of following abnormalities of the heart (Gertsch,2003; Webster, 1995):

1. Unusual rhythmicity of the pacemaker.
2. Transfer of the pacemaker from the snus to another place in the heart.
3. Blocks at varies points in the spread of the impulse through the heart
4. Abnormal pathways of impulse transmission through the heart.
5. Unconstrained generation of spurious impulses in any part of the hearing.

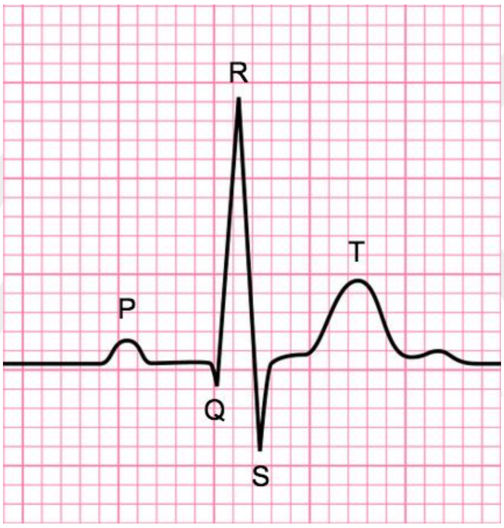

Figure 2.10 Normal electrocardiogram

2.4.1 Ventricular Based Arrhythmias

Premature Ventricular Contractions (PVC): PVC is extra heartbeat that begins in one of the heart’s two ventricles (see Figure 2.11). It is also called ventricular ectopic beats (VEB). They can result from such factors as cigarettes, excessive intake of coffee and lack of sleep. Fluttering or missed beats may be felt and meditation is needed for treatment.

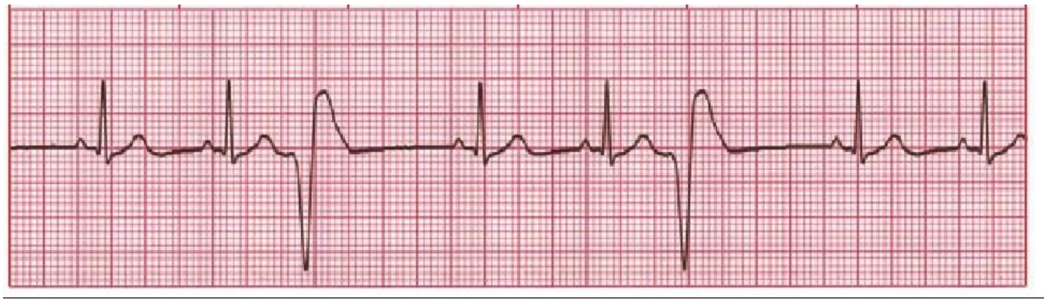

Figure 2.11 An example ECG strip of PVC: PVC every three beats. Inverted peaks can be observed between QRS complexes

Ventricular Flutter: It is applied to a rapid ventricular tachycardia (250 to 350 bpm). It has a sinusoidal QRS complex that makes hard to identify the QRS (see Figure 2.12). rates are too fast. It is dangerous because ventricles cannot fil completely. A prominent consequence of ventricular flutter is low cardiac output. The patients can feel dizziness, nausea, unconsciousness.

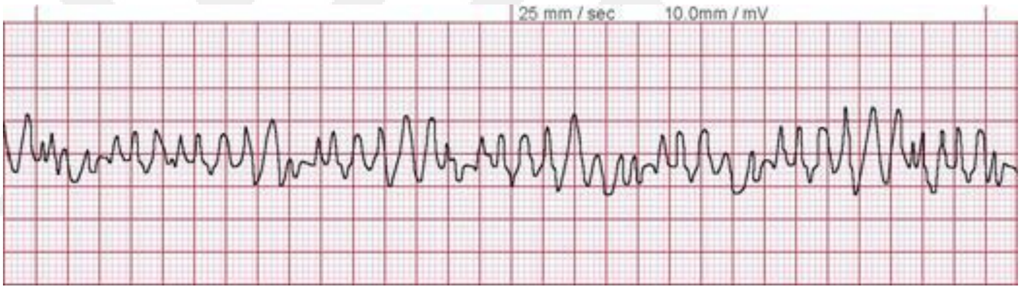

Figure 2.12 An example ECG strip with ventricular flutter. Sinus wave pattern by oscillations

Ventricular tachycardia (VT): In ventricular tachycardia, the ventricles caused to heartbeat to faster than normal due to abnormal electrical signals in ventricles. The heart rate is faster than 100 bpm and the width of the QRS complexes increases (see Figure 2.13). The common symptoms are heart palpitations, chest pain and shortness of breath and meditation are needed to return to normal rhythm.

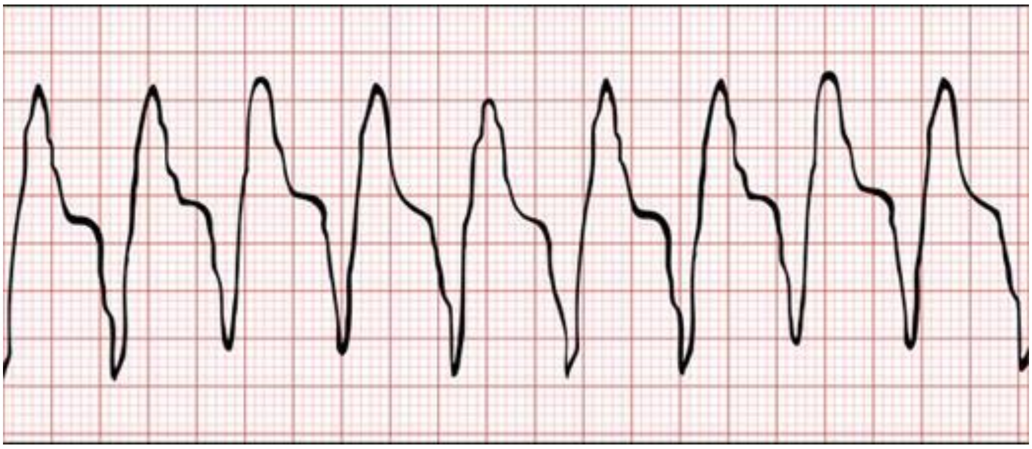

Figure 2.13 An example of tachycardia ECG strip

Ventricular fibrillation: It occurs when the heart beats with rapid and erratic electrical impulses and causes pumping the ventricles to quiver uselessly instead of pumping blood. A common sign of ventricular fibrillation is lack of consciousness. The ECG waveform is chaotic irregular deflections of varying amplitude and no identifiable P waves, QRS complexes of T waves (see figure 2.14). There is a need for defibrillation to return to normal rhythm.

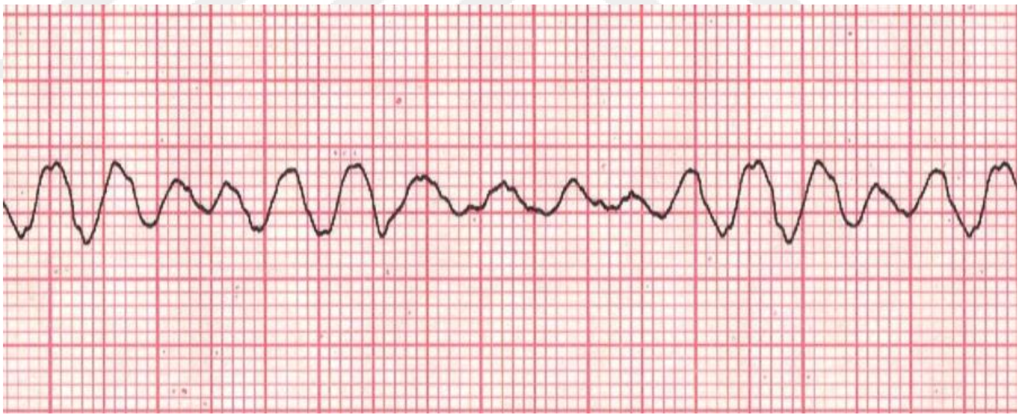

Figure 2.14 An example ECG strip with ventricular fibrillation. Rapid, wide, irregular ventricular complexes

Ventricular Escape Beats: It occurs after a sinus bradycardia with 176 seconds or in which a supraventricular pacemaker failed to fire as a result of hypoxia, excessive parasympathetic stimulation. It becomes a rhythm at a rate of 20-40 bpm. QRS complexes are broad more than 120 ms (see Figure 2.15). Cardiac output decreases due to loss of atrial kick. Meditation is needed to increase the rhythm.

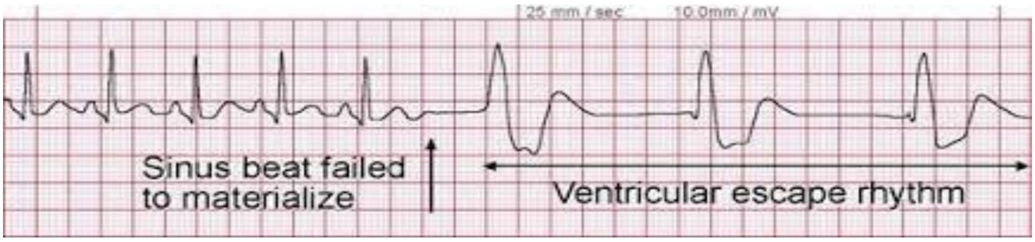

Figure 2.15 An example ECG stripe with ventricular escape beats. The QRS complexes are broad

Bundle Branch Block: There is a blockage along the pathway that electrical impulses transmit to make heat beat. The blockage can occur on the pathway which carries electrical impulses to the lower chambers (left and right ventricles) of heart. The characteristics of Bundle Brunch Block ECG waveform are m-shaped QRS complexes, blurred S-wave and notched R-wave (see Figure 2.16). The treatment may involve medications to reduce high blood pressure or lessen the effects of heart failure.

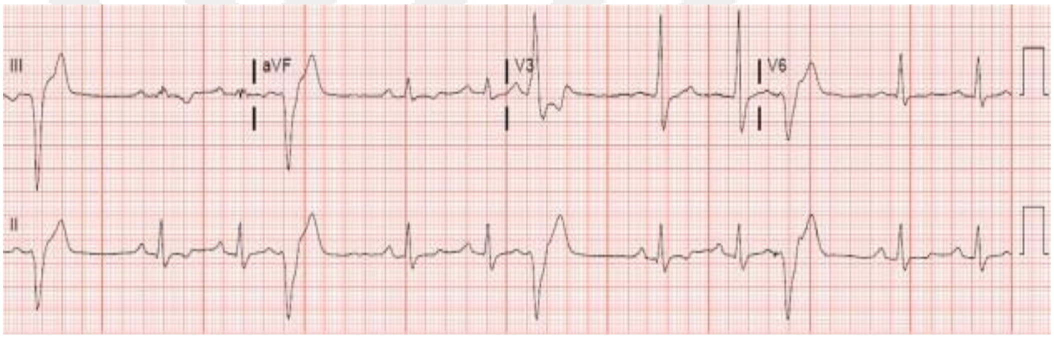

Figure 2.16 An example ECG stripe of right bundle branch block

2.4.2 Supraventricular Based Arrhythmias

Sinus Bradycardia: The bradycardia means a slow heart rate that is defined as fewer than 60 bpm (see Figure 2.17). It results the decreased cardiac output. Vagal stimulation is a cause of bradycardia and a pacemaker is used for treatment.

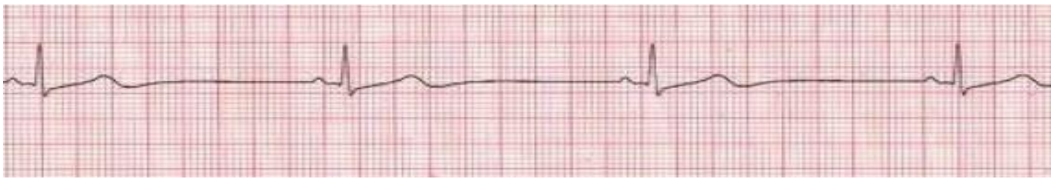

Figure 2.17 An example ECG stripe with sinus bradycardia

Sinus Tachycardia: The tachycardia means fast heart rate that is defined as faster than 100 bpm (see Figure 2.18). Some causes of tachycardia include increased body temperature and stimulation of the heart sympathetic nerves or toxic conditions of the heart. The cardiac output is decreased and medication is needed for treatment.

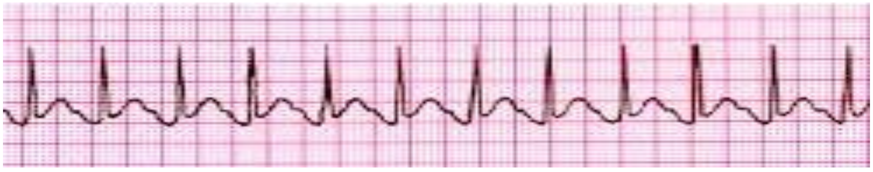

Figure 2.18 An example ECG strip with sinus tachycardia

SA Block: It is divided into three subgroups according to severity:

1. First-degree sinoatrial block: The time interval from the discharge of the electrical impulse in the SA node to start of atrial depolarization is prolonged. It is hard to be discerned due from ECG due to the discharge of impulses in SA node is not perceptible.
2. Second-degree sinoatrial block: It is divided into type I and types II. In type I, there is a delay in the conduction from the SA node to the atrium. Therefore, the P-P interval decreased. In type II, SA impulses are blocked occasionally and the rate is slower than normal due to some unreached pulses to atria (see Figure 2.19).
3. Third-degree sinoatrial block: There is no conduction between the SA node and the atrium. So the conservation of cardiac rhythm will depend on the latent pacemaker. The rhythm is referred to as an escape rhythm, atrial myocardium, the junctional area or in the His-Purkinje network. However, it cannot be detected from surface ECG.

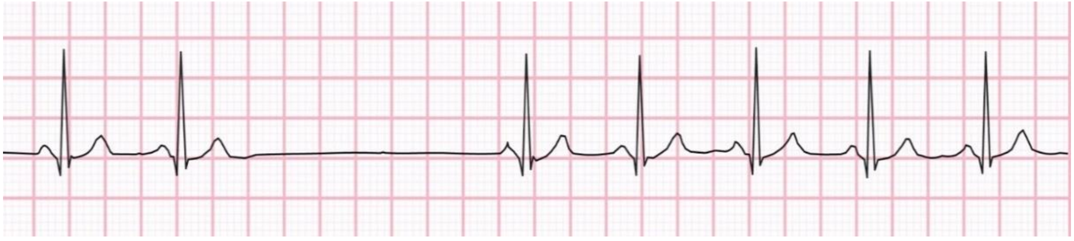

Figure 2.19 An example ECG strip with 2nd degree Type II SA block

Sinus Arrest: It occurs when the sinoatrial node of the heart cases to generate impulses (see Figure 2.20). If there is no electrical activity, there will be neither depolarization nor contraction. The parasympathetic activity or sinus node disease can cause sinus arrest. The blood pressure will be dropped by a longer absence of electrical activity. There is a need for medication to increase S-A node activity. On the other hand, there is an emergency need for a pacemaker, if the underlying reasons cause myocardial infarction.

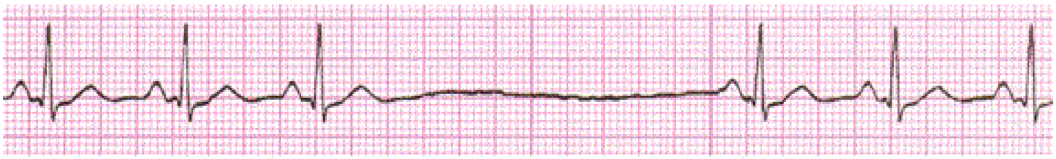

Figure 2.20 An example ECG strip with sinus arrest

Sick Sinus syndrome: It is also known as sinus dysfunction that sinus node which is the natural pacemaker of the heart, does not accomplish properly (see Figure 2.21). Generally, the sinus node produces a balanced pace of regular impulses whereas the signals are abnormally in sinus node syndrome and heart rhythm can be too fast (Tachy) or too slow (Brady). There is a need for both pacemaker and medication for treatment of tachycardia and bradycardia.

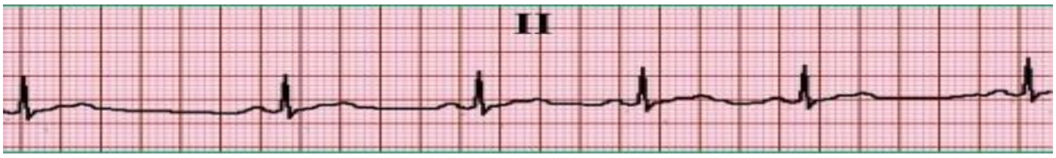

Figure 2.21 An example ECG strip with sick sinus syndrome

AV block: AV block is grouped into three groups according to intensity:

1. Prolonged P-R Interval – First Degree Block: If the P-R interval increases to greater than 0.20 second, it is said to be prolonged P-R interval and the patient have first-degree heart block (see Figure 2.22). So first-degree block is defined as a delay in conduction from the atria to the ventricles but it is not defined as an actual blockage of conduction.
2. Second Degree Block: It is divided into two different subgroups type I, also called Wenckebach, and type II, also called Mobitz. In type I second-degree block, the

PR interval expands with each beat until the impulse is not conducted to ventricles and the QRS complex is dropped (see Figure 2.23). In type II second degree block, the blockage occurs at the His bundle and the bundle of branches. Some symptoms are light-headedness, presyncope and, syncope and in high-grade AV block, a pacemaker is indicated and the PR interval is constant and QRS complexes are dropped (see Figure 2.24).

3. Complete A-V Block – Third-Degree Block: If condition causing poor conduction in the A-V node becomes more severe, it will be a complete block of the impulses from the atria into the ventricles. So the ventricles spontaneously set up their own signal which is originating in the A-V node. In ECG waveform, P waves become dissociated from the QRS-T complexes (see Figure 2.25).

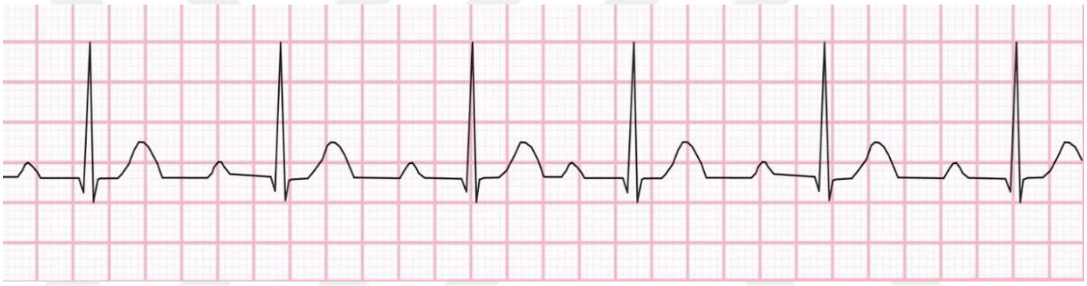

Figure 2.22 An example ECG strip with 1st degree AV block with long PR interval. PR interval is increased

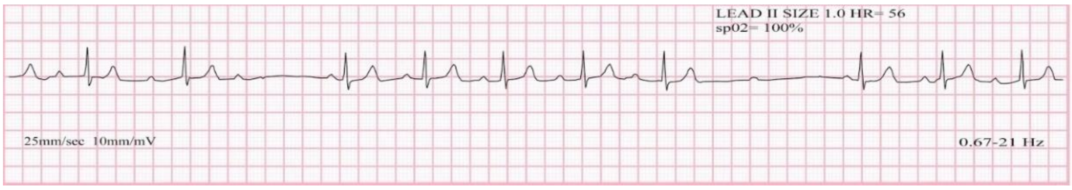

Figure 2.23 An example ECG strip with second degree AV block type 1(Wenckebach)

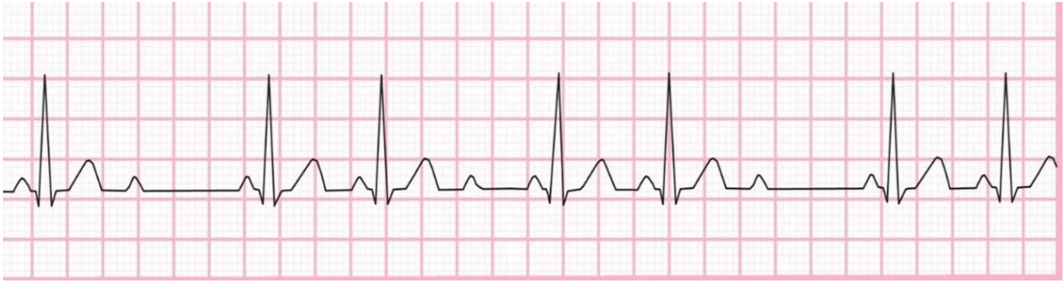

Figure 2.24 An example ECG strip with second degree AV block type two (Mobtiz)

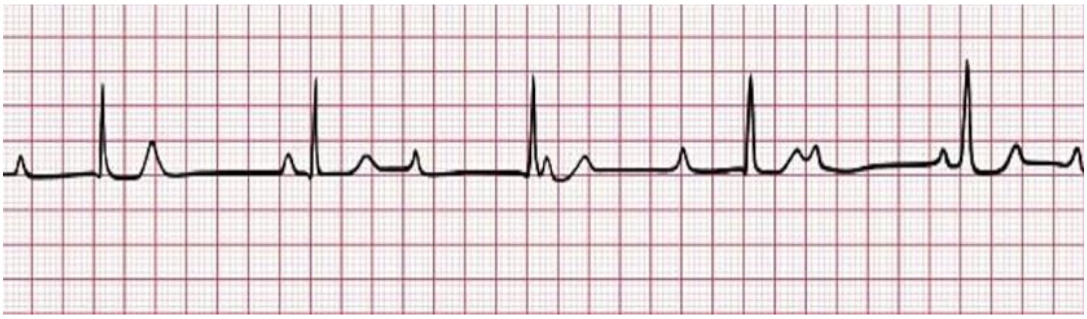

Figure 2.25 An example ECG stripe with complete A-V block

Premature Atrial Contractions (PAC): It occurs due when the sinus node discharged late in the premature cycle that made the succeeding sinus node discharge also late in appearing. In ECG records, the P wave occurs soon in the heart cycle and the P-R interval is decreased which is indicated that the ectopic trigger origin of the beat is in the atria near the A-V node (see figure 2.26).

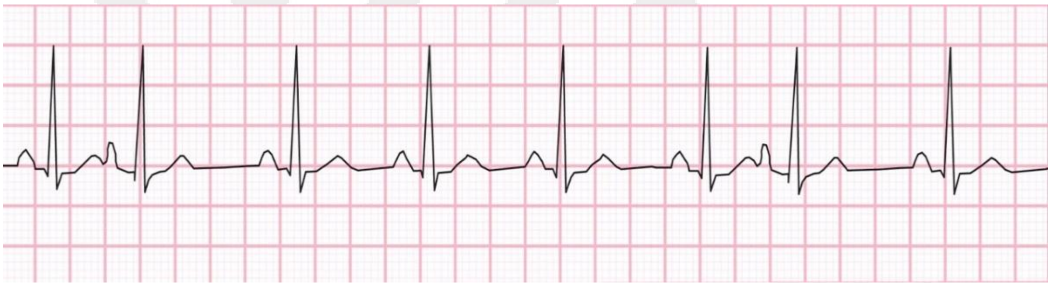

Figure 2.26 An example ECG strip with PAC. The second and the 7th are PACs

Atrial Tachycardia: The impulses come from an ectopic pacemaker in the atria rather than SA node and the rate is between 140-250 bpm. Some symptoms are dizziness, syncope, chest pain and palpitation, shortness of breath (see Figure 2.27).

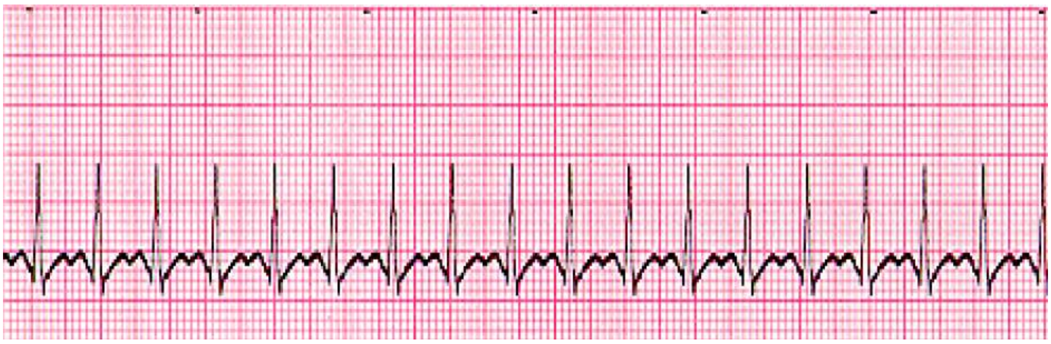

Figure 2.27 An example ECG stripe with atrial tachycardia. Abnormality of P wave

Atrial Flutter: The electrical impulse transmits along a pathway within the right atrium. The beat is produced faster than 250 bpm and it is nor coordinated with ventricular (see figure 2.28). some symptoms may include heart palpitation, shortness of breath, dizziness and, discomfort in the chest.

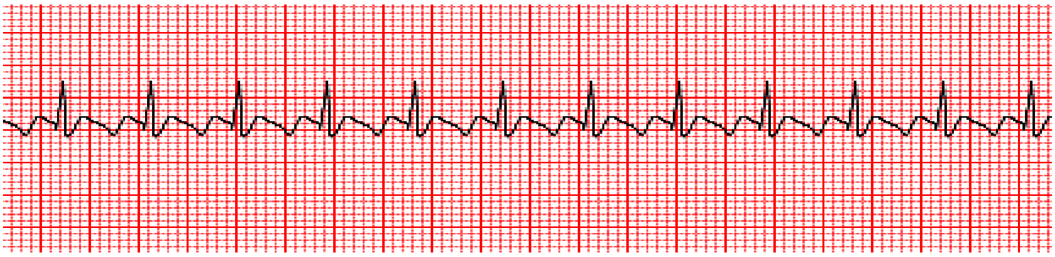

Figure 2.28 An example ECG stripe with atrial flutter. Sawtooth pattern is noticeable

2.4.3 Atrial Fibrillation

Atrial fibrillation or supraventricular is the main interest of this thesis. The mechanism of atrial fibrillation is an irregular heartbeat. The heart’s two upper chambers beat irregularly and out of coordination with the ventricles. A frequent cause of AF is atrial enlargement resulting from heart valve lesions. It prevents the atria from emptying into the ventricles or from ventricular failure with increase damming of blood in upper chambers. The dilated atrial walls make an ideal condition of a long conductive pathway as well as slow conduction. Both of them make susceptible to atrial fibrillation. The rate of AF is 300-650 bpm which cannot be conducted to the lower chambers due to the recovery period of AV node. Therefore, the ventricles have abnormal depolarization and heart rate might be 100-175 bpm (Webster, 1995). Some symptoms of AF are palpitations, dizziness, chest pain, confusion and sometimes with no symptoms (Cown et al., 2014). ECG record with AF with the absence of P wave and irregular QRS complexes is represented (see Figure 2.29).

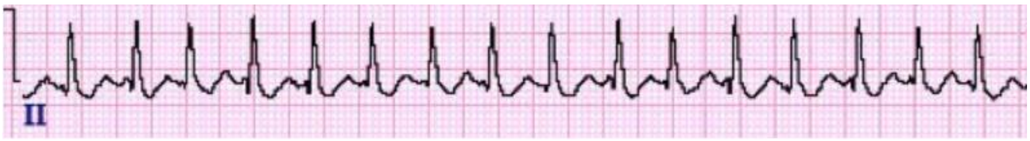

Figure 2.29 An example ECG with AF

The atrial fibrillation is divided into three subgroups because of different treatment (Camm et al., 2010; Fuster et al., 2006; Gallagher & Camm, 1998; Lévy et al., 2003; Lip & Tse, 2007):

1. Paroxysmal Atrial Fibrillation (PAF): An irregular rhythm that takes along within 48 hours and ceases by itself then heart rate returns to normal. It may last for seconds or minutes. There is sinus rhythm on ECGs strip with PAF (Hoshino et al., 2013; Page et al., 1994).
2. Persistent Atrial Fibrillation: Thus type of AF lasts more than 7 days. Medication and cardioversion are needed to aid the heart return to normal rhythm.
3. Permanente Atrial Fibrillation: Neither medication nor cardioversion can correct the heart rate.

In paroxysmal atrial fibrillation, atria contract uncoordinated with ventricles which are caused reduce the transfer of blood to the ventricles completely. Therefore, remaining blood inside the atria can lead to form clots. Forming clots in the aria, increasing the risk of stroke. Studies elicit that the stroke risk of PAF patients is identical to persistent/permanent AF patients (Hohnloser et al., 2007; Lip & Li Saw Hee, 2001).

The three goals of paroxysmal atrial fibrillation are (Lip, 1999):

1. Maintaining long-term sinus rhythm and conquering paroxysms of atrial fibrillation with proper medication
2. Controlling heart rate by increasing the refractoriness of AV node with suitable medication and also ventricular rate can be controlled by implanting pacemaker (Heist, Mansour, & Ruskin, 2011)
3. Intercepting the paroxysmal atrial fibrillation complications; i.e. stroke with antithrombotic therapy.

CHAPTER THREE
HEART RATE VARIABILITY ANALYSIS

3.1 Introduction

Heart rate variability (HRV) is an approach for time-series analysis which can be useful when analyzing biosignals particularly in studying the cardiovascular blood pressure regulatory system. Biosignal time series, are very difficult to analyze due to the thousands of data points, which may be impossible to achieve and if the time series is long, significant changes in biological system will occur and the time series will start to drift in the parameter space of the system because of both internal and external influences. Other disturbing factors in such signals are the notable noise level, strongly separated of the signals and regular strong periodic signals.

HRV analysis is the prominent point of this thesis. The features on which PCA algorithm were run had previously been obtained from HRV analysis by Hilavin et al (2015).

3.2 ECG Recording and Generation of RR interval

3.2.1 RR Interval Time Series

HRV analysis that is based on the RR interval time series, the result of intervals between fiducial points of R peaks of QRS complexes in the ECG. The RR interval time series is actually an event series not continuous signal that can be noted when performing frequency domain analysis. HRV does not measure the rhythm of SA node because it is not based on P-P intervals, it indicates fluctuations in the AV conduction superimposed on the P-P interval while beat to beat changes in RR intervals indicate the variability of the SA node quite an accuracy. Practically, it is better to use R-R intervals for HRV analysis due to the low amplitude of the P wave, especially in presence of noise.

3.2.2 Time and Amplitude Resolution of ECG Recording

The important parameters of HRV are the sampling frequency and amplitude of ECG recording. The sampling frequency not only determines the temporal resolution of R peak recognition of QRS complexes but also determines the accuracy of the measurement of RR intervals. The frequency of 200 Hz is enough because too low sampling rate produces noises and errors in all HRV measures and, higher sampling rates do not give better results. In some cases, such as heart failure, if the overall variability of RR interval is low, HRV analysis needs a higher sampling rate due to the changes in the length of RR intervals could not be taken below that resolution.

The amplitude resolution of the analog-digital conversion ECG signal may not be as a limiting factor. The 12 bits amplitude resolution that is commonly used is enough.

3.2.3 Duration of the Recording

The duration of the recording is determined based on the aim of the study and stationarity issues. Generally, frequency domain methods are used for short-time measurements while time domain methods are used for long-term measurements of HRV. Non-linear methods are appropriate for short-time and others for long-term analysis. It is necessary to be noted that if the same mathematical method is used for the analysis of both short-term and long-term ECG recording, the physiological interpretation can be different. It is not appropriate to compare the HRV measures obtained from recordings of dissimilar durations, with each other.

3.2.4 Stationarity of the Recording

Stationary means that the all parameters which define the working point of the system remain constant in time. It is strongly linked to the duration of the recording. Therefore, the longer recording is less stationary than shorter recording because of the changes in the physiological state of the subject. The best approach for the analysis of long recording is to divide the recording into shorter segments and perform discrete analysis on them.

3.2.5 Removing Trends

The trend in RR interval time series can be seen easily and also interpreted as a sign of nonstationary, which can be possibly removed by subtracting the trend from the data. Practically, if there is no significant trend in the data irrespective of its origin and circumstances, HRV analysis method will be useful. It is necessary to remove the trend from the data. The linear trend is removed commonly because the removal of non-linear trends may cause significant bias. From the point of view of spectral analysis, the contribution of the lowest frequencies will be decreased by removing of the trend, therefore further analysis is concentrated on faster oscillations. It should be noted that the removal of a trend does not restore the stationary of data but the only way that can overcome the problem of non-stationary is to reduce both internal and external disturbances during the study.

3.2.6 Ectopic Beats, Arrhythmias and Noise

The ECG signal can contain QRS complexes originating outside the SA node which is referred to errors because HRV analysis is based on the assessment of the variability of sinus rhythm. Since errors can affect the HRV analysis, it cannot be ignored.

Technical errors, such as missed beats or electrical noise, can be edited by cleaning the data by a sufficient interpolation based on successive QRS intervals. Cumulative time is not changed during interpolating the RR interval time series.

The editing of ectopic beats is difficult due to the ectopic beats produce a very short RR interval followed by a delay and prolonged RR interval. It can be edited by lengthening the first interval and shortening the second interval. The sequences of the ectopic beats and arrhythmias are stroke volume reduction and cardiac output that are leading to temporary drops in the blood pressure. Therefore, it activates autonomic reflexes and convinces changes in the efferent autonomic activity. These physiological responses can last 10-30 beats, so the editing of the two intervals on either side of beats does not remove all of the changes in HRV.

The best option to HRV analysis is to select an error-free recording otherwise editing can be considered.

3.3 Time Domain Analysis

The standard deviation of RR intervals is the most common time domain estimate of HRV. Normal to the normal deviation of intervals (SBNN) measured between following sinus beats. SDNN can be calculated over 5 min segments which are called SDANN or over 24 hours. These are not compatible due to the HRV is not stationary process in which the mean and the variance are independent of the record length. The low-frequency (LF) variation impart a significant proportion of the overall HRV power and also to the SBNN in long-term recordings. Against the effect of HRV reduction at higher levels of the heart rate, SDNN can be normalized by dividing it by the mean of RR interval.

The most common used HRV measures are based on the differences between RR intervals. For instance, the root mean square of successive differences of RR intervals (RMSSD), the number of pairs of neighboring RR intervals differing by more than 50 ms (NN50 count), and the ratio of the NN50 count to the count of all RR intervals demonstrate as a percentage (pNN50). They elicit high frequency (HF) variations of heart rate and are almost independent of long-term trends because all of the measurements use RR intervals differences.

The advantages of time domain HRV estimates are easily calculated, do not require need time-consuming computation and stationary. The main disadvantage of time domain methods is poor discrimination between effects of sympathetic and parasympathetic autonomic branches.

3.4 Frequency Domain Analysis

The main idea of frequency domain analysis of HRV is the examination that HRV is composed of clearly defined rhythms that are related to different regulatory mechanisms of cardiovascular control. There is a need for advanced analysis methods, such as power spectral density (PSD) analysis, in order to get more specific information on the dynamics and frequency components of HRV. PSD decomposes the signal into its frequency and power. There are two methods for computing the PSD function: Fourier transform and Autoregressive (AR).

3.4.1 Fourier Transform

The Fourier transform converts the time domain data into frequency domain and back. It is a one-to-one transform which means no information is lost or added. The original Fourier transform should be replaced with the discrete version of the transform in the case of RR interval time series.

$$X(f_n) = \Delta \sum_{k=0}^{N-1} x(t_k)e^{-2\pi i f_n t_k}$$

$$= \Delta \sum_{k=0}^{N-1} x(t_k)e^{-\frac{2\pi i k_n}{N}}, \ where\ f_n = \frac{n}{N\Delta} \qquad (3.1)$$

The discrete Fourier transform has several important features:

1. The frequency scale is discrete because only f_n components are possible
2. The resolution of frequency scale depends inversely on the number of data samples and interval Δ.

3.4.2 Autoregressive Modeling

“The Autoregressive (AR) model is based on the idea that the feature values of a time series depend linearly on the previous values” (Akaike, 1969). It is totally generic and can be used to model large sets of different systems. The AR is defined by the following equation:

$$x_k = \sum_{j=1}^{p} a_j x_{k-j} + u_k \tag{3.2}$$

Where the x_k is the sample and, a_j is the model parameter and p is the model order. u_k represents the noise. The important task is to find optimal p while the a_j of the model delineates the system as well as possible. If we have such a model, it can be computed the corresponding PSD from the following formula:

$$PSD(f) = \frac{a_0}{\left|1 + \sum_{j=1}^{p} a_j z^j\right|^2}, \quad z = e^{-2\pi i f \Delta} \tag{3.3}$$

Where f is the frequency and Δ is the sampling interval. The Equation 3.3 has the ability to model sharp peaks because all free parameters are in the denominator. There are no restrictions for the maximal number of frequency since the AR spectrum is based on a modeling approach. PSD can be assessed utilizing as high a frequency resolution as one desires. By contrast, the Fourier transform that the frequency is defined by the number of samples and also AR can resolve the frequency location of each peak.

3.4.3 Resampling

The data can be described as a hypothetical continuous function in which the data has been sampled unevenly in time at the moments of R peaks and also the input data has been assumed to be the data that sampled unevenly in time in both Fourier transform and AR modeling analysis. The original RR interval should be converted to a form by interpolating every interval and also resampling the evidently continuous function. Interpolation can be linear or splines. The approach of interpolation does not have a difference in HRV analysis.

The frequency of resampling should be higher than the effective sampling rate of the RR interval data which is equal to the mean of heart rate around 60 bpm (1Hz). Practically, a resampling rate of 2-5 Hz is sufficient.

3.4.4 Spectral Power

The area under the PSD curve from zero to the highest frequency represents the total power (TP) of the RR interval data and it is equal to the variance of the signal. The spectral power is divided into three frequency band typically in the short-time spectral analysis:

1. High frequency (HF; 015 – 0.40 Hz): It corresponds to heart rate variations that are related to the respiratory sinus arrhythmia and fluctuations are mediated by fluctuations of efferent parasympathetic activity.
2. Low frequency (LF; 0.04 – 0.15 Hz): It is important during characterizing baroreflex sensitivity utilizing spectral approaches. The sympathetic nervous system has a central role on the L, but fluctuations are influenced by the sympathetic nervous system.
3. Very low frequency (VLF; 0 – 0.04 Hz): Oscillations at frequencies are related to the vasomotor of thermoregulation or hormonal system.

The unit of PSD function depends on the unite of RR interval time series, so the magnitude of HRV is defined as power and its unit is ms^2 or s^2. Due to the various frequency band among healthy subjects of equal age, comparing the two subjects can be deceptive. At the sequence of this, powers are defined in normalized units by dividing any power by TP less VLF power.

An example of HRV analysis is represented in Figure 3.1. It is performed by FFT and AR modeling. It consisted of 349 intervals with the length of 5 min RR interval and the mean of RR interval was 856 ms which is corresponding to a heart rate of 70.3 bpm. The linear trend was eliminated and interpolation and resampled was done before analyzing.

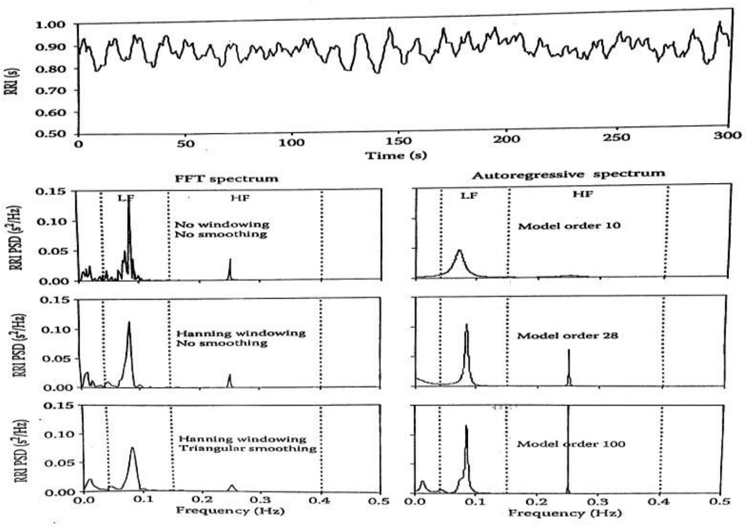

Figure 3.1 An example of RR interval series and corresponding FFT and AR spectrum

3.4.5 Effects of Respiration

The respiratory element of HRV depends on the breathing tone and its frequency. The HF component is in direct proportion with breathing volume that means the HF component decreases as breathing volume decreases. In addition, if the power of respiration peak increases, its frequency will be decreased. The HF element cannot be measured as HF component if the breathing frequency is below 0.5 Hz. In this case, it will be as a part of LF component. In Figure 3.2, the effect of breathing frequency on HRV is shown. In the upper right panel in Figure 3.2, when the breathing frequency was fixed at 0.10 Hz the amplitude of the respiratory was 0.75 Hz. (the upper left panel) and it is noted that the breathing volume was fixed in all exhibitions.

In the situations that the breathing frequency and volume are controlled, the HF element can be used as an estimate pf parasympathetic and vagal activity. Fixing respiration rate can decrease stress level that can affect the function of the autonomic nervous system. Therefore, it interferes with HRV. Using a respiratory frequency which is close to normal breathing rate, can decrease the stress and it can be done by first measuring the breathing rate and adjusting the frequency of metronome. Alternatively, white noise breathing can be used (see the right bottom panel in Figure 3.2).

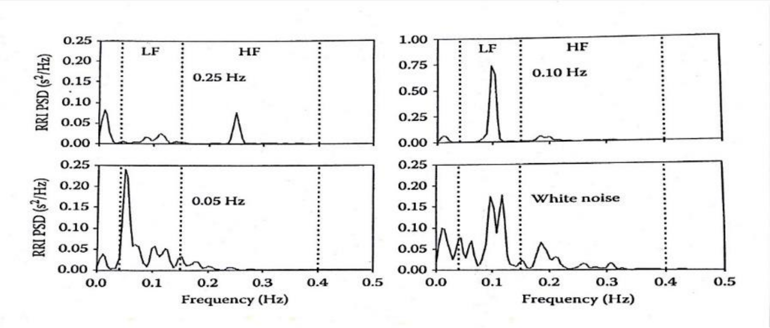

Figure 3.2 RR interval FFT as three different breathing frequencies of 0.25, 0.10, 0.05, and white noise breathing

3.4.6 Time-Frequency Analysis

Analyzing spectral methods in steady state are not useful. Because they cannot define exactly when interpretative frequency elements change in amplitude or frequency. The spectral features of the slowly changing system can be analyzed in deep if the information about temporal variations of frequency elements are available. There are many different methods to achieve a time-frequency analyzing such as Windowed Fourier Transform, Wigner-Ville Distribution, Complex Demodulation. All of them are based on various integral transforms of the original data. The principle disadvantage on all of them is the trade-off between temporal resolution and statistical reliability of spectral elements and it is not able to be define the amplitude of the certain frequency in specific local due to there is a need to a time window to measure it. A shorter time window can represent more quick changes within the power spectra. On the other hand, a longer time window can represent a more valid estimate of the spectral power.

3.5 Non-linear Analysis

In frequency domain analysis, certain predefined pattern used for recognition such as Fourier transforms which is the pattern is a sinusoidal wave and wavelet which is the pattern is in certain wavelet function. There is another alternative way of distinguishing the variability of heart rate with the measuring of the regulatory or complexity of the fluctuation.

3.5.1 Approximate Entropy and Sample Entropy

Entropy is an approach for quantifying the regularity of the data (Pincus and Goldberger, 1994). there are many processes to define the entropy of a time-series, but most of them required the noise-free data and long recordings. Approximate entropy (ApEn) is one of the methods for measuring the complexity of short-time series and also it is useful for HRV analysis because noise-free recording is hard to obtain (Pincus and Goldberger,1994) and it does not base on specific assumption.

In ApEn calculating, in the first step we construct the pseudo phase vectors from the beginning time series $x(i)$, in which i=1 …N, N represents the number of data point

$$u(i) = [x(i), x(i+1), x(i+2), \dots, x(i+m-1)], \qquad (3.4)$$

m refers to embedding dimension, $u(i)$, vectors $u(i)$ can be defined as m-point patterns. First, we choose one m-point pattern and look for similar m-point patterns. The similarity of two patterns is based on the maximum distance d between the components is less than the tolerance r (see Figure 3.3).

$$d[u(i), u(j)] = \max\{|u(i+k) - u(j+k)| : 0 \le k \le m-1\} \le r, \qquad (3.5)$$

The number of close vectors $u(j)$, which are at a distance r from $u(i)$, is

$$C_i^{(m)}(r) = \frac{\{the\ number\ of\ index\ j\ for\ which, j \le N-m+1, d[u(i), u(j)] \le r\}}{N-m+1} \qquad (3.6)$$

The maximum value of C is 1 due to normalization and also it can be considered as the probability of detecting similar m-point patterns. The same analysis can be accomplished for all m-patterns. The logarithmic average of probabilities over all m-patterns is

$$\emptyset^{m}(r) = (N-m+1)^{-1} \sum_{i=1}^{N-m+1} lnC_i^m(m) \tag{3.7}$$

Approximate Entropy is defined as

$$ApEn(m,r,N) = \emptyset^{m}(r) - \emptyset^{m+1}(r) \tag{3.8}$$

ApEn is depended on three parameters which are the length m of the vectors that are being compared, the tolerance and the number of data points. It elicits that the direct comparisons always need for fixing of parameters. The m=2 is the ordinarily used value in HRV analyzing. ApEn methods its final value asymptotically by the increasing of the number of data. N>900 and m=2 give a valid result practically.

ApEn is very sensitive to the smallest trends in data since comparison of patterns is based on the absolute values of data. The trend can be eliminated before ApEn. Alternatively, slow trends can be eliminated by using the differentiated data or the difference of successive RR interval whereas it behaves like a high-pass filter in the frequency domain. If the tolerance parameter is constant, ApEn will not sensitive to changes in single data values although the situation could well be different if the tolerance parameter is bounded to the SD especially ectopic beats that if not edited, they can change the SD remarkably. The similarity of the vector to itself is included in the calculation during the calculation the number of similar vectors to get ApEn. This makes certain that $C_i^{(m)}(r)$ is non-zero which is necessary for calculating the logarithm and this causes ApEn to give a result that implicates greater regularity of the signal that may be present.

Sample Entropy (SampEn) is calculated for removing the bias (Richman and Moorman, 2000). A comparison to the vector itself is prevented:

$$C_i'^m(r) = \frac{\{the\ number\ of\ index\ j\ for\ which, j \neq i, j \leq N-m+1, d[u(i),u(j)] \leq r\}}{N-m+1} \tag{3.9}$$

The average of probabilities ∅ is also explained without logarithms

$$\emptyset^m(r) = (N-m+1)^{-1} \sum_{i=1}^{N-m+1} C_i'^m(r) \tag{3.10}$$

Now SamEn defined as

$$SampEn(m,n,N) = \ln\left(\frac{\emptyset^m(r)}{\emptyset^{m+1}(r)}\right) \tag{3.11}$$

The explanation and utilize of SampEn stand exactly the same as for ApEn. On the other hand, the dependence on *r* (tolerance parameter) and *N* (data points) are different. ApEn is maximum in the certain value of *r* while SampEn decreases as *r* increases. SampEn basically is independent of *N* while with small values of *N*, its statistical reliability is poor. SampEn and ApEn have the same result when *r* and *N* are large enough. SampEn supplies a more reliable estimate of the complexity of a signal in comparison to Apen and it can be used for considerably shorter time series than ApEn, (<200 points).

3.5.2 Detrended Fluctuation Analysis

Analyzing a long-standing time series which are lasting several hours, identification of repeated pattern is not the best method for examining the HRV data. The represented method is to distinguish the internal correlations of the signal which is scaling properties and fractal structures. Detrended Fluctuation Analysis (DFA) presents a probability for distinguishing this as a function of correlation distance (Peng et al., 1993,1995; Iyengar et al., 1996).

For calculating DFA, first, we should construct an integrated version of the original time series $x(i)$, where *i=1…N,* that gives us

$$y(k) = \sum_{i=1}^{k} (x(i) - < x >) \tag{3.12}$$

Where <*x*> referees to the mean of the original time series and k=1…N. Secondly, we divide the time series y(k) into spaced segments with length *n* (see Figure 3.3). We calculate the local trend for each segment by fitting a regression line $y_n(k)$ to the segment. Removing the linear trend of each segment calculates the root-mean-square fluctuation of the integrated time series. Therefore,

$$DFA(n) = \sqrt{1/N \sum_{k=1}^{N} [y(k) - y_n(k)]^2} \tag{3.13}$$

In the final statement, we should take into account that $y_n(k)$ should be updated when operating into the next segment when the index *k* is stepped. Typically, DFA is increased by increasing the segment length. If logarithm DFA increases as a function of log(n), the time series will follow the scaling law, and in this situation, the slope α of the linear changes the scaling exponent that defines the type of scaling.

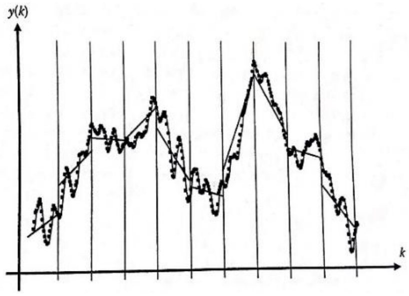

Figure 3.3 The integrated time series and the local trends

3.5.3 Correlation Dimension

The dynamics of a system can be defined by measuring its attractor dimension or the path toward which the system converges. In the chaotic system, the attractor can be fractal, so its dimension is not an integer. Having knowledge about the dimension of the attractor can also be helpful to get useful information about the traces of underlying systems. The correlation dimension (CD) is one of the easiest approaches for estimating the attractor dimension (Grassberger and Procaccia, 1983; Kanz and Schreiber, 1995; Yum et al., 1999) and it also referred to with the designation D2.

The starting point for the calculation is in the modification of the time series in the multi-dimensional phase space $x(i)$, where $i = 1 \dots N$, by utilizing the vectors of the pseudo phase space $u(i) = [x(i), x(i+1), x(i+2), \dots, x(i+m-1)]$, where m is the embedding dimension. Then we calculate how many attractor points are at a distance r as measured from observation points for each vector $u(i)$

$$C_i^m(r) = \left\{ the\ number\ of\ index\ j\ for\ which, j \leq N-m+1, \frac{d[u(i), u(j)]}{N-m+1} \right\} \tag{3.14}$$

where the *d* is defined as a normal Euclidian distance

$$d[u(i), u(j)] = \left(\sum_{k=1}^{m} |u(i;k) - u(j;k)|^2\right)^{1/2} \tag{3.15}$$

and then we calculate the mean of quantities $C^m(r)$ over all vectors that we compute the so-called correlation integral

$$C^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} C_i^m(r) \tag{3.16}$$

CD is described as a limit

$$CD(m) = \lim_{r \to 0} \lim_{N \to \infty} (\log C^m(r) / \log r) \tag{3.17}$$

Practically, with limited data sets, the limits cannot be calculated confidently. Therefore, the CD is described as the slope of the regression line which is calculated from a log-log representation.

The embedding dimension m should be selected when calculating the CD so it is at least 2D. The number of data points should be over 10^m to describe the attractor accurately. The calculation of the valid CD for biosignals is hard to achieve because it is not possible to discover such a range of the distance r, in which $\log C^m(r)$ alters linearly as a function of $\log r$ due to the noise included in the data and non-stationary of the data, but calculation of the correlation dimension can still be functional. Optimistic results have been achieved by utilizing m=20 and by probing for the mean slope within 0.01<$C^m(r)$<0.1 which is called *modified correlation dimension*. This cannot explain the actual dimension of the system but it gives a measure of the complexity of the system when CD and the system come to be more complicated.

3.5.4 Return Map

Dynamical systems are defined by a group of differential equations. If the variables obtain values only at specific moments in time in the cases with the RR interval time series, the differential equations can be substituted with discrete equations, for instance,

$$\begin{aligned} x_{i+1} &= F(x_i, y_i, z_i) \\ y_{i+1} &= G(x_i, y_i, z_i) \\ z_{i+1} &= H(x_i, y_i, z_i) \end{aligned} \tag{3.18}$$

where x, y, z and etc are dynamical variables of the system. $F, G, H, \ldots$ are functions which describe the dynamics. These functions are not recognized, but we can attempt to solve for them by inspecting the measured time series. The equation can be clearly expressed if there is one variable:

$$x_{i+1} = F(x_i) \tag{3.19}$$

This expression joints the new value x_{i+1} to its predecessor value x_i. The function F can be solved principally by pairing successive values of the time series for $i = 1$ to $N - 1$ and then plot them on a two-dimensional graph which is called a return map.

If the dynamical system is not one dimensional, a return map cannot solve functions F, G and so on. Even in this cases, a single variable return map may prove to be utilized and it is a certain type of projection of the multidimensional system into one dimension. Figure 3.4 elicits the return map of an RR interval time series. The points on the graph scatter in order to configure an ellipsoid but it can also configure complex structures. If the return map is an ellipsoid, it can be characterized by two quantities: first the SD in the direction of the diagonal SD2 and second the SD in a direction perpendicular to SD2. These deviations measure the variability because they calculate the movement of the system in a phase space. However, the parameters would not describe the variability very well if the return map has a complicated shape.

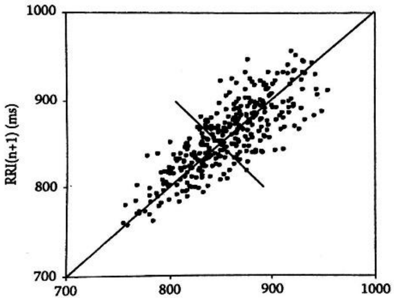

Figure 3.4 A return map from an RR interval time series with SD along the diagonal SD2=5 ms and SD perpendicular to the diagonal SD1=18 ms

CHAPTER FOUR
FEATURE EXCTRACTION WITH PCA

We have many variables in machine learning. The higher number of features, the harder it to visualize and work on it. Sometimes, some of these features are redundant and correlated and this is where the dimensionality reduction become active. It means that the transformation of high dimensional data into reduced dimensionality that corresponds to the constitutional dimensionality of the data. The minimum number of features or dimensions that we need to observe the properties of the data are referred to the constitutional dimensionality of data (Fukunaga, 2013). Dimension reduction is referring to process of transmitting a set of data into the low dimension space. Also, it is necessary that the new space conveys similar information concisely and it is important in many areas since it makes easy visualization and classification (Jimenez & Landgrebe, 1998).

The advantages of the feature extraction are (Guyon & Elisseeff, 2003):

- Facilitating data understanding
- Reducing training time
- Reducing the memory requirements
- Improving the precious of measurement and prediction

There are two components of dimension reduction: feature selection and feature extraction. Feature selection approaches refer to a process whereby a data space is transformed into a feature space that, in theory, has exactly the same dimension as the original data space while feature extraction reduces the dimension of data and transform it to a lower dimension space.

Feature selection methods are involved in three ways:

- Filter: It acts as preprocessing to rank the features
- Wrapper: It finds a subset that gives the highest performance in prediction
- Embedded: It selects variables as part of the training process without dividing the data into training and testing sets (Chandrashekar & Sahin, 2014).

Feature extraction approaches used for dimensionality reduction include:

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Generalized Discriminant Analysis (GDA)

The primary focus of this thesis is the Principal Component Analysis transformation to compare the results in PAF patient's detection.

4.1 Principal Components Analysis (PCA)

A key problem in statistical pattern recognition is feature extraction. On the other hand, the transformation is designed in such a way that the data set would be represented by a reduced number of effective features and maintain most of the intrinsic information content of the data. Principal Components Analysis, which is also known as the Karhunen- Loève transformation, maximizes the rate of decrease of variance. Principal Components Analysis (PCA) is the oldest technique in multivariate analysis.

Let X denoted a p-dimensional random vector representing the data and also we assume that the mean of the random vector C is zero.

$$E[X] = 0$$

E is referring to statistical expectation operator. If X has a nonzero mean, so we subtract the mean from it. Let u elicit a unit vector, also of a dimension p, onto the vector that X is to be projected. This projection is described by the inner product of the vectors X and u that is shown by

$$a = X^T u = u^T X \tag{4.1}$$

Subject to the constraint

$$\|u\| = (u^T u)^{1/2} = 1 \tag{4.2}$$

The projection a is a random variable with a mean and variance related to the data vector X. Under the presumption that the random data X has zero mean, it accompanies that the mean value of the projection a is zero too:

$$E[a] = u^T E[x] = 0 \tag{4.3}$$

The variance of the a is the equal as its mean-square value so:

$$\begin{aligned} \sigma^2 &= E[a^2] \\ &= [(u^T X)(X^T u)] \\ &= u^T E[XX^T]u \\ &= u^T R u \end{aligned} \tag{4.4}$$

The p-by-p matrix R is the correlation matrix of the data vector. It is described as the expectation outer product of the vector X by itself, as shown by:

$$R = E[XX^T] \tag{4.5}$$

The correlation matrix R is symmetric that means:

$$R^T = R \tag{4.6}$$

From this property, it accompanies that if a and b are any p-by-1 vectors, so:

$$a^T Rb = b^T Ra \tag{4.7}$$

From Eq.(4.4) we can say that the variance σ^2 of the projection a is a function of the unit vector u, therefore we can write:

$$\begin{aligned}\Psi(u) &= \sigma^2 \\ &= u^T R u \end{aligned} \tag{4.8}$$

We may be of the view of $\Psi(u)$ as a variance probe (Lowe, D & A.R. Webb, 1991b; Personnaz & Guyon & Dreyfuse, 1985).

4.1.1 Eigenstructure of Principal Components Analysis

The next subject that should be considered is that of looking the unit vectors u along which $\Psi(u)$ has stationary values, subject to a restriction on the Euclidean norm of u. The solution to this problem is in the eigenstructure of the correlation matrix R (Preisendorfer, 1988). If u is a unit vector such that the $\Psi(u)$ has stationary value, so for any small perturbation δu of the unit vector u, we discover that to first-order in δu,

$$\Psi(u + \delta u) = \Psi(u) \tag{4.9}$$

From the definition of the variance probe in Eq. (4.7) we have:

$$\begin{aligned}\Psi(u + \delta u) &= \Psi(u + \delta u)^T R(u + \delta u) \\ &= u^T R u + 2(\delta u)^T R u + (\delta u)^T R \delta u \end{aligned} \tag{4.10}$$

Where, in the second line, we have made use of Eq. (4.7). by ignoring the second-order term $(\delta u)^T R \delta u$ and involving the definition of Eq. (4.7) we can write:

$$\begin{aligned}\Psi(u + \delta u) &= u^T R u + 2(\delta u)^T R u \\ &= \Psi(u) + 2(\delta u)^T R u \end{aligned} \tag{4.11}$$

Hence, the use of both Eq. (4.9) and Eq. (4.10) implies that

$$(\delta u)^T R u = 0 \tag{4.12}$$

Any perturbation δu of u are not allowable; rather, we are limited to utilize only those perturbations that the Euclidean norm of the perturbed vector $u + \delta u$ maintains same to unity; that is,

$$\|u + \delta u\| = 1$$

Or,

$$(u + \delta u)^T (u + \delta u) = 1$$

So, in Eq. (4.2), we need to the first order in δu,

$$(\delta u)^T u = 0 \tag{4.13}$$

This elicits that the perturbation δu should be orthogonal to u, and one alter in the direction of u is allowed.

Practically, the elements of the unite vector u have no dimensions in physical perception. If we combine Eq. (4.11) and Eq. (4.12), we must suggest a scaling factor λ into the letter equation. λ has same dimensions as the entries I the correlation matrix R. Performing all of this, may then write:

$$(\delta u)^T R u - \lambda(\delta\, u)^T u = 0$$

$$(\delta u)^T\, (R u - \lambda u) = 0 \tag{4.14}$$

For the situation of Eq. (4.13) to hold, it is essential and enough that we have

$$R u = \lambda u \tag{4.15}$$

This Equation rules the unite vectors u for which the variance probe Ψ(u) has stationary values. Eq. (4.14) is identified as the eigenvalue problem that is encountered in linear algebra. The issue has not trivial solutions (i.e., $u \neq 0$) just for unique values of λ that are named Eigenvalues of the correlation matrix R. The related values of u are called eigenvectors. The correlation matrix is indicated by real and positive eigenvalues. The corresponding eigenvectors are special. Assume that the eigenvalues are discrete and the eigenvalues of the p-by-p matrix R be represented by $\lambda_0, \lambda_1, \dots, \lambda_{p-1}$, and the corresponding eigenvectors are represented by $u_0, u_1, \dots, u_{p-1}$, respectively.

We can write

$$Ru_j = \lambda_j u_j, \quad j = 0, 1, \dots, p - 1 \tag{4.16}$$

Let the associating eigenvalues be organized in decreasing order:

$$\lambda_0 > \lambda_1 > \dots > \lambda_j > \lambda_{p-1} \tag{4.17}$$

Therefore $\lambda_0 = \lambda_{max}$. Let the corresponding eigenvectors be utilized to establish a p-by-p matrix:

$$U = [u_0, u_1, \dots, u_j, \dots, u_{p-1}] \tag{4.18}$$

Λ is diagonal matrix elicited by the eigenvalues of matrix R:

$$\Lambda = \mathrm{diag}[\lambda_0, \lambda_1, \dots, \lambda_j, \dots, \lambda_{p-1}] \tag{4.19}$$

The matrix U is an orthogonal matrix in the perception that its column vectors provide for the conditions of orthonormality:

$$u^T u_j = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases} \tag{4.20}$$

Eq. (4.19) needs separate eigenvalues. Equivalently, we can write

$$U^T U = I$$

In conclusion, the inverse of matrix U is equal to its transpose, as shown by

$$U^T = U^{-1} \tag{4.21}$$

We can rewrite Eq. (1.18) in a form that is known as the orthogonal transformation:

$$U^T R U = \Lambda \tag{4.22}$$

Or

$$u_j^T R u_k = \begin{cases} \lambda, & k = j \\ 0, & k \neq j \end{cases} \tag{4.23}$$

From Eqs (4.8) and (4.22) the result follows that the variance probes and eigenvalues are same, as shown by

$$\Psi(u_j) = \lambda_j, \qquad j = 0, 1, \dots, p-1 \tag{4.24}$$

In summary, we can summarize the two prominent inferences that have got from the eigenstructure of principal components analysis:

- The eigenvectors of the R (correlation matrix) belonging to the zero-mean data vector X explain the unit vectors u_j, denoting the principle directions that the variance have their stationary values.
- The corresponding eigenvalues describe the stationary values of the variance probes.

4.1.2 Basic Data Representation

There is p possible solution for the unit vector u and p possible projections of the data vector X. from Eq. (4.1) we explain that

$$a_i = u_j^T X = X^T u_j, \qquad j = 0, 1, \ldots, p-1 \tag{4.25}$$

a_i represents the projections of X onto the principal directions which are represented by the unit vector u_j and it also called the principal components; they possess the equal physical dimensions as the X. The Eq. (4.24) can be convinced as one of the analysis.

To recreate the original data X from the projection a_i, we follow the process that is shown below.

We collaborate the set of projection $\{a_i | j = 0, 1, \ldots, p-1\}$ into a single vector.

$$\begin{aligned} a &= \left[a_0, a_1, \ldots, a_{p-1}\right]^T \\ &= \left[X^T u_0, X^T u_1, \ldots, X^T u_{p-1}\right]^T \\ &= U^T X \end{aligned} \tag{4.26}$$

And then we premultiply both sides of Eq. (4.25) by the U, and the original data vector can be recreated as follows:

$$\begin{aligned} X &= Ua \\ &= \sum_{j=0}^{p-1} a_j u_j \end{aligned} \tag{4.27}$$

The Eq. (4.26) is a coordinate transformation, according to X in data space is transformed into a corresponding a in feature space (Lee, T & Peterson & Tsai, 1990).

4.1.3 Dimensionality Reduction

Principal components analysis provides an effective technique for dimensionality reduction. Particularly, we can reduce the number of features for representing the effective data by discarding those combinations (linear combination) in Eq. (4.26) that have small variance and maintain those terms that have large variances. Let $\lambda_0, \lambda_1, \ldots, \lambda_{m-1}$ elicits the largest eigenvalues of the correlation R. We can estimate the data X by abbreviating the expansion of Eq. (4.26) after m terms as shown below:

$$X' = \sum_{j=0}^{m-1} a_j\, u_j, \qquad m < p \tag{4.28}$$

It is important to note that the largest eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_{m-1}$ do not go into the calculation of the estimating vector X'; they purely decide the number of the terms in the real expansion utilized to calculate the X' that is shown below.

$$e = X - X' \tag{4.29}$$

Substitution Eqs. (4.25) and (4.26) in (4.27) yields

$$e = \sum_{j=m}^{p-1} a_j\, u_j \tag{4.30}$$

The vector e is orthogonal to estimating data vector X' or the product of the vectors X' and e is equal to zero (see Figure 4.1). This property is represented below by utilizing Eqs. (4.27) and (4.28).

$$
\begin{aligned}
e^T X' &= \sum_{i=m}^{p-1} a_i \, {u_i}^T \sum_{j=0}^{m-1} a_j \, u_j \\
&= \sum_{i=m}^{p-1} \sum_{j=0}^{m-1} a_i a_j u_i^T \, u_j \\
&= 0
\end{aligned} \tag{4.31}
$$

The Eq. (4.29) is known as the principle of the orthogonality. The entire variance of the p components of the components of the random vector X is, via Eq. (4.8) and the first line of Eq. (4.22),

$$
\sum_{j=0}^{p-1} \sigma_j^2 = \sum_{j=0}^{p-1} \lambda_j \tag{4.32}
$$

Here σ_j^2 is the variance of the j^th principle component a_j and the total variance of the m elements of the estimating X' is represented below.

$$
\sum_{j=0}^{m-1} \sigma_j^2 = \sum_{j=0}^{m-1} \lambda_j \tag{4.33}
$$

Therefore, the entire variance of the (m-p) elements in the estimation error vector $(X - X')$ is,

$$
\sum_{j=m}^{p-1} \sigma_j^2 = \sum_{j=0}^{p-1} \lambda_j \tag{4.34}
$$

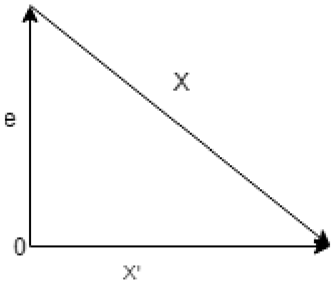

Figure 4.1 Relationship between vector X, its recreated version X' and vector e

The eigenvalues $\lambda_m, \ldots, \lambda_{p-1}$ are the smallest eigenvalues of the correlation matrix R. All these eigenvalues are close to zero, the effective eigenvalues will be the dimensionality reduction that preserves the information content of the original data. Therefore, to accomplish dimensionality reduction on original data, we calculate the eigenvalues and corresponding eigenvectors of the correlation matrix of original data vector and then project the input data onto the subspace that is extended by eigenvectors associating with the large eigenvalues.

4.1.4 The curse of Dimensionality

The curse of dimensionality was coined by Bellman in 1961. It refers to the problems associated with multivariate data analysis as the dimensionality increases. In practice, the curse of dimensionality means that the maximum number of features which the performance of our classifier will deteriorate rather than improve. In most case, the additional information that is lost by scraping some features is compensated by more accurate mapping in the lower-dimensional space (Bellman, 1961).

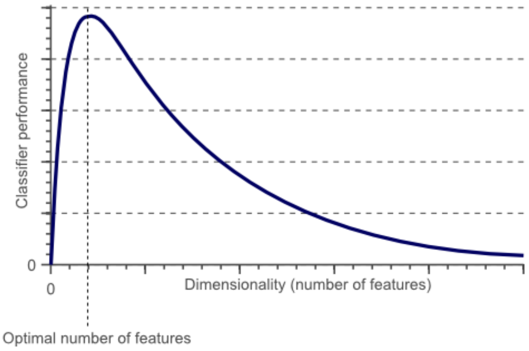

Figure 4.2 The curse of dimensionality (Bellman, 1961)

CHAPTER FIVE
METHODS

Paroxysmal atrial fibrillation patients can be diagnosed when an atrial fibrillation episode is identified in ECG record. PAF is a rhythm disturbance that happens in short time spontaneously. At the sequence of this, it is hard to obtain an ECG record during a PAF episode in a normal physical investigation in healthcare facilities. On the other hand, there is a need for 24 hours or more for diagnosis by utilizing Holter which is confining for the patients and also increase the economic charges on the system (Cowan et al., 2014). The approach proposed in this thesis utilizes ECG records that is taken during normal sinus rhythm of the heart to detect PAF patients. The records were preprocessed and 31 HRV features were obtained. Atrial and ventricular ectopic beat number was also used as separate features and the dimension of the feature space was 33” (Hilavin et al., 2016). The HRV analyzing was explained in Chapter 3 and these 33 features were used as input features to comparing the performance of feature extraction method that is PCA. Feature extraction was accomplished by PCA algorithm as explained in Chapter 4. As the result of PCA, different subset of dimensions were extracted. After dimensionality reduction, classification of PAF and non-PAF patients was performed utilizing K-Nearest Neighbor (KNN) by using the whole dimension as given in the following section entirely. The researching approach is carried through the software environment of MATLAB version 2015b. During the study, an Intel(R) Core(TM) i7-6500U @ 2.50GHZ computer with 8 GB memory was used.

5.1 Data Acquisition

In this study, the experiments were done using features extracted by PCA method for PAF from ECG records freely accessible by PhysioNet. PhysioNet provides an extensive range of physiologic signals and is an open-source software to the researchers (Goldberger et al., 2000). The original data includes two-channel ECG recordings that have been generated for use in the Computers in Cardiology Challenge 2001, where the main destination of the challenge was expanding automated methods for forecasting PAF (Moody et al., 2001). The signals were typified at 128 Hz and digitized with 12-bit resolution and also it includes automatically generated QRS occurrence times of each record. On the other hand, those timings are not asserted by experts or false detection (Moody et al., 2001). The RR interval series obtained from these QRS occurrence times were utilized in this comparative study due to the reduction of the sensitivity of the system to false QRS detections in actual applications.

The database includes 100 ECG records came by 98 different subjects. Any ECG records include two 30 minutes' records from the one and the same subject. These 100 ECG records have been divided into two subgroups. The first group contains 53 subjects who have been diagnosed with a PAF attack which is called as prior-to-PAF and the second group contains 47 subjects who are at least 45 minutes far from any PAF episode which is called as non-PAF records. These non-PAF subjects put in healthy controls those are referred for long-term ambulatory ECG monitoring and who are in-depth care units with no PAF activity. In this thesis, prior-to-PAF records were not utilized to make the performance valuation more achievable and realistic. Using prior-to-PAF records can increase the performances of the classifiers yields to be an invalid bias in real life and also getting ECG just before PAF is impossible. So, 288 five minute segments from distance-from-PAF and 510 from non-PAF were acquired with no overlapping due to the short-term HRV analyzing would be done (Hilavin et al., 2016).

5.2 Classification

The world has massive amounts of data and records on every aspect of human ventures: for instance, performance monitoring. The data come in an extensive variety forms such as numeric, textural, video, audio signals. Understanding and percept the data acquire some automated procedure to aid the analyzing the data.

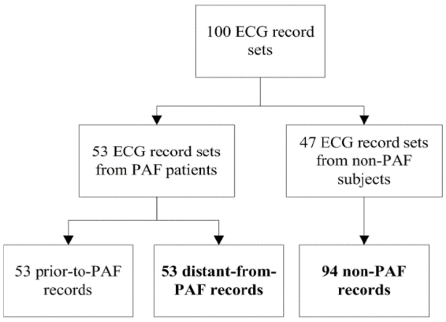
100 ECG record sets
53 ECG record sets from PAF patients
47 ECG record sets from non-PAF subjects
53 prior-to-PAF records
53 distant-from-PAF records
94 non-PAF records

Figure 5.1 The summary of data acquisition. The records which are represented by bold characters were utilized in this thesis

Methods for analyzing the data contain those for signal processing, filtering, dimension reduction, data summarization, etc have been advanced in many kinds of literatures like physics, mathematics, among others. The main purpose of the classification is providing the procedure or description of practical applications of real-world problems. Classification both supervised and unsupervised receive some attention on the statistical theory of discrimination. A large number of applications ranging from varying from the classical ones such as medical diagnosis has attracted considerably with many methods.

In this thesis, the supervised K-Nearest Neighbors was utilized to find the best classification of PAF and non-PAF subjects.

5.2.1 K-Nearest Neighbor Algorithm

Nearest Neighbor is the simplest decision algorithm that can be utilized for classification. It classifies the sample based on its nearest neighbor and also utilizes some or all the patterns which are accessible in the training set. On the other hand, it finds the similarity between the test patterns and any patterns in the training set. The nearest neighbor allocates to a test pattern the class label of its nearest neighbor (Altman, 1992).

In K-Nearest Neighbor, instead of one nearest neighbor as in the NN algorithm, K neighbors can be found. The maximum class of these K nearest neighbor is the class number or label allocated to the new pattern. The number of K is important since, with the valid number of k, the classification accuracy will be improved (Mitchell, 1997).

The training sets in kNN classification include objects whose valid classes are known. The objects are appeared by position vectors in a space with multidimensional space to detect neighbors.

$$d_{Euclidean} = \sqrt{\sum_{i}(x_i - y_i)^2} \tag{5.1}$$

Here, x_i elicits the test set and y_i elicits the train set.

There are many distance measures, for instance, the Minkowski, city-block, cosine but the most common one is Euclidean distance.

$$d_{Minkowski} = \sqrt[p]{\sum_{j=1}^{n}\left|x_{sj} - y_{tj}\right|^p} \tag{5.2}$$

The City Block distance is a special case of the Minkowski distance, where p=1.

$$d_{City\ Block} = \sum_{j=1}^{n}\left|x_{sj} - y_{tj}\right|^p \tag{5.3}$$

In training phase, there is no need for any specific operation. Both storing the feature vectors and class number of the training data are the training phase of the algorithm. In the real classification with kNN, the test sample whose class is not known is eliciticing as a vector in the feature space. After calculating the distance between the new vector and stored vectors, k nearest samples are picked, the test sample assigns the number of a class which has the nearest distance in comparison to the others.

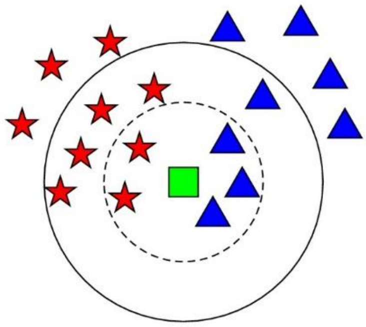

Figure 5.2 A simple diagram of KNN classifier. The test sample that is shown in green to be classified. The two classes train samples elicit with blue and red. If k is selected 3, the test samples will have belonged to blue class

KNN classification does not depend on any assumption, unlike other clustering methods which assume a Gaussian distribution. Because of this reason, KNN permits a great of the non-specific statement in the classification (Duda et al, 2001).

The disadvantages of KNN are:

1. It is very time-consuming in the large dataset
2. In presence of the noise, accuracy of the KNN can be dignified

5.3 Model Evaluation

In order to evaluate the performances, following measures were used and explained in this study. The measurements have been recommended by physicians and health-care employees which are good enough to evaluate the performance of all automated system (Valafar, 2000).

In this thesis, following four measures were used:

$$\text{Accuracy (\%)} = \frac{TP+TN}{TP+FP+FN+TN} \times 100 \tag{5.4}$$

$$\text{Sensitivity (\%)} = \frac{TP}{TP+FN} \times 100 \tag{5.5}$$

$$\text{Precision (\%)} = \frac{TP}{TP+FP} \times 100 \tag{5.6}$$

$$\text{Specificity (\%)} = \frac{TN}{TN+FP} \times 100 \tag{5.7}$$

Where TP, TN, FP, and FN refer to true positive, true negative, false positive, and false negative, respectively.

True Positive (TP): PAF episode is classified as PAF episode
True Negative (TN): the non-PAF episode is classified as non-PAF
False Positive (FP): any non-PAF episode that is classified as PAF
False Negative (FN): any PAF episode that classified as non-PAF

AUC: It is a common method to calculate the area under the ROC and its values are separated between 0 to 1.0. for achieving AUC, the true positive rates are plotted against the false positive rates.

CHAPTER SIX
RESULTS AND CONCLUSION

6.1 Results

Detecting of Paroxysmal Atrial Fibrillation form normal sinus rhythm (NSR) ECG records became a necessity since the time of PAF episodes initiates and ends quickly and the patient doesn’t have enough time to go to a health clinic and take his/her ECG at the attack time. Principal Component Analysis (PCA) is used for feature extraction and dimension reduction. In this comparative study, we compared the effectivity of features obtained from PCA for detecting whether the subject has PAF or not. The process of the study is represented in Figure 6.1 entirely. In this thesis, the features on which PCA algorithm was run had previously obtained from HRV analysis (Hilavin et al, 2016). The features with the length of 33 were reduced one-by-one to new feature sets leading to dimension reduction. Then, we used these new feature sets as input to KNN algorithm for classifying the data into PAF or non-PAF. The number of neighbors (K) and distance metrics are the important parameters in KNN.

To compare the clustering and the performance, the number of K was selected as 1,3 and 5 and three different distance metrics, Euclidean, City Block and Minkowski, were used.

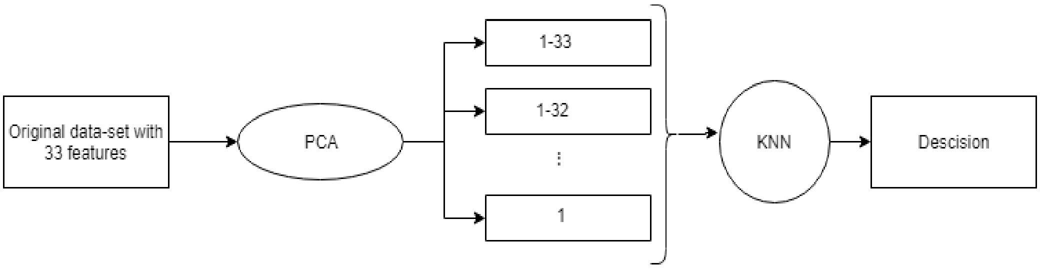

Figure 6.1 Flowchart of the study

Table 6.1 Illustrating the features which were extracted by HRV analyzing

Number	**Feature**	**Description**
1	Mean RR	Mean of RR interval
2	Std RR	Standard deviation of RR intervals
3	Mean HR	Mean of HR interval
4	RMSSD	Square root of the mean squared differences between successive RR intervals
5	NN50	Number of successive RR interval pairs that differ by more than 50 ms
6	PNN 50	NN50 divided by the total number of RR intervals
7	VLF peak	VLF band peak frequency
8	LF peak	LF band peak frequency
9	HF peak	HF band peak frequency
10	VLF power	Absolute power of VLF band
11	VLF power prc	Relative power of VLF band (VLF power/Total power)
12	LF power	Absolute power of LF band
13	LF power prc	Relative power of LF band (LF/Total power)
14	LF power nu	Power of LF band in normalized units(LF power/(Total power – VLF power)
15	HF power	Absolute power of HF band
16	HF power prc	Relative power of HF band (HF power/Total power)
17	HF power nu	Power of HF band in normalized units(HF power/(Total power – VLF power)
18	LF/HF power	Ratio of LF and HF power bands

Table 6.1 continues

19	SD1	Dispersion of points perpendicular to the line of identity in Poincare plot
20	SD2	Dispersion of points along the line of identity in Poincare plot
21	ApEn	Approximation entropy
22	SampEn	Sample entropy
23	DFA Alpha1	Short term fluctuation slope of detrended fluctuation
24	DFA Alpha2	Long term fluctuation slope of detrended fluctuation analysis
25	CorDim D2	Correlation dimension
26	RPA Lmax	Maximum line length in recurrence plot analysis
27	RPA Lmean	Mean line length in recurrence plot analysis
28	RPA REC	Recurrence plot determinism
29	RPA Det	Recurrence plot determinism
30	RPA ShanEn	Recurrence plot Shannon entropy
31	CCM	Complex of point along the line of identity in Poincare plot
32	Atrial ectopic number	The number of atrial ectopic beats
33	Ventricular ectopic number	The number of ventricular ectopic beats

As seen from the results presented in Table 6.2 to Table 6.10, we sorted the 33 features according to the largest Eigenvalues and those corresponding Eigenvectors, then reduced 33 dimensions into 32, 31, 30, …, 2, 1.

We evaluated the KNN on different data sets of varying dimensions' size. Principal Componant Analysis (PCA) was utilized to reduce the dimensionality of dataset, to speed up training and analyze the performance. The number of neighbors (K=1,3 and

5) and distance metrics (Euclidean, City Block and Minkowski) were used to see the efficiency of PCA. For the purpose of cross validation, the training sets were further divided into training sets (In this study, 80% data set partitioned into training datasets with K-fold 10).

In order to visualize the performance of PCA as dimension reduction method and compare the impact on both different number of K and distance functions in the performance of the KNN algorithm, we plotted the results in terms of accuracy and AUC. The results are given in Figure 6. 2 to Figure 6. 7. The results obtained represent that the performance of PCA has not changed significantly until reducing the number of features to approximately 10. However, reducing the dimension further affects the performance significantly.

We first compared KNN classification (with K=1) error rates using Euclidean, City Block and Minkowski distances. We repeatedly reduced the dimension size by PCA and recorded the results in Table 6.2 to Table 6.4. Figure 6.2 summarizes the main results and elicits the performance of PCA in terms of accuracy and AUC. Accuracy and AUC of method were stable approximately in 32 to 10 dimensions but suddenly reduced after 10. The performances of distance metrics were not different significantly.

In continue, we compared KNN (with K=3) error rates using Euclidean, City Block and Minkowski distances and reduced the dimension size and recorded the results in Table 6.5 to Table 6.7 and summarized the results that shown in Figure 6.3. In 3NN the accuracy and AUC results were stable in 32 to 10 dimensions. After 10 dimension, the accuracy and AUC reduced. In the same way, we measured the differences of 5NN error rates using Euclidean, City Block and Minkowski distances and documented the results in Table 6.8 to Table 6.10. The summarization of results represented in Figure 6.4 and the result is similar to the others.

We also computed the error rates using different number of neighbors in same distance metrics. Firstly, we selected Euclidean as a distance metrics and fixed it, then three different number of neighbors (K=1,3 and 5) tried. As mentioned in previous section, the dimensions were reduced by PCA one-by-one and the results represented in Figure 6.5. In this case, reducing dimension by PCA has desire results until approximately 10 dimension, but reducing further effects negatively on both accuracy and AUC. However, K=3 and 5 have better results in comparison to K=1 in terms of AUC.

Secondly, we selected City Block as a distance metrics and three different number of neighbors (K=1,3 and 5) tried. The result of this case is similar to the previous one in that the PCA has acceptable results until 10 dimension, but reducing dimension more than 10 effects negatively on the results as shown in Figure 6.6.

Finally, we determined Minkowski as distance metrics and three different number of neighbors (K=1,3 and 5) checked. The results are represented in Figure 6.7. similarly, decreasing the number of input dimension (i.e. the number of features) further after 10 dimension negatively affects the results.

It should be noted that we down-sampled the dimension of data from 33 to 1 dimension and used PCA to obtain these dimensions. Training and test datasets were created by randomly sampling of 798 included PAF and non-PAF patients. The results in all cases were averaged over three experiments with different random of each data set. Because of this reason, we calculated the average of the results and considered them as final results.

Table 6.2 Evaluating the performances of KNN after reducing dimension with K=1 and Euclidean distance

1NN Euclidean Distance					
Dimension	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**AUC**
33	89.16	83.42	92.76	87.29	0.90
32	88.43	83.16	91.16	82.89	0.90
31	85.83	82.02	87.55	76.36	0.89
30	82.91	72.73	87.66	73.71	0.85
29	86.66	81.12	90.01	80.25	0.89
28	83.75	82.79	84.00	74.93	0.90
27	85.83	82.45	88.04	80.05	0.90
26	85.00	78.18	88.42	77.89	0.87
25	85.00	87.37	83.87	72.92	0.92
24	81.66	79.99	83.47	75.91	0.87
23	90.00	87.07	91.56	82.52	0.92
22	86.66	84.37	86.24	74.37	0.91
21	90.83	88.04	91.98	84.31	0.92
20	89.16	90.29	89.12	83.00	0.93
19	89.16	84.67	91.25	82.77	0.91
18	85.83	82.24	88.14	77.25	0.90
17	80.83	78.23	83.00	74.22	0.87
16	83.33	72.90	88.97	78.66	0.85
15	81.66	77.20	84.78	75.01	0.87
14	88.75	82.44	92.25	88.28	0.90
13	85.41	84.15	86.19	72.98	0.90
12	85.00	81.37	87.21	78.17	0.89
11	82.91	73.31	88.91	80.43	0.83
10	83.33	72.29	87.24	71.02	0.84
9	85.41	80.64	89.35	85.84	0.90
8	84.58	76.34	90.12	82.01	0.86
7	76.66	61.56	85.86	72.80	0.79
6	79.58	74.65	79.20	71.91	0.86
5	75.00	59.72	82.61	63.14	0.78
4	75.00	63.08	81.78	66.26	0.80

Table 6.2 continues

Dimension	Accuracy	Sensitivity	Specificity	Precision	AUC
3	73.33	58.14	82.91	68.27	0.77
2	69.58	59.14	76.05	60.53	0.78
1	63.33	43.01	74.64	48.82	0.70

Table 6.3 Evaluating the performances of KNN after reducing dimension with K=1 and City Block distance

1NN City Block Distance					
Dimension	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**AUC**
33	89.16	87.26	90.08	81.23	0.93
32	87.91	83.12	90.15	82.28	0.90
31	87.08	87.80	86.92	76.48	0.92
30	81.00	84.73	90.25	81.34	0.91
29	87.08	85.69	87.87	77.31	0.91
28	90.41	82.22	95.85	92.57	0.89
27	87.91	85.83	89.22	83.05	0.91
26	88.75	89.73	87.94	79.32	0.93
25	89.66	86.32	91.51	81.63	0.92
24	87.50	81.15	90.44	81.37	0.89
23	85.00	81.95	86.77	76.22	0.89
22	85.83	83.14	87.32	77.45	0.90
21	90.83	92.12	90.04	80.98	0.95
20	88.33	84.61	90.37	82.63	0.91
19	88.18	82.56	92.56	86.48	0.89
18	85.83	82.04	88.08	79.88	0.89
17	91.25	87.78	93.45	85.74	0.92
16	85.41	81.38	87.15	80.08	0.89
15	87.08	85.49	87.56	77.48	0.91
14	85.83	79.53	89.38	78.82	0.88

Table 6.3 continues

Dimension	Accuracy	Sensitivity	Specificity	Precision	AUC
13	85.00	77.69	90.02	83.26	0.87
12	79.58	65.37	86.88	71.85	0.81
9	82.91	73.18	89.01	76.18	0.85
8	79.98	67.94	86.35	73.22	0.82
7	74.58	58.80	85.75	74.69	0.78
6	73.75	64.85	78.38	59.41	0.81
5	76.25	66.08	82.20	67.91	0.75
4	69.16	59.96	75.31	61.40	0.78
3	75.83	76.05	75.59	60.53	0.87
2	63.33	50.66	69.80	46.05	0.74
1	62.50	40.03	73.74	43.57	0.68

Table 6.4 Evaluating the performances of KNN after reducing dimension with K=1 and Minkowski distance

1NN Minkowski Distance					
Dimension	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**AUC**
33	85.41	79.87	87.85	75.06	0.88
32	86.25	84.99	87.26	80.32	0.91
31	85.41	84.43	85.59	70.74	0.90
30	89.16	88.09	89.74	82.56	0.92
29	85.83	86.66	85.02	78.21	0.92
28	83.33	80.12	85.07	74.93	0.88
27	89.16	82.59	93.08	88.37	0.89
26	86.66	80.47	80.38	76.50	0.88
25	75.41	83.57	86.68	79.05	0.90
24	83.75	79.66	85.94	75.92	0.88
23	85.83	83.02	87.20	75.06	0.90
22	89.58	86.96	90.78	84.93	0.92
21	87.91	83.25	90.80	83.94	0.90

Table 6.4 continues

Dimension	Accuracy	Sensitivity	Specificity	Precision	AUC
20	86.66	84.97	87.57	79.49	0.91
18	82.98	76.22	85.03	71.59	0.87
16	90.83	86.64	65.41	85.05	0.92
15	83.75	77.01	86.86	74.64	0.87
14	81.66	72.75	86.65	71.98	0.85
13	87.91	79.16	92.05	82.34	0.88
12	83.33	80.60	85.16	78.94	0.88
11	85.41	86.24	84.37	79.18	0.91
10	83.33	79.39	85.71	75.98	0.88
9	78.75	73.65	82.37	71.19	0.85
8	85.41	81.22	88.08	75.22	0.89
7	80.00	71.50	83.73	65.85	0.84
6	78.75	71.53	82.18	67.37	0.84
5	73.75	66.62	77.60	63.02	0.82
4	69.16	61.70	72.82	54.64	0.79
3	70.41	58.87	77.75	63.74	0.78
2	68.75	61.51	71.93	54.73	0.79
1	64.16	53.53	69.93	50.45	0.75

Table 6.5 Evaluating the performances of KNN after reducing dimension with K=3 and Euclidean distance

3NN Euclidean Distance					
Dimension	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**AUC**
33	87.91	83.40	89.82	78.20	0.90
32	86.25	81.29	89.16	80.01	0.89
31	87.07	84.00	88.92	78.64	0.90
30	88.33	87.65	88.69	76.71	0.92
29	87.91	83.40	89.82	78.20	0.90
28	85.00	82.39	85.94	77.49	0.89
27	85.00	85.30	85.52	70.47	0.91
26	84.58	81.53	86.14	75.13	0.89

Table 6.5 continues

Dimension	Accuracy	Sensitivity	Specificity	Precision	AUC
25	82.91	78.62	84.18	68.18	0.88
22	88.33	86.27	89.19	78.35	0.91
21	86.66	82.65	87.89	75.21	0.90
20	88.33	88.89	87.97	76.93	0.93
19	89.58	94.44	86.66	74.94	0.96
18	88.33	88.58	88.20	78.78	0.93
17	84.16	87.87	83.17	66.45	0.92
16	86.25	86.22	86.24	75.68	0.92
15	85.14	78.54	89.05	80.08	0.88
14	85.41	83.20	86.49	74.76	0.90
13	87.91	85.29	86.21	80.47	0.91
12	82.91	84.59	82.28	66.71	0.91
11	87.91	85.41	89.07	80.78	0.91
10	86.25	78.96	88.87	75.21	0.87
9	84.16	80.18	86.44	76.93	0.88
8	85.00	78.44	88.35	76.68	0.88
7	77.90	66.76	83.98	69.48	0.82
6	79.58	74.38	82.44	70.72	0.85
5	79.58	68.76	83.07	62.84	0.83
4	77.50	66.45	83.45	67.87	0.87
3	74.16	64.37	78.90	61.29	0.81
2	74.58	62.24	80.69	61.33	0.82
1	62.91	46.92	70.46	43.98	0.72

Table 6.6 Evaluating the performances of KNN after reducing dimension with K=3 and City Block distance

3NN City Block Distance					
Dimension	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**AUC**
33	87.50	87.97	86.86	78.72	0.93
32	90.41	86.02	93.02	86.80	0.90
31	87.08	80.59	90.47	84.55	0.89
30	89.16	94.43	86.98	75.12	0.96
27	91.66	87.68	93.73	87.72	0.92
26	90.41	94.89	88.34	79.32	0.96

Table 6.6 continues

Dimension	Accuracy	Sensitivity	Specificity	Precision	AUC
25	90.41	89.49	90.2	80.85	0.93
24	88.33	87.07	88.82	76.07	0.92
23	87.91	87.85	88.00	76.64	0.92
22	86.25	85.59	86.50	69.35	0.91
21	89.16	92.53	87.22	79.30	0.95
20	93.33	93.31	93.52	85.11	0.95
19	86.66	84.28	88.67	81.92	0.90
18	88.33	89.75	87.63	74.91	0.93
17	88.33	91.70	86.57	79.88	0.94
16	85.83	83.97	86.55	74.93	0.90
15	89.58	87.68	90.15	77.32	0.92
14	85.41	84.77	86.18	75.47	0.91
13	83.33	77.96	85.13	68.32	0.87
12	81.25	71.04	85.52	68.29	0.84
11	85.83	81.72	87.79	73.34	0.89
10	82.08	71.97	87.17	72.24	0.84
9	84.33	78.64	85.66	73.35	0.88
8	85.00	73.78	89.46	73.07	0.85
7	75.83	71.11	78.45	59.29	0.84
6	74.58	63.33	81.02	65.53	0.80
5	78.75	73.25	81.63	67.34	0.85
4	73.33	66.48	77.16	60.69	0.82
3	73.33	72.50	73.80	56.09	0.85
2	71.25	61.36	75.35	50.88	0.79
1	67.91	46.95	77.61	48.66	0.72

Table 6.7 Evaluating the performances of KNN after reducing dimension with K=3 and Minkowski distance

3NN Minkowski Distance					
Dimension	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**AUC**
33	84.58	85.04	85.03	69.28	0.91
32	87.08	85.97	87.88	78.96	0.91
31	88.33	83.96	89.96	77.70	0.90
30	86.25	89.62	84.72	70.55	0.93
29	89.58	85.95	91.21	79.50	0.91
28	87.91	86.08	88.71	77.44	0.91
27	83.75	88.73	85.19	74.14	0.88
26	87.91	88.30	87.69	74.56	0.93
25	88.75	92.86	86.30	80.98	0.95
24	87.91	83.95	89.79	83.73	0.90
23	91.66	94.45	90.02	81.83	0.96
22	89.58	91.25	89.51	79.04	0.94
21	86.25	81.50	86.39	75.53	0.91
20	90.41	85.52	92.65	83.74	0.91
19	82.91	89.35	78.77	69.80	0.93
18	89.16	86.97	90.22	77.96	0.92
17	85.83	83.72	86.89	73.26	0.90
16	85.00	80.78	87.33	77.33	0.89
15	87.91	90.84	86.84	71.99	0.94
14	86.25	79.55	89.15	72.90	0.88
13	88.75	82.39	90.76	82.59	0.90
12	84.58	82.13	85.54	71.84	0.90
11	87.08	91.66	84.67	76.44	0.94
10	86.25	84.77	87.41	80.32	0.91
9	85.83	78.52	89.74	80.63	0.87
8	85.83	80.72	88.28	79.59	0.89
7	75.78	69.91	80.07	69.45	0.83
6	80.33	72.56	82.71	63.87	0.85
5	75.41	70.16	78.36	62.59	0.83
4	70.00	70.23	69.71	52.45	0.83
3	76.66	71.13	78.74	57.85	0.84

Table 6.7 continues

Dimension	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**AUC**
2	74.16	63.50	80.20	66.40	0.80
1	71.69	62.06	75.95	53.34	0.80

Table 6.8 Evaluating the performances of KNN after reducing dimension with K=5 and Euclidean distance

5NN Euclidean Distance					
Dimension	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**AUC**
33	86.25	89.00	85.31	73.25	0.92
32	87.91	90.58	86.31	76.53	0.93
31	89.58	90.79	89.39	78.04	0.94
30	89.16	89.65	88.95	76.76	0.93
29	88.75	86.32	89.60	76.37	0.92
28	85.83	92.79	82.65	69.60	0.95
27	85.00	85.25	84.54	64.29	0.91
26	85.00	85.25	84.54	64.29	0.91
25	88.33	91.58	86.62	78.56	0.94
24	82.91	91.32	79.42	65.33	0.94
23	90.00	90.81	89.60	77.45	0.94
22	84.58	88.76	82.89	63.75	0.93
21	90.00	91.47	89.37	78.39	0.94
20	87.08	95.10	83.83	72.53	0.96
19	86.66	84.01	87.60	73.56	0.90
18	85.83	89.13	85.21	68.69	0.93
17	84.16	88.52	82.36	70.18	0.93
16	88.75	89.41	88.50	75.26	0.93
15	85.41	83.54	86.52	77.20	0.90
14	87.91	83.53	90.22	80.51	0.90
13	86.25	87.39	85.90	72.18	0.92
12	84.58	84.63	85.51	71.68	0.91
11	87.08	85.87	87.47	77.36	0.91
10	84.16	77.88	85.33	65.73	0.87
9	85.41	83.00	86.78	77.29	0.90

Table 6.8 continues

Dimension	Accuracy	Sensitivity	Specificity	Precision	AUC
8	85.83	79.67	88.95	77.57	0.87
7	80.00	69.45	85.83	73.31	0.83
6	77.08	72.33	80.11	66.14	0.85
5	77.50	66.27	81.94	59.88	0.81
4	79.58	72.32	83.53	66.63	0.84
3	75.83	67.77	79.62	63.26	0.82
2	72.50	60.00	78.70	58.17	0.79
1	65.00	50.74	71.63	45.99	0.74

Table 6.9 Evaluating the performances of KNN after reducing dimension with K=5 and City Block distance

5NN City Block Distance					
Dimension	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**AUC**
33	87.91	88.13	87.33	75.22	0.92
32	88.75	87.17	89.54	79.73	0.92
31	88.33	86.49	89.48	77.44	0.92
30	87.91	91.76	86.44	74.28	0.95
29	87.50	87.97	87.39	76.00	0.92
28	90.33	94.59	90.58	81.48	0.96
27	87.91	96.03	84.12	73.77	0.97
26	89.16	85.90	90.22	76.85	0.92
25	87.50	85.18	88.39	74.38	0.91
24	85.00	87.04	84.30	66.22	0.92
23	90.00	97.26	86.47	77.01	0.97
22	90.83	93.26	89.98	75.59	0.95
21	86.66	87.35	86.73	77.97	0.92
20	90.00	90.60	89.79	76.25	0.94
19	88.33	90.97	87.42	81.12	0.94
18	88.33	91.31	86.87	75.50	0.94
17	89.16	89.88	88.88	72.98	0.93
16	85.83	87.97	85.06	72.12	0.92
15	84.58	80.32	85.83	66.14	0.88

Table 6.9 continues

Dimension	**Accuracy**	**sensitivity**	**Specificity**	**Precision**	**AUC**
14	91.25	95.06	89.96	76.25	0.96
13	86.25	80.34	88.97	77.16	0.89
12	81.66	76.19	84.17	63.23	0.86
11	85.83	80.11	87.91	74.26	0.89
10	83.75	74.89	87.05	73.23	0.86
9	85.41	89.29	83.99	67.43	0.93
8	80.83	74.47	83.92	69.11	0.86
7	80.41	77.70	81.82	68.54	0.87
6	75.00	77.22	74.56	59.10	0.87
5	77.08	74.72	77.98	57.24	0.86
4	79.16	62.21	81.68	58.16	0.80
3	73.33	61.88	78.83	56.77	0.79
2	72.08	58.67	78.33	58.75	0.78
1	67.08	58.50	70.69	46.32	0.78

Table 6.10 Evaluating the performances of KNN after reducing dimension with K=5 and Minkowski distance

5NN Minkowski Distance					
Dimension	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**AUC**
33	82.91	92.26	79.33	65.57	0.95
32	89.16	91.16	88.50	71.74	0.94
31	84.58	83.91	84.79	69.93	0.90
30	90.00	91.33	89.05	82.28	0.94
29	85.00	82.08	85.67	65.09	0.90
28	88.33	93.99	84.82	77.70	0.95
27	89.58	86.08	91.33	83.04	0.92
26	88.75	87.40	89.20	74.31	0.92
25	85.41	88.84	84.50	71.33	0.92
24	85.41	88.67	84.31	70.15	0.92
23	90.41	86.75	91.72	81.86	0.92
22	83.33	93.06	77.87	67.58	0.95
21	88.33	85.97	89.13	77.13	0.91

Table 6.10 continues

Dimension	Accuracy	Sensitivity	Specificity	Precision	AUC
20	85.41	84.31	85.99	70.88	0.90
19	85.41	84.78	85.76	74.24	0.91
18	88.75	87.84	90.13	71.72	0.92
17	88.75	90.43	88.39	74.21	0.94
16	86.66	78.60	88.87	73.82	0.88
15	86.66	89.56	85.71	70.63	0.93
14	89.16	95.96	86.05	75.67	0.96
13	84.58	87.20	83.44	75.65	0.92
12	89.58	84.37	92.32	85.14	0.90
11	84.58	81.59	85.75	71.37	0.89
10	83.33	83.55	83.15	73.20	0.90
9	82.08	78.31	83.15	66.57	0.87
8	83.75	82.96	84.03	84.08	0.90
7	80.41	83.27	78.89	69.49	0.90
6	76.25	70.87	79.34	65.13	0.84
5	79.58	73.56	82.63	69.50	0.85
4	79.58	71.04	83.12	58.02	0.84
3	75.41	69.51	79.02	67.36	0.83
2	68.33	57.88	74.18	54.25	0.77
1	66.25	52.06	70.89	37.67	0.75

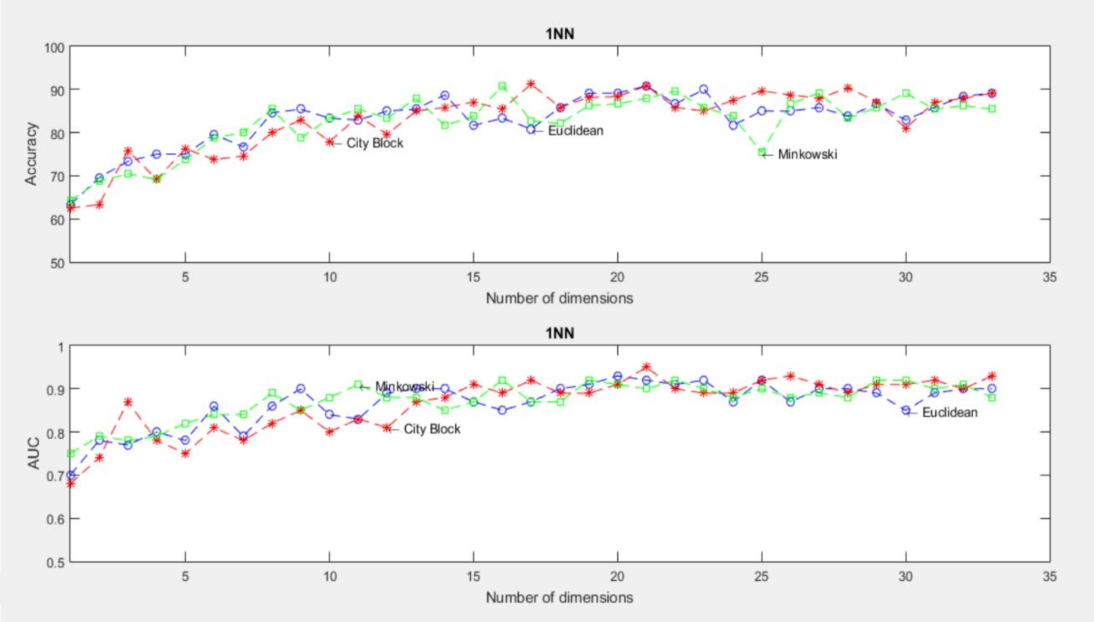

Figure 6.2 The results of dimension reduction in 1NN classifier with 3 different distance functions

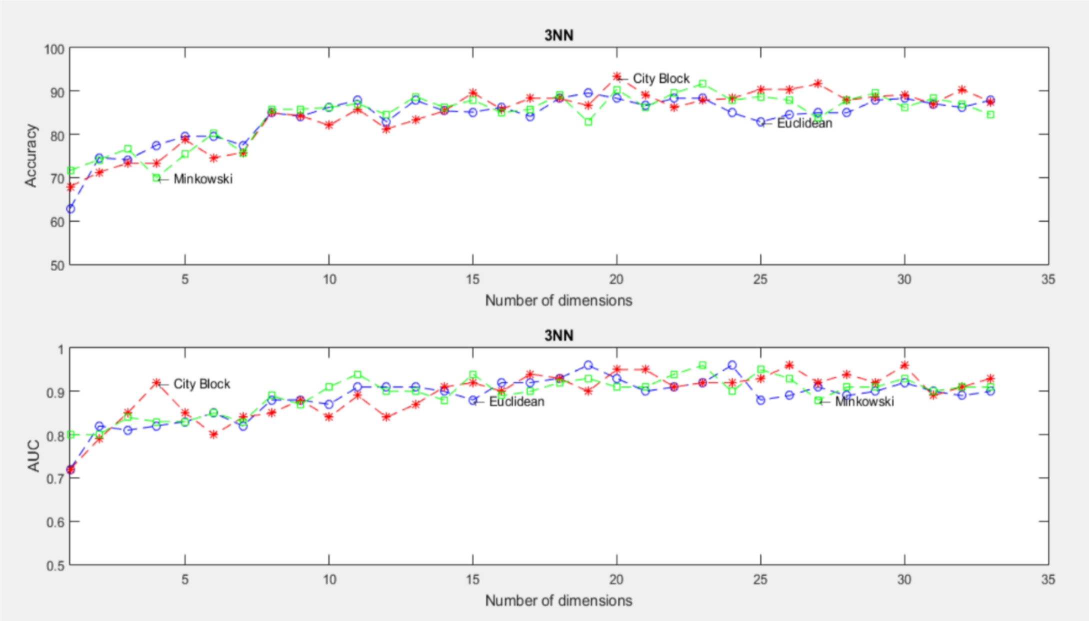

Figure 6.3 The results of dimension reduction in 3NN classifier with 3 different distance functions

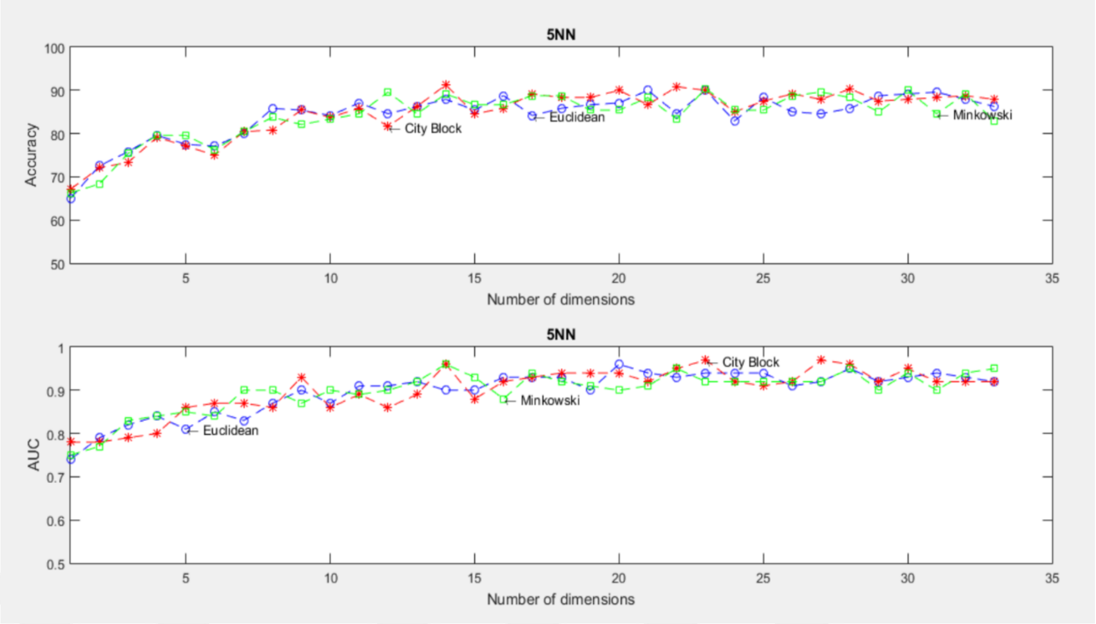

Figure 6.4 The results of dimension reduction in 5NN classifier with 3 different distance functions

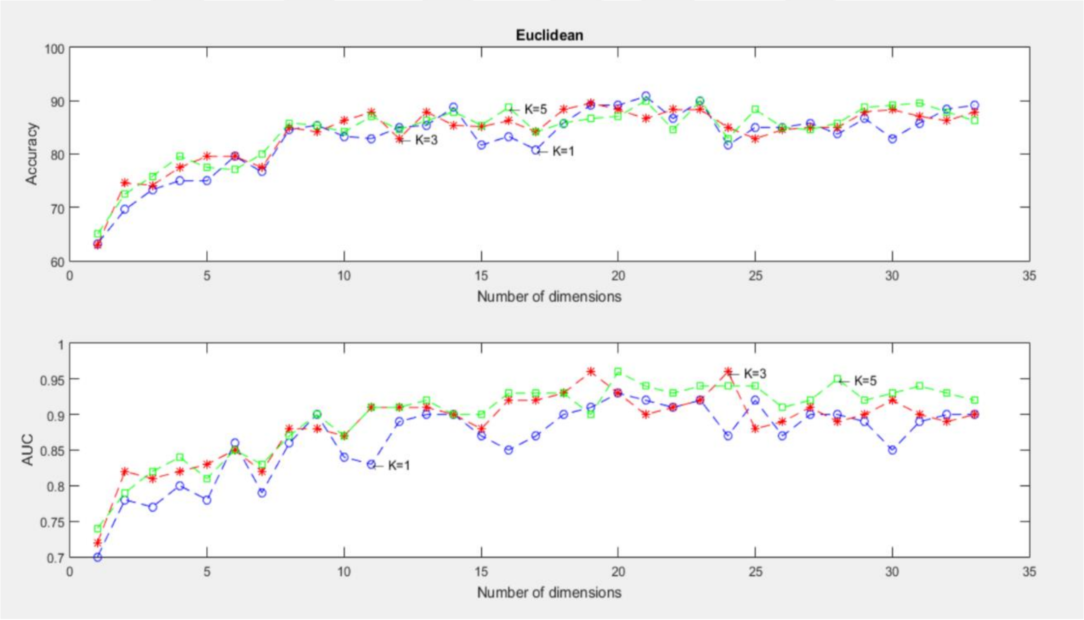

Figure 6.5 The performance evaluation of KNN with different K number in Euclidean distance

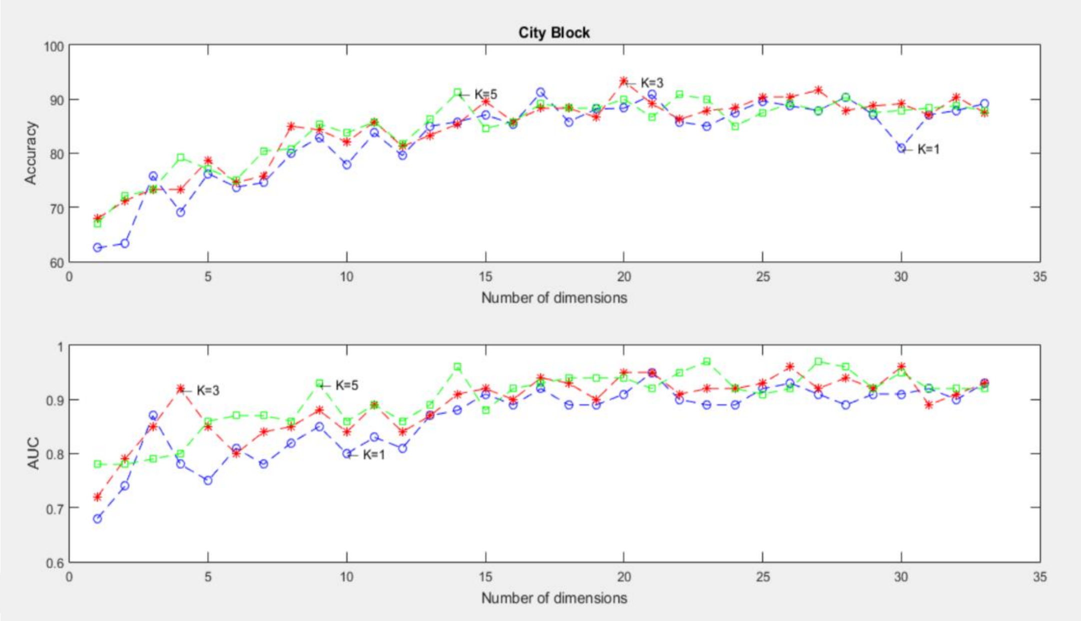

Figure 6.6 The performance evaluation of KNN with different K number in City Block distance

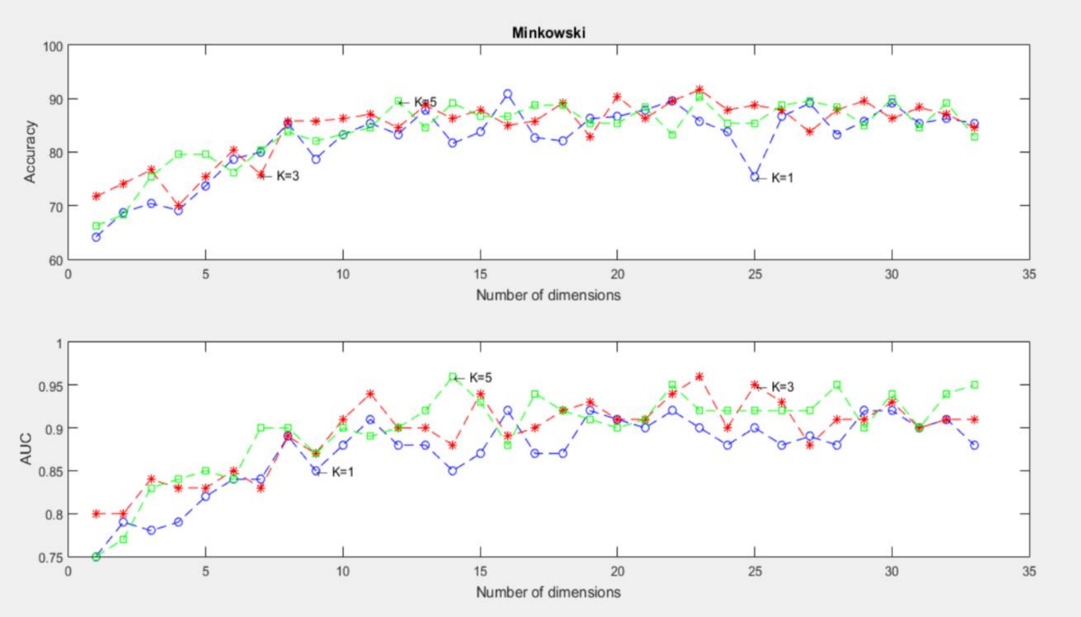

Figure 6.7 The performance evaluation of KNN with different K number in Minkowski

Performance results for PCA in each number of K and different distance metrics can be seen in Table 6.11 to Table 6.19.

Table 6.11 Performance result of PCA with K=1 and Euclidean distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	**Accuracy**	**AUC**
1	**63.23**	**0.70**
2	69.58	0.78
3	73.33	0.77
4	75	0.80
5	75	0.78
6	79.58	0.86
7	76.66	0.79
8	84.58	0.86
9	85.41	0.90
10	83.33	0.84
11	82.91	0.83
12	85	0.89
13	85.41	0.90
14	88.75	0.90
15	81.66	0.87
16	83.33	0.85
17	80.83	0.87
18	85.83	0.90
19	89.16	0.91
20	89.16	0.93
21	**90.83**	**0.92**
22	86.66	0.91
23	90	0.92
24	81.66	0.87
25	85	0.92
26	85	0.87

Table 6.11 continues

Number of features	Accuracy	AUC
27	85.83	0.90
28	83.75	0.90
29	86.66	0.89
30	82.91	0.85
31	85.83	0.89
32	88.43	0.90
33	89.16	0.90
Average	**83.01**	**0.86**

Table 6.12 Performance result of PCA with K=1 and City Block distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	**Accuracy**	**AUC**
1	**62.50**	**0.68**
2	63.33	0.74
3	75.83	0.87
4	69.16	0.78
5	76.25	0.75
6	73.75	0.81
7	74.58	0.78
8	79.98	0.82
9	82.91	0.85
10	77.91	0.80
11	83.75	0.83
12	79.58	0.81
13	85	0.87
14	85.83	0.88

Table 6.12 continues

Number of features	Accuracy	AUC
15	87.08	0.91
16	85.41	0.89
17	**91.25**	**0.92**
18	85.83	0.89
19	88.18	0.89
20	88.33	0.91
21	90.83	0.95
22	85.83	0.90
23	85	0.89
24	87.50	0.89
25	89.66	0.92
26	88.75	0.93
27	87.91	0.91
28	90.41	0.89
29	87.08	0.91
30	81	0.91
31	87.08	0.92
32	87.91	0.90
33	89.16	0.93
Average	**82.86**	**0.86**

Table 6.13 Performance result of PCA with K=1 and Minkowski distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	**Accuracy**	**AUC**
1	**64.16**	**0.75**
2	68.75	0.79
3	70.41	0.78
4	69.16	0.79
5	73.75	0.82
6	78.75	0.84
7	80	0.84
8	85.41	0.89
9	78.75	0.85
10	83.33	0.88
11	85.41	0.91
12	83.33	0.88
13	87.91	0.88
14	81.66	0.85
15	83.75	0.87
16	**90.83**	**0.92**
17	82.75	0.87
18	82.08	0.870
19	86.25	0.92
20	86.66	0.91
21	87.91	0.90
22	89.58	0.92
23	85.83	0.90
24	83.75	0.88
25	75.41	0.90
26	86.66	0.88

Table 6.13 continues

Number of features	**Accuracy**	**AUC**
27	89.16	0.89
28	83.33	0.88
29	85.83	0.92
30	89.16	0.92
31	85.41	0.90
32	86.25	0.91
33	85.41	0.88
Average	**82.32**	**0.87**

Table 6.14 Performance result of PCA with K=3 and Euclidean distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	**Accuracy**	**AUC**
1	**62.91**	**0.72**
2	74.58	0.82
3	74.16	0.81
4	77.50	0.82
5	79.58	0.83
6	79.58	0.85
7	77.50	0.82
8	85	0.88
9	84.16	0.88
10	86.25	0.87
11	87.91	0.91
12	82.91	0.91
13	87.91	0.91
14	85.41	0.90
15	85.14	0.88

Table 6.14 continues

Number of features	**Accuracy**	**AUC**
16	86.25	0.92
17	84.16	0.92
18	88.33	0.93
19	**89.58**	**0.96**
20	88.33	0.93
21	86.66	0.90
22	88.33	0.91
23	88.33	0.92
24	85	0.96
25	82.91	0.88
26	84.58	0.89
27	85	0.91
28	85	0.89
29	87.91	0.90
30	88.33	0.92
31	87.07	0.90
32	86.25	0.89
33	87.91	0.90
Average	**83.95**	**0.88**

Table 6.15 Performance result of PCA with K=3 and City Block distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	Accuracy	AUC
1	**67.91**	**0.72**
2	71.25	0.79
3	73.33	0.85
4	73.33	0.92
5	78.75	0.85
6	74.58	0.80
7	75.83	0.84
8	85	0.85
9	84.33	0.88
10	82.08	0.84
11	85.83	0.89
12	81.25	0.84
13	83.33	0.87
14	85.41	0.91
15	89.58	0.92
16	85.83	0.90
17	88.33	0.94
18	88.33	0.93
19	86.66	0.90
20	**93.33**	**0.95**
21	89.16	0.95
22	86.25	0.91
23	87.91	0.92
24	88.33	0.92
25	90.41	0.93
26	90.41	0.96

Table 6.15 continues

Number of features	**Accuracy**	**AUC**
27	91.66	0.92
28	87.91	0.94
29	88.75	0.92
30	89.16	0.96
31	87.08	0.89
32	90.41	0.91
33	87.50	0.93
Average	**84.25**	**0.89**

Table 6.16 Performance result of PCA with K=3 and Minkowski distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	**Accuracy**	**AUC**
1	**71.69**	**0.80**
2	74.16	0.80
3	76.66	0.84
4	70	0.83
5	75.41	0.83
6	80.33	0.85
7	75.83	0.83
8	85.83	0.89
9	85.83	0.87
10	86.25	0.91
11	87.08	0.94
12	84.58	0.90
13	88.75	0.90
14	86.25	0.88
15	87.91	0.94

Table 6.16 continues

Number of features	**Accuracy**	**AUC**
16	85	0.89
17	85.83	0.90
18	89.16	0.92
19	82.91	0.93
20	90.41	0.91
21	86.25	0.91
22	89.58	0.94
23	**91.66**	**0.96**
24	87.91	0.90
25	88.75	0.95
26	87.91	0.93
27	83.75	0.88
28	87.91	0.91
29	89.58	0.91
30	86.25	0.93
31	88.33	0.90
32	87.08	0.91
33	84.58	0.91
Average	**84.52**	**0.89**

Table 6.17 Performance result of PCA with K=5 and Euclidean distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	**Accuracy**	**AUC**
1	65	0.74
2	72.5	0.79
3	75.83	0.82
4	79.58	0.84
5	77.50	0.81
6	77.08	0.85
7	80.00	0.83
8	85.83	0.87
9	85.41	0.90
10	84.16	0.87
11	87.08	0.91
12	84.58	0.91
13	86.25	0.92
14	87.91	0.90
15	85.41	0.90
16	88.75	0.93
17	84.16	0.93
18	85.83	0.93
19	86.66	0.90
20	87.08	0.96
21	**90.00**	**0.94**
22	84.58	0.93
23	90.00	0.94
24	82.91	0.94
25	88.33	0.94
26	85.00	0.91

Table 6.17 continues

Number of features	**Accuracy**	**AUC**
27	84.58	0.92
28	85.83	0.95
29	88.75	0.92
30	89.16	0.93
31	89.58	0.94
32	87.91	0.93
33	86.25	0.92
Average	**84.22**	**0.89**

Table 6.18 Performance result of PCA with K=5 and City Block distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	**Accuracy**	**AUC**
1	**67.08**	**0.78**
2	72.08	0.78
3	73.33	0.79
4	79.16	0.80
5	77.08	0.86
6	75	0.87
7	80.41	0.87
8	80.83	0.86
9	85.41	0.93
10	83.75	0.86
11	85.83	0.89
12	81.66	0.86
13	86.25	0.89
14	**91.25**	**0.96**
15	84.58	0.88

Table 6.18 continues

Number of features	**Accuracy**	**AUC**
16	85.83	0.92
17	89.16	0.93
18	88.33	0.94
19	88.33	0.94
20	90	0.94
21	86.66	0.92
22	90.83	0.95
23	90	0.97
24	85	0.92
25	87.50	0.91
26	89.16	0.92
27	87.91	0.97
28	90.33	0.96
29	87.50	0.92
30	87.91	0.95
31	88.33	0.92
32	88.75	0.92
33	87.91	0.92
Average	**84.64**	**0.90**

Table 6.19 Performance result of PCA with K=5 and Minkowski distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	**Accuracy**	**AUC**
1	**66.25**	**0.75**
2	68.33	0.77
3	75.41	0.83
4	79.58	0.84
5	79.58	0.85
6	76.25	0.84
7	80.41	0.90
8	83.75	0.90
9	82.08	0.87
10	83.33	0.90
11	84.58	0.89
12	89.58	0.90
13	84.58	0.92
14	89.16	0.96
15	86.66	0.93
16	86.66	0.88
17	88.75	0.94
18	88.75	0.92
19	85.41	0.91
20	85.41	0.90
21	88.33	0.91
22	83.33	0.95
23	**90.41**	**0.92**
24	85.41	0.92
25	85.41	0.92
26	88.75	0.92

Table 6.19 continues

Number of features	Accuracy	AUC
27	89.58	0.92
28	88.33	0.95
29	85	0.90
30	90	0.94
31	84.58	0.90
32	89.16	0.94
33	82.91	0.95
Average	**84.11**	**0.89**

With respect to the average results in Table 6.11 to Table 6.19, the best features are selected based on the nearest percentage to the average values. In KNN with K=1, the dimension reduction is acceptable until the accuracy values are 6% below the average value which are the 8 features in Euclidean and City Block and 5 features in Minkowski distance metrics. In KNN with K=3, the acceptable values are 4% below the average accuracy in which the all distance metrics (Euclidean, City Block and Minkowski) 8 features is the best dimension. And finally, in KNN with K=5, the dimension reduction is acceptable until the accuracy is 3% below the average value of accuracy that is the 8 features in Euclidean and Mikowski and 9 features in City Block.

In summarization, in 1NN the best features are 8 in Euclidean and City Block distance metrics and 5 features in Minkowski. In 3NN, the best feature is 8 in all distance metrics and in 5NN the best feature is 8 in Euclidean and Minkowski and 9 feature in City Block.

6.2 Using the PCA algorithm on another dataset

There are two different methods for dimension reduction, which are feature selection and feature extraction. Genetic Algorithms (GA) is one of the useful tools for selecting the features which gives the highest predictor performance (Goldberg, 1989; Holland, 1975). GAs are systems that can successfully apply in many feature selection cases. The main principle task of GAs is transformation of dimensions into m-dimensional space (m<d) that can maximize a set of optimization (Raymer, Punch, Goodman, Kuhn, & Jain, 2000). In a previous Ph.D. Study, best eight HRV indices were chosen by a Genetic algorithm from among the above-mentioned 33 HRV indices (Hilavin, 2016). These best HRV indices constitute the new data set for us. Similar to our first study, within this framework, eight different size (1 to 8) sub datasets were obtained from these 33 HRV indices by the PCA algorithm. Alternatively, we utilized GAs+PCA to reduce 33- features. By this way, 8 main features were extracted by GAs and then these new 8 features used as new input features to PCA to reduce them further. As we mentioned before, we tried different distance matrices and different K values to analyze the performance of PCA and the results are summarized in Figure 6.8 to Figure 6.13.

It can be seen from these figures that reducing dimensions more than 5 decreased the performance for KNN in terms of accuracy and AUC. The most successful classification results came by 5 dimensions with which the classifier can detect PAF patients with an accuracy of approximately 89%.

In Figure 6.8 to Figure 6.10, we compared the different distance metrics vs constant number of neighbors (K=1,3 and 5). The results are similar to each other and performance of the system decreases after reducing dimension more than 5.

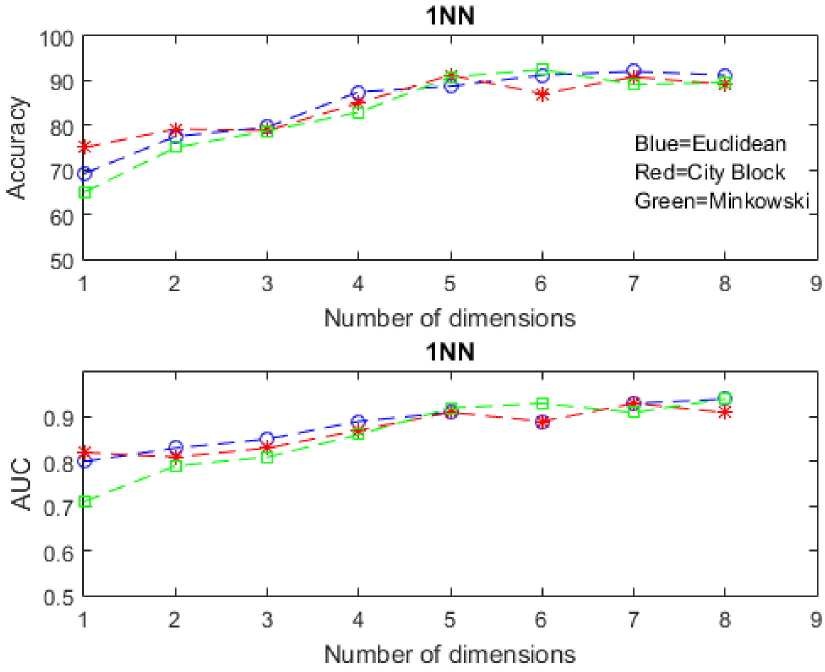

Figure 6.8 The results of PCA using GAs new dimensions as a new input in 1NN classifier with 3 different distance functions

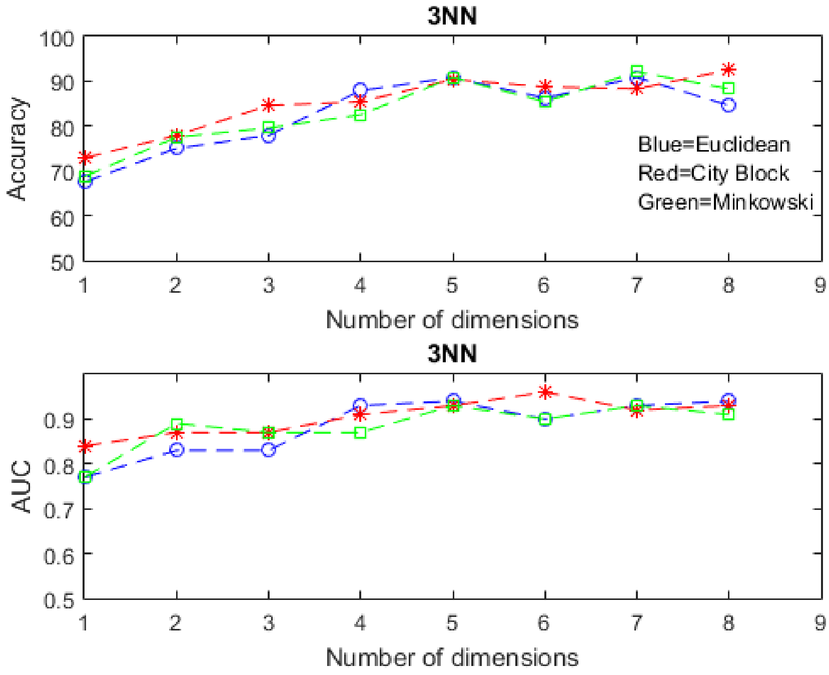

Figure 6.9 The results of PCA using GAs new dimensions as a new input in 3NN classifier with 3 different distance functions

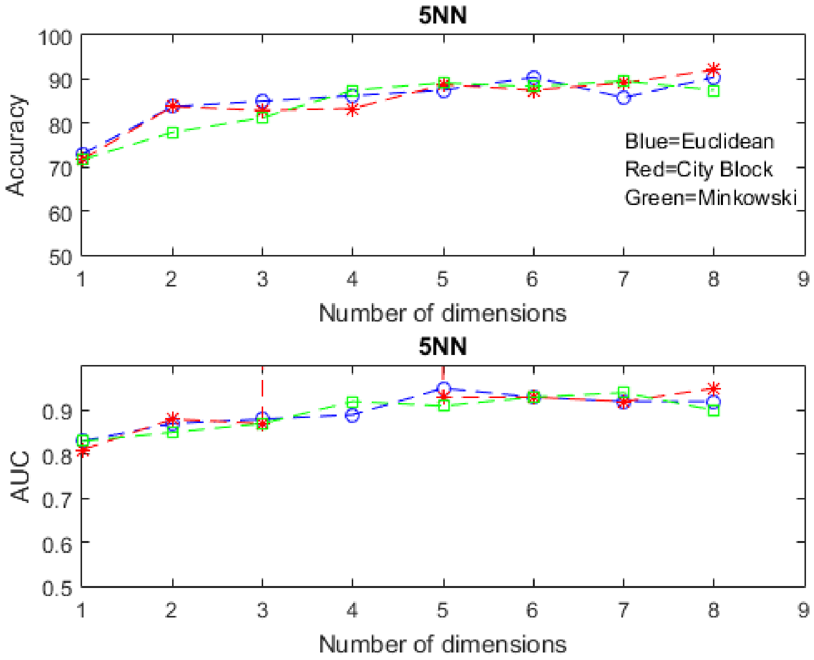

Figure 6.10 The results of PCA using GAs new dimensions as a new input in 5NN classifier with 3 different distance functions

In Figure 6.11 to Figure 6.13 we compared the different number of K (1,3 and 5) in constant distance metrics. First we selected Euclidean distance as a distance metric parameter and compared K=1,3 and 5. Then selected City Block as a constant distance metrics and again compared K=1,3 and 5 and applied same for Minkowski distance metrics.

The results elicit that reducing dimension by PCA is satisfactory until 4 whereas reducing more effects negatively on the results. Another interesting result is about the performance of different number of K after 4 dimensions. In Euclidean distance, the performance of K=5 is better than K=1 and it’s better than K=3 in terms of accuracy and AUC. For City Block distance metric, the performance of K=1,3 and 5 are very close to each other. Finally, for Minkowski distance metric, the performance of K=5 is better than K=3 and it’s also better than K=1.

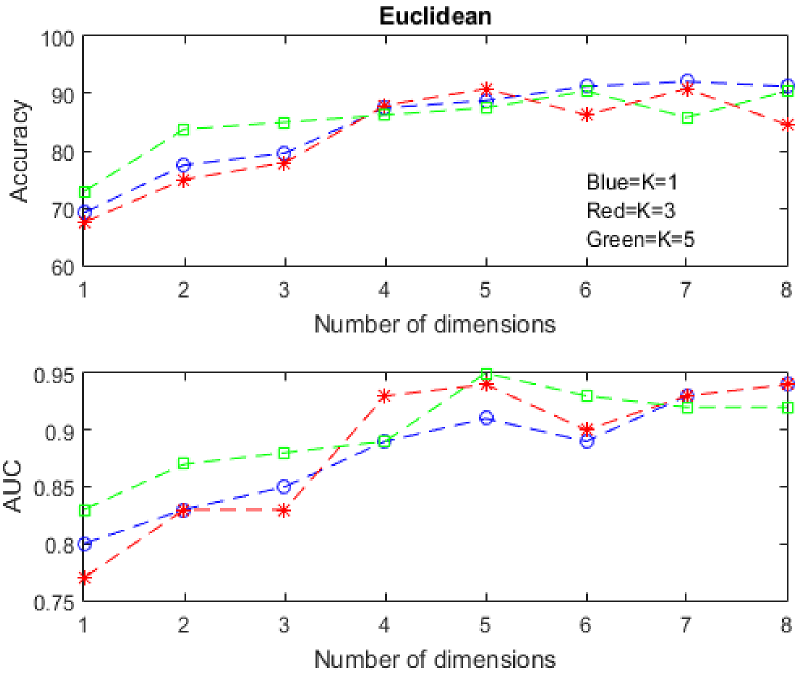

Figure 6.11 The performance evaluation of KNN with different K number in Euclidean distance by using GAs new features as a new input for PCA

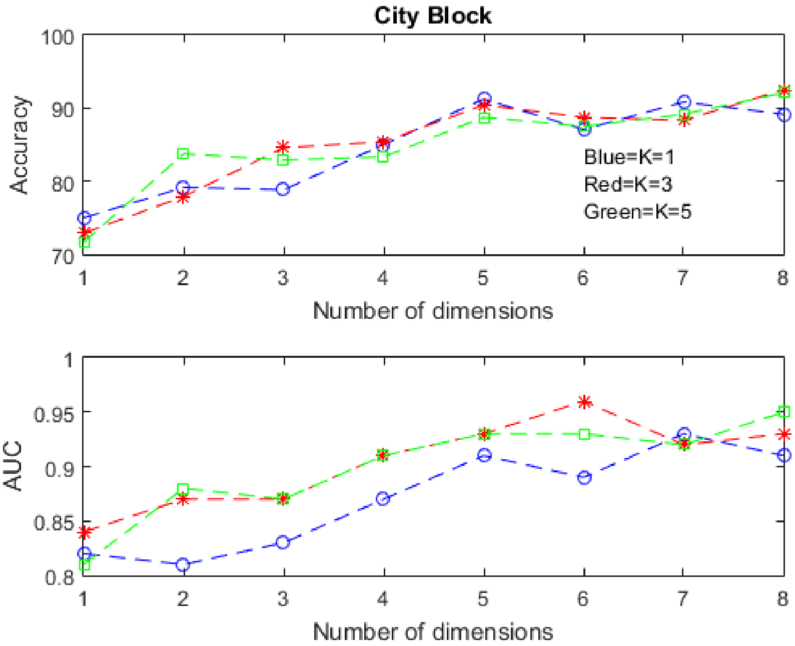

Figure 6.12 The performance evaluation of KNN with different K number in City Block distance by using GAs new features as a new input for PCA

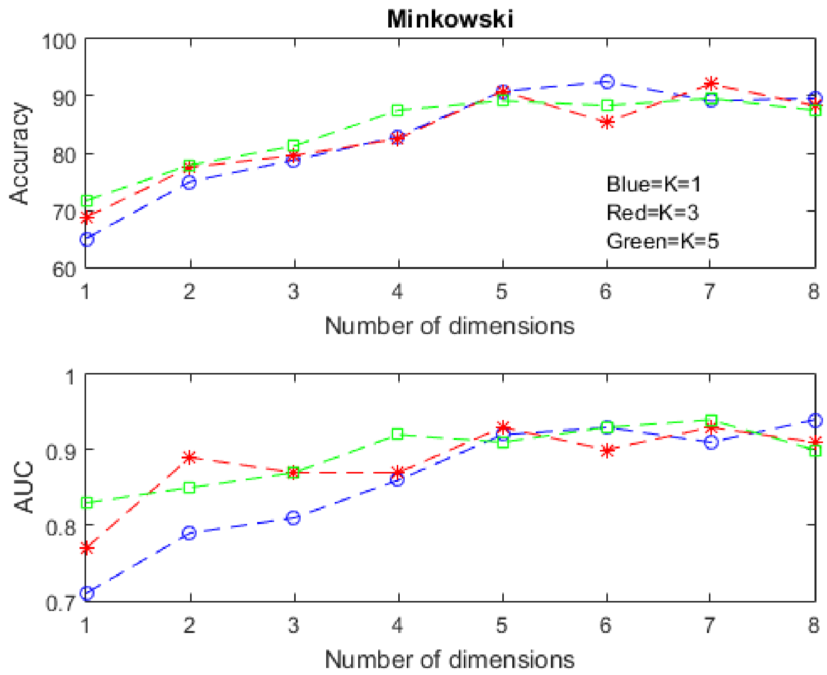

Figure 6.13 The performance evaluation of KNN with different K number in Minkowski distance by using GAs new features as a new input for PCA

Performance results for GAs+PCA in each number of K and different distance metrics can be seen in Table 6.20 to Table 6.28.

Table 6.20 Performance result of GAs+PCA with K=1 and Euclidean distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	Accuracy	AUC
1	**69.16**	**0.80**
2	77.50	0.83
3	79.58	0.85
4	87.50	0.89
5	88.75	0.91
6	91.25	0.89
7	**92.08**	**0.93**
8	91.25	0.94
Average	**84.63**	**0.88**

Table 6.21 Performance result of GAs+PCA with K=1 and City Block distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	**Accuracy**	**AUC**
1	**75**	**0.82**
2	79.16	0.81
3	78.83	0.83
4	85	0.87
5	**91.25**	**0.91**
6	87.08	0.89
7	90.83	0.93
8	89.16	0.91
Average	**84.53**	**0.87**

Table 6.22 Performance result of GAs+PCA with K=1 and Minkowski distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	**Accuracy**	**AUC**
1	**65**	**0.71**
2	75	0.79
3	78.75	0.81
4	82.91	0.86
5	90.83	0.92
6	**92.50**	**0.93**
7	89.16	0.91
8	89.58	0.94
Average	**82.96**	**0.85**

Table 6.23 Performance result of GAs+PCA with K=3 and Euclidean distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	**Accuracy**	**AUC**
1	**67.50**	**0.77**
2	75.00	0.83
3	77.91	0.83
4	87.91	0.93
5	**90.83**	**0.94**
6	86.25	0.90
7	90.83	0.93
8	84.58	0.94
Average	**82.60**	**0.88**

Table 6.24 Performance result of GAs+PCA with K=3 and City Block distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	**Accuracy**	**AUC**
1	**72.91**	**0.84**
2	77.91	0.87
3	84.58	0.87
4	85.41	0.91
5	90.41	0.93
6	88.75	0.96
7	88.33	0.92
8	**92.50**	**0.93**
Average	**85.10**	**0.90**

Table 6.25 Performance result of GAs+PCA with K=3 and Minkowski distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	Accuracy	AUC
1	**68.75**	**0.77**
2	77.50	0.89
3	79.58	0.87
4	82.50	0.87
5	90.83	0.93
6	85.41	0.90
7	**92.08**	**0.93**
8	88.33	0.91
Average	**83.12**	**0.88**

Table 6.26 Performance result of GAs+PCA with K=5 and Euclidean distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	Accuracy	AUC
1	**72.91**	**0.83**
2	83.75	0.87
3	85.00	0.88
4	86.25	0.89
5	87.50	0.95
6	**90.41**	**0.93**
7	85.83	0.92
8	90.41	0.92
Average	**85.25**	**0.89**

Table 6. 27 Performance result of GAs+PCA with K=5 and City Block distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	Accuracy	AUC
1	**71.66**	**0.81**
2	83.75	0.88
3	82.91	0.87
4	83.33	0.91
5	88.75	0.93
6	87.50	0.93
7	89.16	0.92
8	**92.08**	**0.95**
Average	**84.89**	**0.90**

Table 6.28 Performance result of GAs+PCA with K=5 and Minkowski distance. The best case, the worst case and the average values of metrics are highlighted

Number of features	Accuracy	AUC
1	**71.66**	**0.83**
2	77.91	0.85
3	81.25	0.87
4	87.50	0.92
5	89.16	0.91
6	88.33	0.93
7	**89.58**	**0.94**
8	87.50	0.90
Average	**84.11**	**0.89**

With regarding the Table 6.20 to Table 6.28, the best features are selected best on the nearest percentage to the average values. In Euclidean distance in KNN with K=1,3 and 5, the dimension reduction should be ceased until the accuracy values are below the average value. In this way, the best features are selected 4 since after 4 dimension the accuracy value is less than the average value.

In City Block distance metrics in KNN with K=1,3 and 5, if the accuracy value is 1% over the average value, the feature will be acceptable. In this way, the best feature is 5 since after 5 dimension the accuracy is less than 1% over the average value.

Finally, in Minkowski distance metrics in KNN with K=1,3 and 5, the dimension reduction can be accepted until the value of accuracy is 2% lower than the average value. By this way, the only acceptable features are 4 since by reducing further, the accuracy values are more than 2% below the average value.

In summarization, the best features in KNN (K=1,3 and 5) in Euclidean, City Block and Minkoeski distance metrics are 4,5 and 4, respectively.

6.3 Conclusion

Atrial fibrillation is an irregular heartbeat wherein the heart’s two upper chambers beat irregularly and out of coordination with the ventricles. A frequent cause of AF is atrial enlargement resulting from heart valve lesions. In PAF, atria contract uncoordinated with ventricles which are caused reduce the transfer of blood to the ventricles completely. Therefore, remaining blood inside the atria can lead to form clots. Forming clots in the aria, increasing the risk of stroke. Also, detecting PAF based on ECG is very difficult since it happens in very short time and has no specific symptoms.

Principal Component Analysis (PCA) is an offered scheme for feature extraction and dimension reduction. It has been used extensively in many applications involving high-dimensional data. In this study, we compared the effectivity of PCA features extracted from 33 short-term Heart Rate Variability (HRV) features obtained from normal sinus rhythm (NSR) ECG records for the diagnosis of Paroxysmal Atrial Fibrillation (PAF) disease. Within this framework, different data sets consisting of 33 to 1 features obtained from PCA were used as input to the classification algorithm, which is chosen as the K-Nearest Neighbor (kNN) algorithm. Different values for K and difference distance metrics were utilized to find the best performance.

In this study, the effectivity of features obtained from PCA for the detection of Paroxysmal Atrial Fibrillation (PAF) patients from their normal sinus rhythm (NSR) ECG records were analyzed. It is important to note that the aim of this study is not developing the best pattern classification system for PAF diagnosis; instead, the aim is comparing the performances of decreasing number of features obtained from PCA transformation. For this purpose, new features extracted by PCA from 33 short-term HRV indices (time domain, frequency domain and nonlinear) obtained from the ECG signals of both PAF and non-PAF subjects were used as the input. KNN classifier was utilized and different values of K (1, 3, 5) and different distance metrics (Euclidean, City Block, Minkowski) were tried and compared. The performances of different systems were compared in terms of accuracy and AUC.

In order to achieve the best result of dimension reduction, some specific percentage of average values were selected and based on these values, the best features were selected. Within this framework, in 1NN, 8 feature was selected as a best dimension in Euclidean and City Block distance metrics and 5 feature was selected as a best dimension in Minkowski distance metrics. By the way in 3NN, the best feature was selected 8 dimension in all distance metrics (Euclidean, City Block and Minkowski) and finally in 5NN the 8 feature was selected in Euclidean and Minkowski distance metrics and 9 feature was selected in City Block distance metrics.

Beside this study, the same procedure is applied to another HRV dataset. This set consists of 8 best HRV indices chosen from among the 33 HRV indices by a Genetic Algorithm. These 8 features used as new input to PCA and performance of GAs+PCA were analyzed and KNN classifier with different number of K=1,3 and 5 and different distance metrics parameter (Euclidean, City block and Minkowski) were utilized.

The best feature was selected based on the specific average value of accuracy. By this way, in Euclidean distance with different number of K=1,3 and 5 the proper feature is 4 and in City Block with different number of K (1,3 and 5) this feature was selected as 5 and finally in Minkowski distance metrics with different number of K (1,3 and 5) the proper dimension was selected as 4.

The obtained results from both studies elicit that it is possible to further reduce the number of input dimension of a classification system by using PCA algorithm without a reduction in the performance of the system. The advantages of the PCA in reducing input dimension are:

- Facilitating data understanding
- Reducing training time
- Reducing the memory requirements

In this thesis, the system constructed has some benefits because of using only HRV obtained features. Here, only 5-minuts RR interval data were used. On the other hand, morphological features are very sensitive to noise even though the R-R interval series data is resilient to noise. In the same way, utilizing only the R-R interval series decreases the processing time in comparison to ECG based methods.

Although the classification results wrapped up in this thesis appear sufficiently good, there is no demographic knowledge, for instance physiological activity, drug use, and psychic conditions that had to be considered during HRV analysis. Such information is neglected since there is no available details in the data which is used in this thesis. Moreover, only short-time HRV analysis were used in this study due to the fact that the length of the ECG records is 30 minutes.

6.4 Future Work

The goal of this work is to compare the performance of PCA as a dimension reduction tool for detecting Paroxysmal Atrial Fibrillation patients from normal sinus rhythm (NSR) ECG records by using short-term HRV features and also analyzing the effects of different distance matrix (Euclidean, City Block, Minkowski) and number of K (1,3 and 5) in KNN algorithm. For future work, different types of PCA such as Hebbian-Based PCA and Kernel PCA can be tried to see if better performances can be obtained.

- APEX algorithm also can be tried.
- The various PCA algorithms using neural networks may be categorized into two classes: reestimation algorithms and decorrelating algorithms can be checked.

Different types of LDA such as:

- Non-parametric LDA (Fukunaga)
- Orthonormal LDA (Okada and Tomita)
- Generalized LDA (Lowe)
- Multilayer Perceptrons (Webb and Lowe)

can be tried.

REFERENCES

Altman, N. S. (1992). An Introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46* (3), 175–185.

Bellman, R. (1961). Adaptive Control Processes. A Guide Tour. Princeton University Press.

Berne, R. M., & Levy, M. N. (1997). *Cardiovascular physiology* (7th edition). St. Louis: Mosby.

Betterman H, van Leeuwen P. (1998) Evidence of the dispersional analysis method for fractal time series. *Annals of Biomedical Engineering*, 23:49505.

Camm, A. J., Malik, M., Bigger, J., & Breithardt, G. (1996). Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, *17* (3), 354–381.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40* (1), 16–28.

Cowan, C., Campbell, J., V-Lin, C., Chung, G., Fay, M., Fitzmaurice, D., & Lip, G. (2014). *Atrial Fibrillation: the management of atrial fibrillation. NICE clinical guideline 180.*

Cysarz D, Bettermann H, Van Leeuwen P. (2005) Entropies of short binary sequences in heart period dynamics. *American Journal of Physiology*. *278* (6):H2163-H2172.

De Chazal, P., & Heneghan, C. (2001). Automated assessment of atrial fibrillation. In *Proceedings of Computers in Cardiology 2001*, *28,* 117–120.

De Vos, C. B., Pisters, R., Nieuwlaat, R., Prins, M. H., Tieleman, R. G., Coelen, R.-J. S., et al. (2010). Progression from paroxysmal to persistent atrial fibrillation: clinical correlates and prognosis. *The Journal of the American College of Cardiology, 55* (8), 725–731.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York: John Wiley & Son.

Evans, W. (1959). The management of paroxysmal atrial fibrillation. *Progress in Cardiovascular Diseases, 60* (2), 480–484.

Friberg, L., Hammar, N., & Rosenqvist, M. (2010). Stroke in paroxysmal atrial fibrillation: report from the Stockholm Cohort of Atrial Fibrillation. *European Heart Journal, 31* (8), 967–975.

Furberg, C. D., Psaty, B. M., Manolio, T. A., Gardin, J. M., Smith, V. E., Rautaharju, P. M., et al. (1994). Prevalence of atrial fibrillation in elderly subjects (the Cardiovascular Health Study). *The American Journal of Cardiology, 74* (3), 236–241.

Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. San Diego: Academic Press.

Getsch,M. (2003). *The ECG: a two-step approach to diagnosis.* Berlin: Springer-Verlage.

Gladstone, D. J., Spring, M., Dorian, P., Panzov, V., Thorpe, K. E., Hall, J., et al. (2014). Atrial fibrillation in patients with cryptogenic stroke. *New England Journal of Medicine, 370* (26), 2467–2477.

Go, A., Hylek, E., Phillips, K., & Chang, Y. (2001). Prevalence of diagnosed atrial fibrillation in adults - National implications for rhythm management and stroke prevention: The AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) study. *Journal of the American Medical Association*, *285* (18), 2370 – 2375.

Goldberg, D. E. (1989). *Genetic algorithms in search optimization and machine learning*. Menlo Park: Addison-Wesley Reading.

Goldberger, A., Amaral, L., Glass, L., & Hausdorff, J. (2000). PhysioBank, PhysioToolkit, and PhysioNet - Components of a new research resource for complex physiologic signals. *Circulation*, *101* (23), 215–220.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, *3*, 1157–1182.

Hall, John E. (John Edward). (1946)-Guyton and Hall textbook of medical physiology (12th edition). John Hall.

Hart, R. G., Pearce, L. A., Rothbart, R. M., McAnulty, J. H., Asinger, R. W., & Halperin, J. L. (2000). Stroke with intermittent atrial fibrillation: incidence and predictors during aspirin therapy. *Journal of the American College of Cardiology, 35* (1), 183–187.

Heykin, S. (1999). *Neural Network: A comprehensive foundation* (2nd ed). India: Sai PrintoPack Pvt. Ltd.

Hilavin, I. (2016). *Development of a system to diagnose Paroxysmal Atrial Fibrillation patients from arrhythmia free ECG records*. PhD Thesis, University of Dokuz Eylul, Izmir (in Turkey).

Holland, J. H. (1975). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.* Cambridge: MIT Press.

Hoshino, T., Ishizuka, K., Nagao, T., Shimizu, S., & Uchiyama, S. (2013). Slow sinus heart rate as a potential predictive factor of paroxysmal atrial fibrillation in stroke patients. *Cerebrovascular Diseases*, *36* (2), 120–125.

Iynegar N, Peng C-K, Morin R, Goldberger AL, Lipsiz LA. (1996) Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics. *American Journal of Physiology, Integrative and Comparative Physiology*. 271: R1078-R1084.

Jimenez, L. O., & Landgrebe, D. A. (1998). Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *28* (1), 39–54.

Kerr, C. R., Humphries, K. H., Talajic, M., Klein, G. J., Connolly, S. J., Green, M., et al. (2005). Progression to chronic atrial fibrillation after the initial diagnosis of paroxysmal atrial fibrillation: results from the Canadian Registry of Atrial Fibrillation. *American Heart Journal*, *149* (3), 489–496.

Klabunde, R. (2011). *Cardiovascular physiology concepts*. New York: Lippincott Williams & Wilkins.

Lee, T.-C., A.M. Peterson, & J.-C. Tsai, (1990). “A multi-layer feed-forward neural network with dynamically adjustable structures.” IEEE International Conference on Systems, Man, and Cybernetics, Los Angeles, CA.

Lévy, S., Camm, A. J., Saksena, S., Aliot, E., Breithardt, G., Crijns, H., et al. (2003). International consensus on nomenclature and classification of atrial fibrillation. *Europace*, *5* (2), 119–122.

Lip, G. Y. H., & Li Saw Hee, F. (2001). Paroxysmal atrial fibrillation. *QJM: an international journal of medicine*, *94* (12), 665–678.

Lowe, D., and A.R. Webb, 1991b. "*Optimized feature extraction and the Bayes decision in feed-forward classifier networks*." IEEE Transactions on Pattern Analysis and Machine Intel-ligence.

Majeed, A., Moser, K., & Carroll, K. (2001). Trends in the prevalence and management of atrial fibrillation in general practice in England and Wales, 1994-1998: analysis of data from the general practice research database. *Heart, 86* (3), 284–288.

Maier, C., Bauch, M., & Dickhaus, H. (2001). Screening and prediction of paroxymal atrial fibrillation by analysis of heart rate variability parameters. *In Proceedings of Computers in Cardiology 2001, 28,* 129–132.

Martínez, A., Alcaraz, R., & Rieta, J. J. (2014). Morphological variability of the P-wave for premature envision of paroxysmal atrial fibrillation events. *Physiological Measurement, 35* (1), 1–14.

Mitchell, T. M. (1997). *Machine learning*. Boston: McGraw-Hill Companies Press.

Moody, G., Goldberger, A., McClennen, S., & Swiryn, S. (2001). Predicting the onset of paroxysmal atrial fibrillation: the Computers in Cardiology Challenge 2001. *In Proceedings of Computers in Cardiology 2001, 28,* 113–116.

Page, R. L., Wilkinson, W. E., Clair, W. K., McCarthy, E. a, & Pritchett, E. L. (1994). Asymptomatic arrhythmias in patients with symptomatic paroxysmal atrial fibrillation and paroxysmal supraventricular tachycardia. *Circulation, 89* (1), 224–227.

Peng C-K, Havlin S, Stanley HE, Goldberger AL. (1995) Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos*. *5*:82-87.

Personnaz, L., I. Guyon, and G. Dreyfus, (1985). “Information storage and retrieval in spin-glass like neural network. *Journal de Physique*, *Lettres*, *46* (8), 359-365 (Orsay, France).

Pincus SM, Goldberger AL. (1994) Physiological time-series analysis: What does regularity quantify? *American Journal of Physiology-Heart and Circulatory Physiology*. 266:H1643-H1656.

Purves, D., Augustine, G. J., Fitzpatric, D., Hall, W. C., LaMantia, A.-S., & White, L. E. (2012). Autonomic control of cardiovascular function. In R. D. Mooney & M. L. Platt (Eds.), *Neuroscience* (5th ed.). Sunderland: Sinauer Associates, Inc.

Raymer, M. L., Punch, W. F., Goodman, E. D., Kuhn, L. A., & Jain, A. K. (2000). Dimensionality reduction using genetic algorithms. *IEEE Transactions on Evolutionary Computation*, *4* (2), 164–171.

Richman JS, Moorman JR. (2000) Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology. 278* (6):H2039-2049.

Sanna, T., Diener, H.-C., Passman, R. S., Di Lazzaro, V., Bernstein, R. A., Morillo, C. A., et al. (2014). Cryptogenic stroke and underlying atrial fibrillation. *New England Journal of Medicine*, *370* (26), 2478–2486.

Savelieva, I., & Camm, A. J. (2000). Clinical relevance of silent atrial fibrillation: prevalence, prognosis, quality of life, and management. *Journal of Interventional Cardiac Electrophysiology*, *4* (2), 369–382.

Sherwood, L. (2015). *Human physiology: from cells to systems (*9th edition). Boston: Cengage learning.

Todd, W. (2013). *Conduction Anatomy*. Retrieved April 20, 2016, from 134 http://wesleytodd.blogspot.com.tr/2013/06/triangle-koch.html.

Valafar, F. (2000). Applications of Neural Networks in Medicine and Biological Sciences. In A. Zilouchian & M. Jamshidi (Eds.), *Intelligent Control Systems Using Soft Computing Methodologies*. Boca Raton: CRC Press, Inc.

Van Gelder, I. C., & Hemels, M. E. W. (2006). The progressive nature of atrial fibrillation: a rationale for early restoration and maintenance of sinus rhythm. *Europace : European Pacing, Arrhythmias, and Cardiac Electrophysiology : Journal of the Working Groups on Cardiac Pacing, Arrhythmias, and Cardiac Cellular Electrophysiology of the European Society of Cardiology*, *8* (11), 943–949.

Webster, J. (1998). *Medical instrumentation: application and design*. New York: John Wiley & Sons, Inc.

Yanowitz, F. G. (2012). *Introduction to ECG interpretation V8. 0*. Salt Lake City: Intermountain Healthcare.