

DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**CLUSTERING OF HRV DATA OBTAINED FROM
NORMAL SINUS RHYTHM ECG RECORDS OF
PAF AND NON-PAF SUBJECTS USING A NEW
ENSEMBLE METHOD**

by
Omid ALIGHOLIPOUR

July, 2018
İZMİR

CLUSTERING OF HRV DATA OBTAINED FROM NORMAL SINUS RHYTHM ECG RECORDS OF PAF AND NON-PAF SUBJECTS USING A NEW ENSEMBLE METHOD

**A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfilment of the Requirements for the Degree of Master of
Science in Electrical and Electronics Engineering**

**by
Omid ALIGHOLIPOUR**

**July, 2018
İZMİR**

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled “**CLUSTERING OF HRV DATA OBTAINED FROM NORMAL SINUS RHYTHM ECG RECORDS OF PAF AND NON-PAF SUBJECTS USING A NEW ENSEMBLE METHOD**” completed by **OMID ALIGHOLIPOUR** 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.



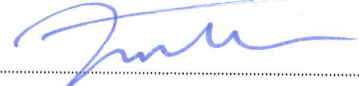
Prof. Dr. Mehmet KUNTALP

Supervisor



Doc. Dr. Olcay Akay

(Jury Member)



Prof. Dr. Törker Ince

(Jury Member)



Prof. Dr. Latif SALUM

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

First, I would like to express my sincere to my supervisor Prof. Dr. Mehmet KUNTALP for useful comments, remarks and gratitude throughout this study. I have benefited greatly from his indisputable experience on academic background and his very kind behavior.

Also, as I am far away from my home town, I would like to give my special thanks to Zehra Ünalın who really helped me in every aspect. In addition, I'd like to thank my friends Safa, Salih, Kahraman and Batuhan and my other homemates which were always there and motivated me during this time.

Finally, I would like to extend special thanks to my mom and dad, Maryam and Mohammed, for their endless sacrifices, support and patience during my life.

Omid ALIGHOLIPOUR

CLUSTERING OF HRV DATA OBTAINED FROM NORMAL SINUS RHYTHM ECG RECORDS OF PAF AND NON-PAF SUBJECTS USING A NEW ENSEMBLE METHOD

ABSTRACT

Atrial fibrillation (AF) is one of the most common diseases in which the atria cannot completely push the blood to ventricles and therefore clot formation could occur which may contribute to serious health problems. In most cases, AF starts with short episodes and progresses to longer and developed form over time. With respect to the acceleration of the number of patients, certain criteria have been used to determine the disease in order to start necessary treatment. Paroxysmal AF is a type of AF in which the episodes could last from minutes to days and ends by itself. If PAF condition continues, it could be converted to Persistence AF. Therefore, it is very important to determine PAF patients and treat them accordingly. The determination of PAF condition is very easy by recording the ECG of the subjects during an episode. However, since PAF episodes mostly last for a short time period, it is very difficult to record the ECG of the subjects during the arrhythmic event. Therefore, a system that would be able to detect PAF patients based on ECG records taken during non-arrhythmic time periods would be very beneficial. The aim of this study is to analyze the structure of the data obtained from arrhythmia-free ECG records of PAF and non-PAF subjects. In other words, the separability of the two types of data is to be investigated. In this study, the dataset which contains HRV features obtained from normal sinus rhythm (NSR) ECG records of non-PAF and PAF subjects is aimed to be clustered by using two most famous unsupervised classification methods; K-Means and Fuzzy C-Means algorithms. The obtained results show that there is a significant overlap between the two types and thus there is a need for a good classifier.

Keywords: Paroxysmal atrial fibrillation, Clustering, K-means, Fuzzy C-means, hybrid clustering

PAF HASTASI OLAN VE OLMAYAN KİŞİLERİN NORMAL SİNÜS RİTİM EKG KAYITLARINDAN ELDE EDİLEN KHD VERİLERİNİN YENİ BİR TOPLULUK METODU KULLANARAK KÜMELENMESİ

ÖZ

Atriyal fibrilasyon (AF) atriyumun kanı ventriküle tamamen itemediği en yaygın hastalıklardan biridir ve bu nedenle ciddi sağlık sorunlarına katkıda bulunabilecek pıhtı oluşumu meydana gelebilir. Çoğu durumda, AF kısa episodlarla başlar ve zamanla ilerleme kaydedip gelişir. Hasta sayısının hızlanarak artması ile birlikte, belirli kriterler hastalığın seyrini belirlemek ve gerekli tedaviye başlatmak için kullanılmıştır. Paroksizmal AF, atakların dakikalar, saatler bazen de günlerce sürebileceği ve kendiliğinden sona ereceği bir AF türüdür. PAF durumu devam ederse, Persistent AF'ye dönüşecektir. Bu nedenle, PAF hastalarını belirlemek ve onlara göre uygun tedavi uygulamak çok önemlidir. Hastaların aritmi sırasındaki EKG'sini kaydetmek suretiyle bu hastalığın belirlenmesi çok kolaydır. Ancak, PAF episodları genelde kısa sürdüğünden tam aritmik olay sırasında hastanın EKG'sini kaydedebilmek çok zordur. Bu nedenle, aritmik olmayan zaman aralıklarında alınan EKG kayıtlarına dayanarak PAF hastalarını saptayabilecek bir sistem çok yararlı olacaktır. Bu çalışmanın amacı, PAF hastası olan ve olmayan kişilerin aritmi içermeyen EKG kayıtlarından elde edilen verilerin yapısını analiz etmektir. Diğer bir deyişle, iki veri tipinin ayrılabilirliği araştırılacaktır. Bu amaca yönelik olarak, PAF hastası ve PAF hastası olmayan kişilerin aritmi bulunmayan EKG kayıtlarından elde edilen HRV endekslerini içeren veri tabanının, en ünlü iki öğreticisiz sınıflandırma yöntemi olan K-Means ve Bulanık C-Means algoritmalarını kullanarak kümelenmesi baz alınmıştır. Elde edilen sonuçlar, iki sınıf arasında önemli bir örtüşme bulunduğunu ve bu nedenle iyi bir sınıflandırıcıya ihtiyaç olduğunu göstermektedir.

Anahtar kelimeler: Paroksizmal atriyal fibrilasyon, kümeleme, k-means, bulanık c-means, topluluk kümelenme

CONTENTS

	Page
M.Sc THESIS EXAMINATION RESULT FORM.....	ii
ACKNOWLEDGMENTS.....	iii
ABSTRACT	iv
ÖZ.....	v
LIST OF FIGURES.....	ix
LIST OF TABLES	x
CHAPTER ONE - INTRODUCTION	1
CHAPTER TWO - PHYSIOLOGICAL BACKGROUND.....	4
2.1 Inside the Heart.....	4
2.2 Blood Circulation of Heart	6
2.3 Cardiac Electro-Conduction System.....	6
2.3.1 Sinoatrial Node	7
2.3.2 Atrioventricular Node	7
2.3.3 Bundle of His and Purkinje Fibers.....	7
2.4 Cardiac Electrophysiology.....	8
2.5 Electrocardiography.....	9
2.6 ECG Interpretation.....	10
2.7 Arrhythmias	13
2.7.1 Ventricular Arrhythmias	14
2.7.1.1 Ventricle Tachycardia	14
2.7.1.2 Ventricular Flutter	14
2.7.1.3 Premature Ventricular Contractions.....	14
2.7.1.4 Ventricular Fibrillation.....	15
2.7.2 Supraventricular Arrhythmias.....	15
2.7.2.1 Atrial Fibrillation.....	15
2.7.2.2 Atrial Flutter.....	15
2.7.2.3 Supraventricular Tachycardia (SVT)	16
2.7.2.4 Premature Supraventricular Contractions	16

2.7.2.5 Sinus Tachycardia	16
2.8 Atrial Fibrillation	16
CHAPTER THREE - HEART RATE VARIABILITY.....	19
3.1 Time Domain HRV Indices	20
3.2 Frequency Domain HRV Indices.....	22
3.3 Nonlinear methods	23
CHAPTER FOUR - CLUSTERING	25
4.1 Data Clustering	25
4.2 Clustering Algorithms	28
4.2.1 Notation and Terminology	28
4.2.1.1 Data	29
4.2.1.2 Similarity Measure	29
4.2.1.3 Cluster Center Initialization	30
4.2.1.4 Number of Clusters	31
4.2.2 K-Means Clustering	32
4.2.3 Fuzzy C-means Clustering.....	35
4.3 Ensemble Learning	39
4.4 Model Assessment	40
4.4.1 Cross-Validation	40
4.4.2 Cluster Validation	41
4.4.3 Accuracy Measure	43
CHAPTER FIVE - METHODS AND RESULTS	49
5.1 Methods and Parameters.....	50
5.1.1 Determining the Number of Clusters.....	50
5.1.2 Cross-Validation	52
5.2 Experiments	52
5.2.1 Single Clustering Algorithms	53
5.2.2 Ensemble Models.....	53

CHAPTER SIX - CONCLUSION	60
REFERENCES	63



LIST OF FIGURES

	Page
Figure 2.1 Structure of the heart.....	5
Figure 2.2 Blood circulation.....	6
Figure 2.3 Electrical conduction system	7
Figure 2.4 Location of leads.....	10
Figure 2.5 Standard ECG recording paper (mV)	11
Figure 2.6 Each wave represent the statues of the chambers	12
Figure 2.7 Segment representation of the ECG signal	13
Figure 2.8 Typical ECG signals. A) Normal ECG signal; B) AF signal	17
Figure 3.1 Deriving RR interval from ECG	20
Figure 4.1 Classification of clustering algorithms	26
Figure 4.2 Dendogram of dataset with Divisive and Agglomerative methods	28
Figure 4.3 Overall perspective of clustering procedure	29
Figure 4.4 Elbow represents the optimal number of cluster	32
Figure 4.5 A fundamental representation of the combining clustering algorithms. A and B are different algorithms and a, b, c and d are different combinations we can form.....	40
Figure 4.6 The scheme of dividing dataset to 5-fold using cross validation.....	41
Figure 5.1 The Elbow can be distinguished at $k=2$	50
Figure 5.2 Silhouette plots for a) $k=2$ b) $k=3$ c) $k=4$ d) $k=5$ e) $k=6$	51
Figure 5.3 The scheme of dividing dataset to 4-folds using cross validation. Total accuracy is achieved by finding the mean of all estimations	52
Figure 5.4 Procedure of a typical clustering analysis. Box A can be any clustering algorithm	53
Figure 5.5 Ensemble of different algorithms based on membership values	54
Figure 5.6 An ensemble approach based on fuzzy membership restriction rule followed by 3 different clustering methods.....	56
Figure 5.7 Comprehensive ensemble clustering approach.....	57

LIST OF TABLES

	Page
Table 3.1 Time domain HRV measurements	21
Table 3.2 Frequency bands of HRV	22
Table 3.3 HRV features used in this study	24
Table 4.1 K-means algorithm	34
Table 4.2 Fuzzy C-means algorithm	36
Table 4.3 K-means validity indices	44
Table 4.4 Fuzzy C-means validity indices	45
Table 4.4 Continued	46
Table 4.5 The confusion matrix	47
Table 5.1 Average Silhouette values for different number of clusters.....	51
Table 5.2 Results of clustering algorithms after 20 times implementation.....	55
Table 5.3 Ensemble model results after 20 times iteration	55
Table 5.4 Results of ensemble model by considering different threshold values. Threshold values are presented in percentage. For each step, accuracy is calculated	58
Table 5.5 Final performance of ensemble model with restriction on membership values followed by K-means	58
Table 5.6 Amount of number of unassigned data points for various threshold values	59

CHAPTER ONE

INTRODUCTION

Cardiovascular system is one of the essential and crucial systems of the body. Heart produces enough pressure to pump the blood to all parts of the body. Recording of the heart's electrical activity is called Electrocardiography (ECG or EKG) and any abnormal change in cardiac rhythm is known as arrhythmia. Atrial Fibrillation (AF) is a type of arrhythmias caused by abnormal functioning of the atria. Owing to malfunction of the atria, blood doesn't completely fill into the ventricles. This increases the risk of the blood clots in the atria and it may lead to stroke, heart failure, sudden death and other heart related complications (Davies & Scott, 2015; Gacek & Pedrycz, 2012; Najarian & Splinter, 2012). In 2010, it was estimated that about 33.5 million people suffered from AF. By 2030, approximately 14-17 million AF patients are anticipated in European Union (Kirchhof, Benussi, Kotecha, Ahlsson, Atar, Casadei, et al, 2016).

There are three types of AF:

1. Paroxysmal AF
2. Persistent AF
3. Permanent AF

In some cases, AF episodes spontaneously start and terminate within minutes, hours or days without any external interference. In such cases, it is called Paroxysmal AF (PAF). AF episodes lasting longer than 7 days, including those that can only be terminated by drugs or cardioversion are called Persistent AF. To distinguish Paroxysmal and Persistent AF, long term ECG monitoring might be considered. In Permanent AF, episodes last more than 1 year and usually don't stop by itself or external interference.

According to studies (Proietti et al., 2015; De Vos et al., 2010), there is possibility for paroxysmal AF patients to develop into Persistent or Permanent AF over time.

Hence, detecting paroxysmal AF in earlier stages will not only help physicians to avoid dangerous health problems, but help to prevent the development of disease.

Determination of PAF condition is very easy by recording the ECG of the patients during a PAF episode. However, since PAF episodes last for a short period of time, it is usually very difficult to record the ECG of the subjects during the arrhythmic event. Therefore, a system that would be able to detect PAF patients based on ECG records taken during non-arrhythmic time periods will be very beneficial.

In the literature, there are studies related to detecting PAF patients from normal sinus rhythm ECG records. One way to detect PAF patients was proposed by Nicole Kikillus, et al. (2007). In this method, after normalizing the RR intervals, the time domain analysis was done. Then, Poincare Plot, which is produced by using two sequences of RR interval points, was generated. The image and the time domain analysis assess a risk level, which determines whether the patient is suffering from atrial fibrillation. There are about 20 risk levels and when a data has a value more than 2, it defines a PAF subject. Gokana et al. used RMSSD of RR interval and autocorrelation to determine AF (Gokana, Phua, Lissorgues, 2014). Considering 3 different types of features from various regions of ECG and classifying by 3 different algorithms, a study was proposed to detect and predict the PAF subjects and examined best feature group for detection and prediction purposes separately (Pourbabaei, & Lucas 2008). Moreover, Donoso et al. introduced a preliminary study about analyzing the structure of feature space of ECG records in order to discover the separability of the subclasses of AF by using K-means and hierarchical clustering algorithms (Donoso et al., 2013).

This study aims to evaluate the separability of two groups (PAF and non-PAF) with respect to HRV features obtained from normal sinus rhythm (NSR) ECG records produced by Hilavin (Hilavin, 2016) by employing various clustering methods. For this purpose, the features of PAF and non-PAF subjects were clustered by K-means, Fuzzy C-means, GIPF-FCM and s-FCM algorithms. Furthermore, two ensemble clustering techniques were proposed to obtain better examination about the structure of the datasets.

In the literature, various studies have been conducted in order to compare K-means and Fuzzy C-means algorithms. In a recent study (Shedthi, Shetty, Siddappa, 2017), unsupervised classification algorithms were implemented on agricultural data. In addition, some studies evaluated the performance of K-means and Fuzzy C-means algorithms with respect to iteration, speed, time and accuracy (Panda, Sahu, Jena, Chattopadhyay, 2012; Ghosh & Dubey, 2013; Bora, Gupta, Kumar, 2014; Cebeci & Yildiz, 2015). These studies resulted in better performance of K-means than Fuzzy C-means. Furthermore, in the last decades many modifications of these algorithms were introduced to enhance the quality of clustering. Some of these methods will also be discussed in the following sections.

Chapter 2 briefly explains physiological background of cardiovascular system. HRV methods are discussed in Chapter 3. Unsupervised classification methods and related concepts are stated in Chapter 4. Chapter 5 focuses on the results obtained and performance evaluation of different clustering algorithms. Also, new models of Fuzzy C-means algorithms are introduced in this chapter. Finally, in Chapter 6, discussion and conclusion of this thesis are mentioned.

CHAPTER TWO

PHYSIOLOGICAL BACKGROUND

The location of human heart is in the chest between lungs and behind sternum and it weighs between 200 to 425 grams. The main responsibility of heart is to pump blood to all tissues and organs. At each heartbeat, heart sends blood to all tissues and organs and back again. In a healthy person heart beats approximately 100000 times in a day and about 2.5 billion times during his or her lifetime. The effectiveness of heart can be measured by the muscles around the heart, myocardium. Therefore, any changes in the pumping system can affect our life cycle. In the following sections, detailed information about structure of heart and its activity will be discussed.

2.1 Inside the Heart

Heart is the main component of our circulatory system. Beside heart, blood vessels and blood are other component of this system. A normal structure of heart is shown Figure 2.1. Heart wall is made of 3 layers:

- Epicardium
- Myocardium
- Endocardium

Epicardium is the thinnest and outer layer of the heart and it lubricates and protects outside of the heart. Myocardium is the thickest wall and includes cardiac muscle tissue. Cardiac muscle tissue is a specialized muscle which is found only in heart. Myocardium makes heart to contract. The performance of the pumping depends upon the contraction and relaxation of myocardium. The thickness of the myocardium determines the strength of the heart pumping. The endocardium is the layer that lines inside of the heart and it is responsible for preventing blood from leaking to the inside of the heart and forming potentially deadly blood clots.

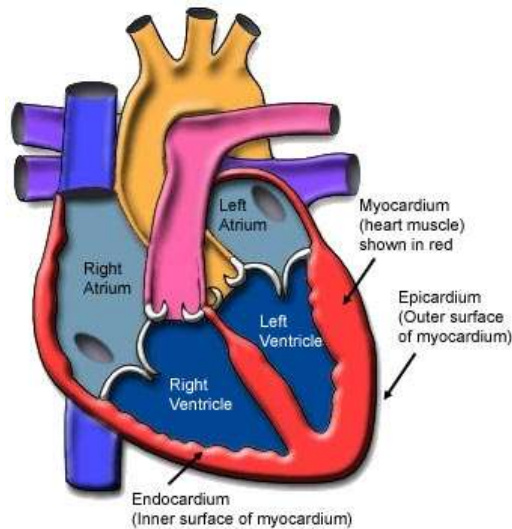


Figure 2.1 Structure of the heart

The heart consists of 4 chambers: Right Atrium, Left Atrium, Right Ventricle and Left Ventricle. Atria are generally smaller, thinner and less muscular than ventricles. Blood circulation of the heart is illustrated in Figure 2.2 where atriums receive dirty blood from all parts of the body (left atria receives fresh blood from lungs) and ventricles pump it.

Furthermore, there are 4 valves which make the chambers connected:

- Tricuspid valve: between right atrium and right ventricle
- Mitral valve: between left atrium and left ventricle
- Pulmonic valve: between right ventricle and lungs
- Aortic valve: between left ventricle and aorta

Blood goes through the valve from chambers. The main function of valves is to enforce the blood to go in one direction. When a valve is open in one side, other is closed and without closing of an opened valve, the other valve doesn't open. The valves are made of leaflets and are so strong to avoid any shape change. When pressure in atria is higher than ventricle, the atrioventricular valves open and blood goes to ventricle and when the pressure of ventricle exceeds the atria pressure, these valves close to prevent the blood from flowing back to atrium.

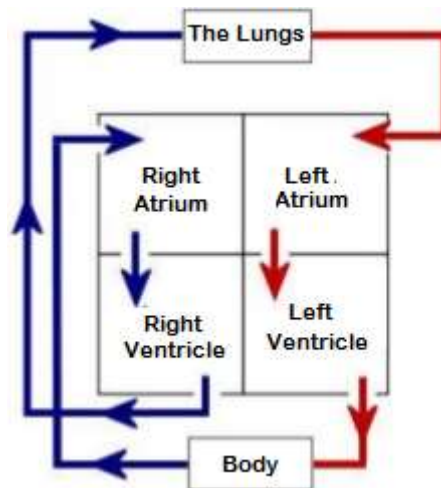


Figure 2.2 Blood circulation

2.2 Blood Circulation of Heart

Because left and right side of the heart generally works at the same time, at each heartbeat, both ventricles or atria contract or relax at the same time.

Blood circulation can be described as follows:

- Step 1: Oxygen-poor blood enters right atrium through vena cava, while oxygenated blood enters to left atrium.
- Step 2: As the atrium contracts, blood flows from right atrium to right ventricle through the tricuspid valve. In the left hand side, oxygen-rich blood goes to left ventricle through open mitral valve.
- Step 3: When the ventricle reaches its maximum, tricuspid and mitral valves are closed to prevent blood from following to atrium. After ventricle contracts, oxygen-poor blood goes to lung through pulmonic valve and oxygen-rich blood flows to the body through the aortic valve.

2.3 Cardiac Electro-Conduction System

Heart beat is a result of generation and conduction of the electrical impulses. There are a group of specialized muscle cells in the wall which sends signals to the

heart muscles (Figure 2.3). The electrical system, also known cardiac conduction, includes the components below.

2.3.1 Sinoatrial Node

Sinoatrial node (or SA node) is located in the right upper atrium. Generated impulses by SA node travel throughout the heart wall and make atriums to contract and push blood to ventricles.

2.3.2 Atrioventricular Node

Atrioventricular node (or AV node) is located in the wall of the right atrium of the heart. It receives electrical impulses from the SA node and directs them to the conduction system in the walls of the ventricles. The most important role of AV node is to delay signals reaching from SA node by several milliseconds. This delay ensures that atria completely empty its whole blood to ventricle before contraction of ventricle.

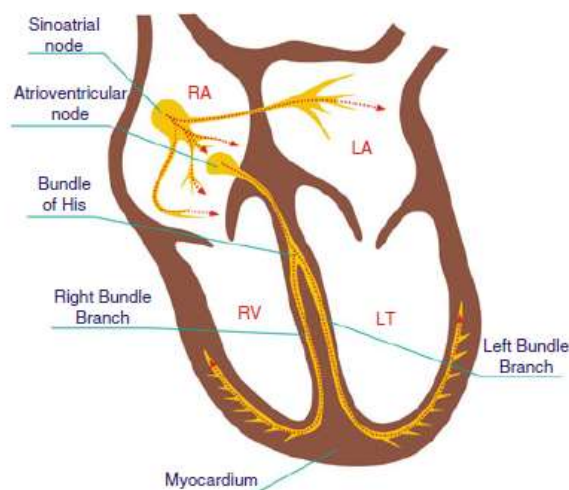


Figure 2.3 Electrical conduction system (Gacek & Pedrycz, 2011)

2.3.3 Bundle of His and Purkinje Fibers

Bundle of His consists of specialized muscle fibers which carry impulses throughout the ventricles. There are 2 bundle branches; the right bundle carries nerve

impulses that cause contraction of the right ventricle and the left bundle carries nerve impulses that cause contraction of the left ventricle. After delay in AV node, signal is transmitted to Purkinje fibers via these bundles. When the signal reaches to Purkinje fibers (which covers throughout the ventricles), signal travels to epicardial surface of myocardium.

2.4 Cardiac Electrophysiology

Cardiovascular system circulates blood throughout the body in order to supply oxygen and other necessary nutrients to tissues and remove wastes from tissues. As discussed before, SA node has an intrinsic behavior that fire the heartbeat, but actually this is modified by The Autonomic Nervous System. The Autonomic Nervous System (ANS) is firstly defined by Langley in 1921 and is the system which controls all automated body functions consisting heart rate, blood pressure, metabolism and digestion. There are two main divisions of autonomic nervous system: Sympathetic and Parasympathetic nerves. These two nerve systems have a very good and accurate control on the organs (Clifford, Azuaje, McSharry, 2006).

Sympathetic Nervous System (SNS) controls the body's autonomic reaction to physical and psychological activities. It increases heart rate and blood pressure. As a consequence, it makes SA node to fire more than normal. On the other hand, Parasympathetic Nervous System (PNS) is responsible for body functions during rest time. It slows down the heart rate, decreases the blood pressure and stimulates digestion. Parasympathetic nervous system resets organ functions after the sympathetic nervous system is activated. These nerves act quickly and decrease the velocity through AV node which makes that heart beat slowly.

The sympathetic and parasympathetic nervous systems usually do opposite things and the level of activation of these nerves is adjusted by body. For a healthy person, in the case where both sympathetic and parasympathetic nerves act normally, heart rate will be 100 bpm. ANS adjusts the heart rate by the information it receives from the sensors in different parts of the body.

2.5 Electrocardiography

Monitoring of the electrical activity of heart is an easy way to determine any change and abnormal activity in the process of heartbeat. The potential changes between depolarization and polarization of cardiac muscle cells can be monitored by placing electrodes on the surface of chest and limb. Recording of these signals over a period of time is called Electrocardiogram (ECG or EKG).

Willem Einthoven's galvanometer was one of the first ECG recording machines which was more accurate and precise than others (Gacek & Pedrycz, 2011). To determine the ECG of an individual, standard 12-lead ECG recording is used to consider 12 different perspectives. ECG leads are mainly grouped into two categories:

- Frontal leads
- Transverse leads

Einthoven introduced bipolar limb leads to determine depolarization of heart. This type of leads is used in the frontal plane and includes positive and negative electrodes. These leads are placed on left arm, right arm, left leg which are known as LA, RA and LL, respectively. Additionally, one lead is placed on right leg (RL) as neutral and helps to minimize ECG artifact. These 3 leads make up a triangle known as Einthoven's Triangle (Figure 2.4). Besides the information directly obtained by these leads, combination of them also gives information about vertical plane of heart:

- Lead I: Information between aV_R and aV_L
- Lead II: Information between aV_R and aV_F
- Lead III: Information between aV_L and aV_F

According to Einthoven's law, if we add all voltage signals from these three leads, the net result is zero volts (Davies & Scott, 2015). According to Wilson central terminal, a resistor with 5k resistance connecting to terminal of each of limb leads to a common point called central terminal.

$$V_w = \frac{Er + Ef + El}{3} \quad (2.1)$$

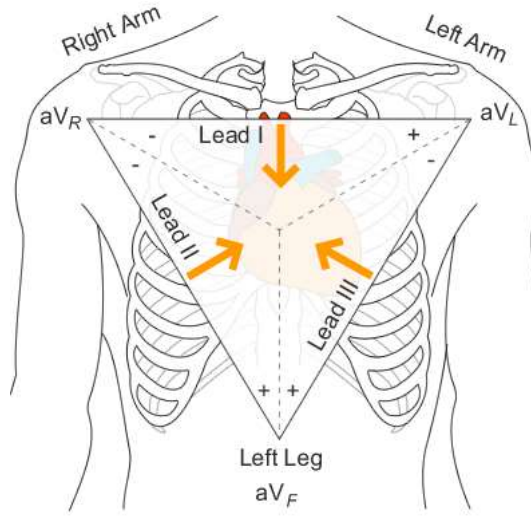


Figure 2.4 Location of leads

Owing to a high impedance voltmeter, Kirchhoff's law can be used

$$I + II + III = 0 \quad (2.2)$$

There are 6 precordial chest leads (V1-V6). These leads are unipolar and located over the left chest. Overall, we have 10 leads with 12 different perspectives.

2.6 ECG Interpretation

A special type of paper is used to describe and record ECG signals over a period of time. Limb leads are represented in the left hand side where as chest leads are shown in the right hand side (Figure 2.5). Every small box has 1 mm width and height of these small boxes measure 0.04 second. In addition, each small square represents 0.1 mV (Davies & Scott, 2015). A normal ECG beat consists of different segments which are described in Figure 2.6 and Figure 2.7.

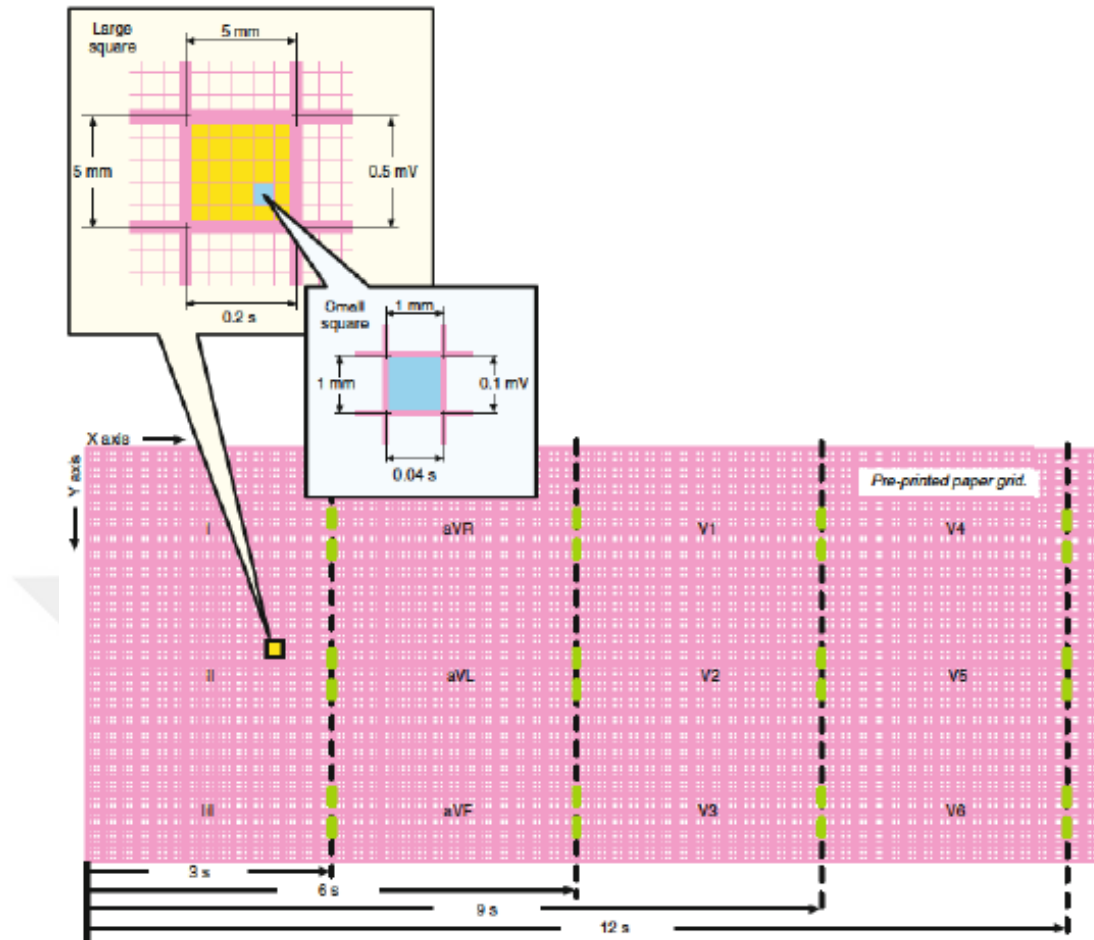


Figure 2.5 Standard ECG recording paper (mV) (Davies & Scott, 2015)

- P-wave: Arterial depolarization is represented by this wave which occurs before QRS complex. Leads II and chest lead V1 are good ones to review and monitor this wave. The amplitude of P-wave should not be more than 2.5 mm.
- QRS complex: This segment includes Q, R and S waves and represents depolarization of ventricles where Q is the downward deflection after P wave, R is upward deflection after Q-wave and finally S is another downwards deflection. It takes between 0.06 to 0.10 second and the voltage varies between 1.5 and 2 mV.

- T-wave: Occurs after QRS complex and represents repolarization of ventricles. The height of the T wave can vary but it should not exceed the height of the R wave.

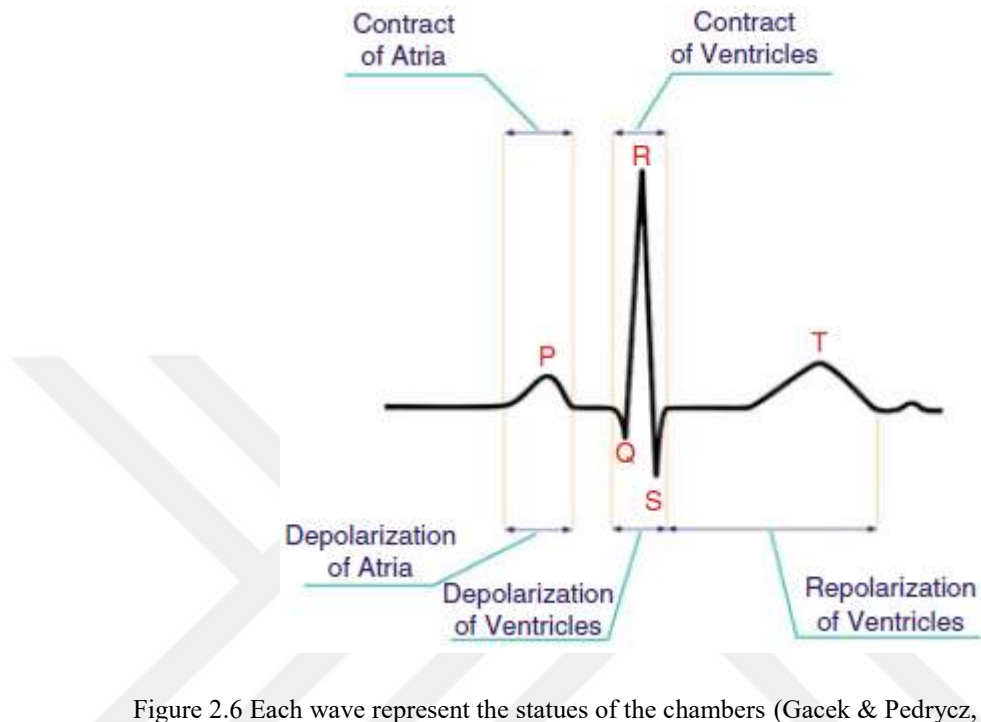


Figure 2.6 Each wave represent the statues of the chambers (Gacek & Pedrycz, 2011)

- ST segment: Delay between depolarization and repolarization of ventricle is ST segment and starts from the end of S wave to start point of T wave. ST segment must be flat. In case of any changes, it can be indication of a condition such as Myocardial ischemia.
- QT interval: Total time from when ventricular depolarization to the ventricular repolarization. This item depends on gender and cardiac rate, but generally is measured from 0.35 to 0.44 seconds.
- RR interval: The time between begining of a QRS complex to beginning of next QRS. Using this span, heart beat can be calculated.

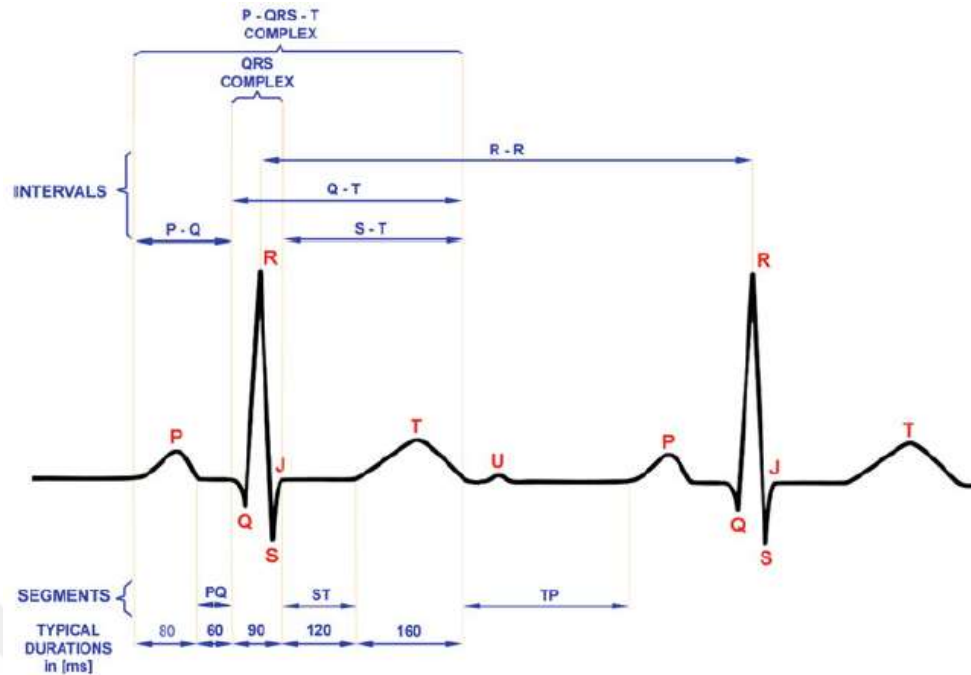


Figure 2.7 Segment representation of the ECG signal (Gacek & Pedrycz, 2011)

2.7 Arrhythmias

The main purpose of ECG monitoring is to identify any changes in the rate and shape of heart beats and figure out which factors cause them. Any abnormal changes in heart's function are defined as arrhythmia. The word arrhythmia is derived from Greek to mean loss or absence of rhythm. Although most arrhythmias are harmless, some arrhythmias can cause serious problems and may require immediate therapy. These abnormal changes occur as a result of a dysfunction in the normal conduction pathways of the heart. Two fundamental rhythm disturbances are abnormalities in impulse generation system and abnormal conduction of these impulses (Gacek & Pedrycz, 2011). Arrhythmias can be classified from different perspectives, such as mechanism, heart rate and duration. However, grouping arrhythmias with respect to their origin will be appropriate. Therefore, they are categorized into Ventricular Based Arrhythmias and Supraventricular Based Arrhythmias. Before discussing about these topics, it is helpful to know effects of speed of heart rate on classification of arrhythmias.

Bradycardia is the slower heart rate than normal (60 bpm). It is seen mostly in elderly people and is happened when SA node doesn't produce the signal or the produced signal is not received by ventricles.

Tachycardia is opposite of Bradycardia; i.e. much faster (> 100 bpm) heartbeats than normal. It can occur in different areas of the heart.

2.7.1 Ventricular Arrhythmias

Ventricular arrhythmias are a group of arrhythmias which begin in ventricle underlying cardiac conditions and may result in cardiac death.

2.7.1.1 Ventricle Tachycardia

Ventricle Tachycardia (VT) is the fast heart rate. Heart rate is over 100 bpm. In short period, it may not result in serious problems, however, longer periods are dangerous. VT contributes to sudden stop of blood flow and may turn into ventricle fibrillation. There are 3 or more consecutive beats originating from the ventricles. This rhythm makes severe shortness of breath.

2.7.1.2 Ventricular Flutter

This is a special kind of VT where heartbeat is between 250 to 350 bpm and because there is not enough time to fill and pump blood to body, it can be dangerous. It can progress to ventricle fibrillation.

2.7.1.3 Premature Ventricular Contractions

Premature Ventricular Contractions (PVC) are the earlier contractions of ventricles and are not serious but in some cases they may need medication.

2.7.1.4 Ventricular Fibrillation

Ventricular Fibrillation is the uncontrolled and irregular beat and is one of the most serious arrhythmias. In this arrhythmia, many impulses are produced that make heart beat faster. This results in chaotic heartbeat which means heart beats more than normal: consequently, very little amount of blood goes to brain and body which makes this arrhythmia to be dangerous and fatal.

2.7.2 Supraventricular Arrhythmias

These types of arrhythmias are tachycardia that start in the atria or atrioventricular (AV) node and are not as serious as ventricular arrhythmias. Here some of important arrhythmias are discussed.

2.7.2.1 Atrial Fibrillation

Atrial fibrillation is the most common arrhythmia. It involves a very fast and irregular contraction of the atria. Due to its importance, and also being the main topic of this thesis, detailed information will be given about it in the next section.

2.7.2.2 Atrial Flutter

In case of complication, this arrhythmia is similar to AF (Atrial Fibrillation). During atrial flutter, the heart's electrical signals spread through the atria in a fast and regular way. In this arrhythmia, heart rate is about 250 to 350 bpm. In atrial flutter, AV node plays a crucial role and acts like a valve, preventing the signal going toward ventricle and makes ventricles not beat as fast as the atria. This results in the ventricles to beat inefficiently as well (Dayan, 2006).

2.7.2.3 Supraventricular Tachycardia (SVT)

This arrhythmia is also called Paroxysmal Supraventricular Tachycardia (PSVT) and occurs when the signal which travelling from atria to ventricle, comes back and produces an extra beat. The rate is usually between about 100 - 250 bpm. It is usually seen in women, anxious young individual and tired people more than others. There is also a special type of SVT where the signal goes to the ventricle through an extra muscle pathway known as the bundle of Kent. and makes ventricle to beat earlier and very fast. This is called Wolff-Parkinson-White syndrome and can be dangerous.

2.7.2.4 Premature Supraventricular Contractions

This arrhythmia is also known as "Premature Atrial Contractions" (PACs). A premature discharge of an electrical impulse in the atrium causes a premature contraction. Thus, this results in an earlier beat than normal. It is similar to premature ventricular contractions (PVC).

2.7.2.5 Sinus Tachycardia

Sinus tachycardia occurs when your heartbeat rate is over 100 bpm. Most often sinus tachycardia happens when the body's demand for oxygen is increased, such as during exercise, stress, infection, blood loss. In this case, it is called normal sinus tachycardia. However, if this high heart rate is not related to exercise, the initiation of impulses from SA node is faster than normal.

2.8 Atrial Fibrillation

Atrial fibrillation (AF) is the most common arrhythmia type. This arrhythmia is associated with an increased risk of blood clots, stroke, heart failure and sudden death. Furthermore, number of patients who suffer from this arrhythmia have increased significantly during recent decades. In 2010, it was estimated that there were 33.5 million people around the world with AF and it is predicted that by 2030, 14-17 million people in Europe will suffer. Owing to considerably growing population of

AF patients, determination, classification and treatment of this arrhythmia are more essential (Kirchhof, Benussi, Kotecha, Ahlsson, Atar, Casadei, et al, 2016).

From the clinical point of view, atrial fibrillation is a kind of supraventricular based arrhythmia. In this arrhythmia, electrical signals are started in an irregular form in different regions in atria. So, atria beat in an irregular and often fast form instead of beating effectively (pushing whole blood content to ventricle). These signals are significantly prevented by AV node. Therefore, the AV node cannot send the signals to the ventricles as fast as they arrive. The rate of the atrial impulses is 300-650 bpm. However, ventricles beat around 100-170 bpm (see Figure 2.9).

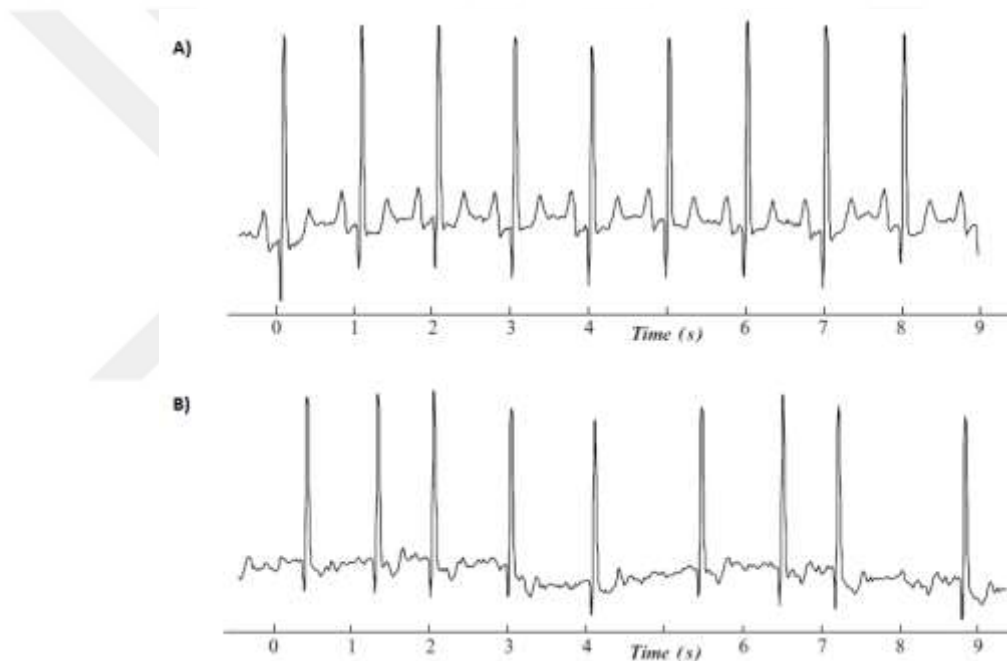


Figure 2.8 Typical ECG signals. A) Normal ECG signal; B) AF signal (Gacek & Pedrycz, 2012)

This weak and irregular contraction of atria leads to the case where the atria could not push completely all its blood to ventricles. Because of this, the chance of developing blood clots in the atria is increased. If the clot reaches the brain, the blood flow to some parts of the brain can be interrupted resulting in life-threatening strokes. People who have AF are five times more likely to have a stroke.

Normally, AF starts with short episodes and improves to longer and developed forms over a time. Hence, determination of AF and finding proper classification is essential in treatment because every class need a unique therapy. According to he

European Society of Cardiology (ESC), there are 3 different classes based on duration, termination and presentation (Kirchhof, Benussi, Kotecha, Ahlsson, Atar, Casadei, et al, 2016):

- Paroxysmal AF: In this case, AF episodes are automatically terminated within 48 hours, however may continue up to 7 days.
- Persistent AF: AF which lasts more than 7 days and terminated by drugs or current cardioversion.
- Permanent AF: If persistent AF occurs more frequently, it may develop into permanent AF and at this time normal heart rhythm cannot be restored with treatment.

In respect to these classes, some treatment procedures have been considered:

- Rhythm control: Aims to terminate the underlying rhythm. Sometimes known as medical cardio version.
- Rate control: Its purposes are to reduce the heart rate.
- Anticoagulants: These drugs are used to prevent to creation of blood clots which cause the stroke.

Generally, in patients with PAF, initially rhythm control therapy is applied, however, in patients with Permanent AF rate control is used. Persistent AF can be treated with either method. In addition, there is another treatment procedure which is known as catheter ablation. This technique consists of scattering the region in atria which is causing the arrhythmic problem (Davies & Scott, 2015).

The determination of PAF condition is very easy by recording the ECG of the subjects during an episode. However, since PAF episodes mostly last for a short time period, it is very difficult to record the ECG of the subjects during the arrhythmic event. Therefore, a system that would be able to detect PAF patients based on ECG records taken during non-arrhythmic time periods would be very beneficial.

CHAPTER THREE

HEART RATE VARIABILITY

Heart Rate Variability (HRV) is one of the most frequently used tools to analyze the ECG and defined as the amount of variation in the length of beat to beat intervals (Xhyheri et al., 2012). HRV is measured by milliseconds and allows mathematical expressions to be easily used for its calculation and interpretation (Samoilov, Teli-shev, Pyanov, 2018). The variation of the length of heartbeat intervals reflects changes in the relative balance between the sympathetic and parasympathetic nervous systems. As discussed in the previous chapter, the sympathetic and parasympathetic nervous systems control organs with opposite effects. Thus, HRV can represent the evaluation of the function of the autonomic nervous system (Long, Fonseca, Haakma, Aarts, Foussier, 2012). HRV ratio could be varied in some cases, such as rest, exercise and disease. Information in HRV is relevant to many cardiovascular and non-cardiovascular diseases such as myocardial infarction, diabetic neuropathy and high blood pressure, (Seyd, Ahamed, Jacob, Joseph, 2008). Since R wave is easy to detect (owing to its considerable amplitude), it is generally examined as the reference point in HRV analysis. Figure 3.1 illustrates RR interval constitution from an ECG signal by calculating the time intervals between successive R points.

The clinical relevance of HRV was first appreciated in 1965 when Hon and Lee noted that fetal distress was preceded by alterations in inter-beat intervals before any appreciable change occurred in heart rate itself. The clinical importance of HRV was noticed in the late 1980s, when it was confirmed that HRV was a strong and independent predictor of mortality after an acute myocardial infarction.

In general, number of heart beats per minute is about 60 to 80 beats and it means the mean value for RR interval is about 600 to 750 ms. Usually the body responds to stresses caused by sources like exercise, psychological events and other internal or external stressors by decreasing or increasing the heart rate. This healthy adaptive behavior leads to changes in heart rate known as HRV. Low HRV has been proposed

as a marker of a number of pathophysiologic conditions, including increased risk of mortality (Gacek & Pedrycz, 2011).

There are many methods for evaluating and describing HRV. Time domain and frequency domain analyses are the most popular methods (Task Force of the European Society of Cardiology, 1996).

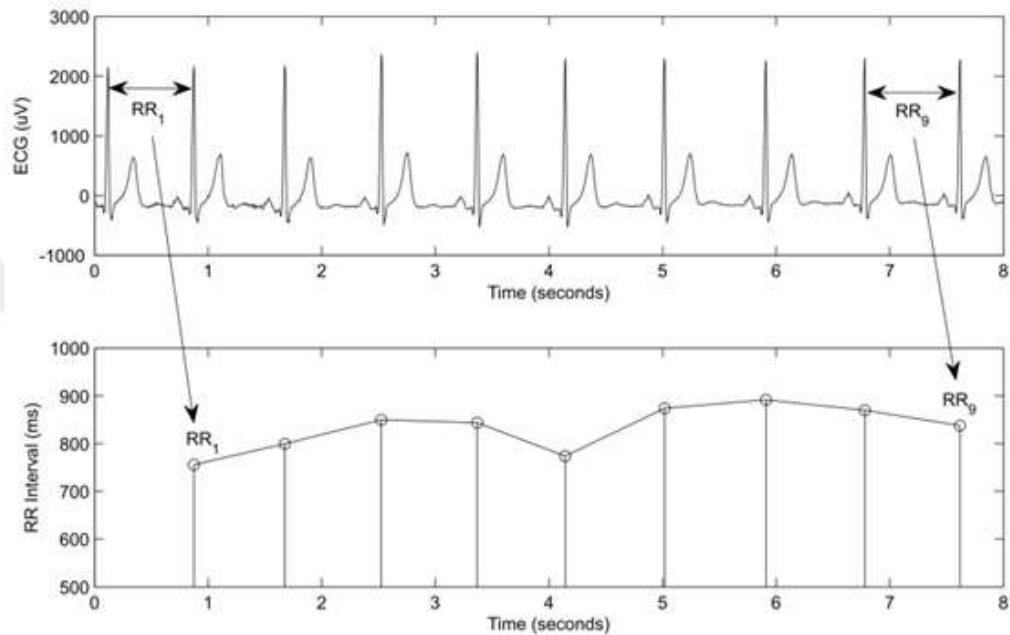


Figure 3.1 Deriving RR interval from ECG

3.1 Time Domain HRV Indices

One of the simplest methods to evaluate the variation of the heart rate is time domain measurements. There are some statistical methods which are used to measure the HRV in time domain (Table 3.1). Standard deviation of the NN intervals (or SDNN) is one of the simplest methods and measured by applying square root of variance (Equation 3.1). Often calculated over a 24-hour period, it describes short term high frequency variations and low frequency components. As the period of monitoring decreases, SDNN estimates shorter and shorter cycle lengths. It should also be noted that the total variance of HRV increases with the length of analyzed recording. Therefore, it depends on the length of recording period (Task Force of the European Society of Cardiology, 1996).

$$SDNN = \sqrt{\frac{1}{n} \sum_{i=1}^n (NN_i - m)^2} \quad (3.1)$$

where NN_i is the duration of the i^{th} NN interval, n total number of NN intervals and m is their mean duration. The bigger the SDNN is, the bigger the HRV is. On the other hand, when the SDNN becomes bigger, it means that the person's emotion variability becomes bigger during measuring his ECG signals.

Square root of the mean squared difference of successive NN intervals (or RMSSD) is based on interval differences and is used only in short term monitoring.

$$RMSSD = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n-1} (NN_{(i+1)} - NN_i)^2} \quad (3.2)$$

The RMSSD represents the difference between two successive R waves. So, when RMSSD is small, it means heart beat becomes quick.

NN50 is the number of pairs of adjacent NN intervals differing by more than 50 ms in the entire recording. pNN50 is the proportion derived by dividing NN50 by the total number of NN intervals.

Table 3.1 Time domain HRV measurements

<i>Method</i>	Unit	Description
<i>SDNN</i>	ms	Standard deviation of all RR intervals
<i>RMSSD</i>	ms	Square root of the mean squared difference of successive NN intervals
<i>NN50</i>	count	Number of pairs of adjacent RR intervals differing by more than 50 ms
<i>pNN50</i>	%	NN50 count divided by the total number of all RR intervals

3.2 Frequency Domain HRV Indices

If we calculate the power spectrum of a typical RR interval over at least 24 hours, 4 frequency regions can be observed (see Table 3.2). But during short term RR intervals, generally low frequency, high frequency and very low frequency regions are considered. Usually, normal activity of SA node affects the high frequency (HF); this means that para-sympathetically mediated activity like respiration can be seen in this frequency region. On the other hand, sympathetically mediated activity like blood pressure regulation is observed in low frequency (LF) region (Clifford, Azuaje, McSharry, 2006).

Table 3.2 Frequency bands of HRV

Band	Frequency Range (Hz)
Ultra low frequency(ULF)	$0.0001 \geq \text{ULF} < 0.003$
Very low frequency (VLF)	$0.003 \geq \text{VLF} < 0.04$
Low frequency (LF)	$0.04 \geq \text{LF} < 0.15$
High frequency (HF)	$0.15 \geq \text{HF} < 0.4$

The distribution of power and central frequencies in HF and LF are not fixed and may change in relation with variations in autonomic modulations of heart period. VLF, LF, and HF power components are usually measured in absolute values of power (milliseconds squared). Also, HF and LF can be calculated in normalized units:

$$LF(u, v) = \frac{LF}{TP - VLF} \quad (3.3)$$

$$HF(u, v) = \frac{HF}{TP - VLF} \quad (3.4)$$

where the total power (TP) is defined as the power in the frequency band going from 0 Hz to 1 Hz. The ratio of LF to HF also indicates the balance of ANS. In healthy adults, the LF/HF ratio is typically between 1.5 and 4.5. With respect to these param-

eters, non-parametric method based upon the FFT algorithm should include the formula of discrete event series interpolation, the frequency of sampling the DES interpolation, the number of samples used for the spectrum calculation, and the spectral window employed.

3.3 Nonlinear methods

Nonlinear methods are used for characterizing the variability of heart rate to measure the regularity and complexity of the fluctuations. Hilavin (Hilavin, 2016) applied 3 nonlinear methods:

Poincare Plot: Poincare Plot is the graphical representation of correlation between successive RR intervals by plotting $RR_{i+\tau}$ over RR_i . The shape of the distribution of the data points is the essential feature. In order to quantify the geometry, an ellipse, which is oriented along the line of identity ($RR_j=RR_{j+1}$), is fitted to the plot. Then, the dispersion of the data points perpendicular to the line of identity is called SD1 and the dispersion of the data points along the line of identity is called SD2. SD1 and SD2 describe the short-term and long-term variability, respectively.

Sample Entropy: The sample entropy (SampEn) is one of methods used for quantifying signal complexity, and is based upon hypothesis that decreasing entropy is a pointer towards loss of heart rate variability (HRV). SampEn is independent of number of data points which may be used for considerably shorter time series.

Detrended Fluctuation Analysis (DFA): Detrended Fluctuation Analysis is useful tool in revealing the extent of long-range correlations in time series. The data's series have been integrated and divided into a series of regular intervals. For each interval, it has been calculated the "local" fluctuation as the difference compared to a straight line of a linear interpolation. Indeed, the "global" fluctuation has been calculated as the square root of the average of the local's fluctuations.

The list of all HRV features obtained from normal sinus rhythm (NSR) ECG records used in this study is given in Table 3.3. All data was provided by Hilavin

(Hilavin, 2016) by employing time domain, frequency domain and nonlinear HRV analysis. Hilavin extracted 32 HRV features obtained from normal sinus rhythm (NSR) ECG records to obtain a full feature set. However, then, by using genetic algorithm, only 8 most important HRV features were selected. We use these 8 features in this thesis study.

Table 3.3 HRV features used in this study (Hilavin, 2016)

Time domain	Mean RR	Mean of RR intervals	$NN = \frac{1}{N} \sum_{i=1}^N NN_i$
	Std RR	Standard deviation of RR intervals	$SDNN = \sqrt{\frac{1}{n} \sum_{i=1}^n (NN_i - m)^2}$
Frequency domain	HF peak	HF band peak frequency	
	LF power prc	Relative power of LF band (LF power/Total power)	$LF \text{ power prc} = \frac{LF \text{ power}}{\text{Total power} - VLF \text{ power}} \times 100$
	HF power prc	Relative power of HF band (HF power/Total power)	$HF \text{ power prc} = \frac{HF \text{ power}}{\text{Total power} - VLF \text{ power}} \times 100$
Nonlinear	SD1	Dispersion of points perpendicular to line of identity in Poincare plot	$SD1^2 = \frac{1}{2} SDSD^2$ $SD2^2 = 2SDRR^2 - \frac{1}{2} SDSD^2$
	SampEn	Sample entropy	$SampEn(m, r, N) = -\ln\left[\frac{Q'^m(r)}{Q'^{m+1}(r)}\right]$
	DFA Alpha1	Short term fluctuation slope of detrended fluctuation analysis	$y(k) = \sum_{i=1}^k [RR_i - RR_{av}]$ $F(n) = \sqrt{\frac{1}{N} \sum_{k=1}^N (y(k) - y_n(k))^2}$

CHAPTER FOUR

CLUSTERING

Machine learning is a subfield of computer science which aims to build algorithms that receive input data and use statistical analysis to predict an output value within an acceptable range (Gan, Ma, Wu, 2007). Machine learning algorithms are divided into 2 types: Supervised and Unsupervised. In supervised learning, we create algorithms to match the input data to output values in order to get the desired output values. But unsupervised algorithms do not need to be trained with desired outcome data. Instead, they use an iterative approach to investigate the data statistically and arrive at conclusions. The classification of the electrocardiogram into different categories of beat types and rhythms, representing one or more underlying pathologies, is essentially a pattern recognition task. The main purpose of this thesis is to analyze the structure of data obtained from PAF and non-PAF subjects by utilizing statistical approaches in unsupervised classification task, where various algorithms with different characteristics are employed and also combined to get more precise results.

4.1 Data Clustering

Classification and clustering are the most widely used techniques for exploratory data analysis. In classification, data points are labeled beforehand. For example, in our dataset (cardiovascular diseases) each sample is known as healthy or unhealthy person. A classifier is then trained with the labeled samples to create a set of rules or a mathematical model that can then be used to look at the features captured from a new person and label the case as healthy or a patient with a particular disease. Then, new samples without any previous information are introduced to classifier to figure out their classes.

The data clustering, also called unsupervised classification or cluster analysis, is an unsupervised method in which there is no label for model training. The exact numbers of clusters are also unknown beforehand. For a given dataset, clustering

methods are expected to divide data points to groups which maximize the intra-subset similarity and inter-subset dissimilarity (Wong, 2015).

Cluster analysis is a major tool in a number of applications in many fields of pattern recognition, business and science. As illustrated in Figure 4.1, data clustering algorithms are mostly divided into two types, Crisp or Fuzzy clustering and Hard clustering (Gan, Ma, Wu, 2007). In Fuzzy clustering, each data point may be assigned to more than one cluster but in hard clustering each data point can only be assigned to one cluster. There are many clustering algorithms with different statistical approaches (Halkidi, Batistakis, Vazirgiannis, 2001). Here, some approaches are discussed.

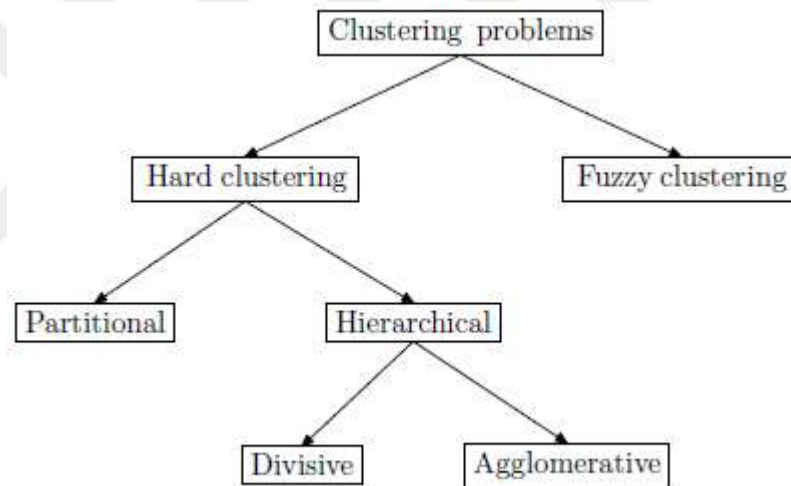


Figure 4.1 Classification of clustering algorithms (Gan, Ma, Wu, 2007)

Hierarchical clustering: There are two major approaches under this category: *agglomerative* and *divisive* methods. In this method, large clusters are split into small ones (top-down or *divisive*) or small clusters are combined to get large clusters (bottom-up or *agglomerative*). In the end, tree of clusters, called dendrogram, shows in which way to illustrate how the points or clusters are related. By cutting the dendrogram at a desired threshold, a clustering of the data items into disjoint groups is obtained (Figure 4.2).

Partitional clustering: this type of clustering simply decomposes a set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one cluster. So, this type of clustering algorithm is useful for large or multi-dimensional datasets and usually has its own objective functions, which define how good a clustering solution is. K-means, one of the most well-known algorithms, has this approach where the objective is to minimize the cost function.

Density-based clustering: Density based clustering algorithms are utilized to discover clusters of arbitrary shape. Unlike partitional clustering, in this technique number of clusters is not required. The clusters are created based on some density and connectivity functions. DBSCAN is one of the most popular and well known algorithms (Ester, Kriegel, Sander, Xu, 1996). There are two parameters in this algorithm:

- radius of the neighborhoods
- minimum number of data points required to form a dense region

DBSCAN uses these parameters to categorize the data points. It starts with a not visited data point and find data points within the ϵ distance (neighborhoods). If there are enough neighborhoods around this point, then clustering process starts. This process continues until all points are marked as visited. In case that the user couldn't detect exact pattern of the data set, choosing a meaningful distance threshold ϵ can be difficult.

Grid-based Clustering: Like density-based clustering, grid-based clustering also is used in large multidimensional spaces. In general, a typical grid-based clustering algorithm consists of the following five basic steps (Grabusts and Borisov, 2002):

1. Partitioning the data space into a finite number of cells
2. Calculating the density for each cell
3. Sorting according to their densities
4. Identifying cluster centers

5. Traversal of neighbor cells

Fuzzy Clustering: This type of the clustering algorithms uses the fuzzy technique to cluster the data and result in crisp clusters, meaning that a data point either belongs to a cluster or not. The most important fuzzy clustering algorithm is Fuzzy C-Means.

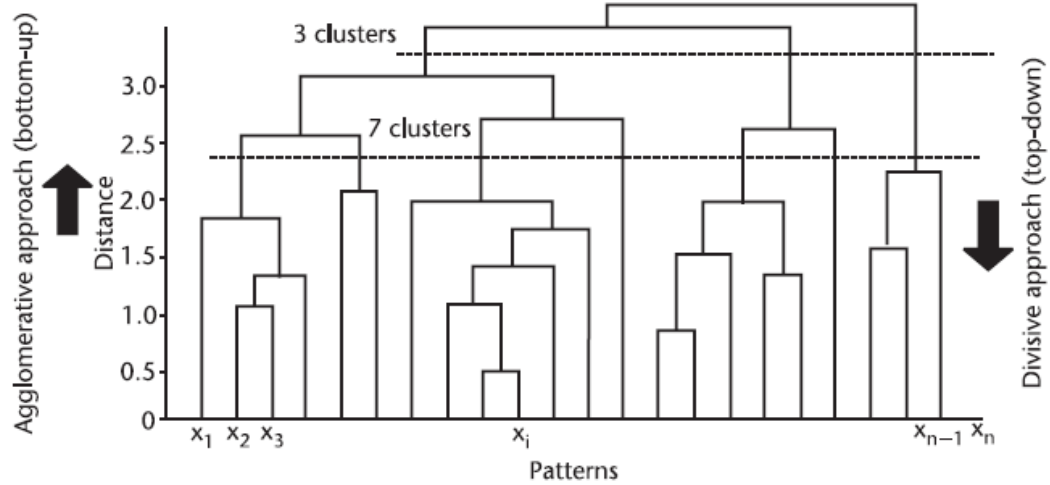


Figure 4.2 Dendrogram of dataset with Divisive and Agglomerative methods

4.2 Clustering Algorithms

The main goal of data cluster analysis is to discover the natural clusters(s) of a set of points or objects. As we discussed before, K-means and Fuzzy C-means are the most popular and well-known algorithms in case of unsupervised classification. Before implementing these algorithms, it is necessary to define some aspects such as distance function, number of cluster, validity index and etc. (Figure 4.3). Each of these parameters can affect the result of clustering (Kaur, 2014). In the next section, basic information about clustering are discussed which includes the initial requirements about clustering. Then, two algorithms will be illustrated in detail.

4.2.1 Notation and Terminology

Before introducing clustering algorithms, it would be necessary to provide basic requirements about methods and parameters.

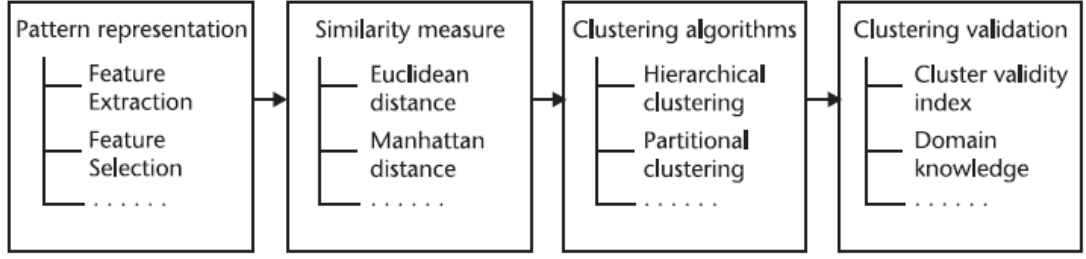


Figure 4.3 Overall perspective of clustering procedure (Gari, Clifford, Azuaje, McSharry, 2006)

4.2.1.1 Data

In this study, we use RR interval series obtained from MIT-BIH PAF Prediction Challenge Database (Moody et al., 2001). The dataset was originally used in Computers in Cardiology Challenge 2001 competition. RR interval series in this database have been extracted from 30 minute ECG signals. The 30 minute ECG signals are extracted from long term ECG records. The database has 100 ECG record sets obtained from 98 different subjects and contains 798 observations and 8 attributes. In case of introducing algorithms, we consider the dataset as follows: $X = \{x_1, x_2, x_3, \dots, x_n\}$ vector in n -dimensional feature space.

4.2.1.2 Similarity Measure

Similarity coefficient represents the relation of the two data points. The more the two data points resemble one another, the larger the similarity is. Different measures of distance or similarity can be used for different datasets. So the result of the algorithms is affected by choosing the appropriate distance measure. Depending on the type of the dataset, different similarity or dissimilarity measures can be used. One of the standard measures for distance function in numeric datasets is Euclidean distance. Euclidean distance computes the root of square difference between co-ordinates of pair of objects. For two vectors, $P = [p_1, p_2, p_3, \dots, p_n]$ and $Q = [q_1, q_2, q_3, \dots, q_n]$ in n -dimensional data space Euclidean distance is calculated as:

$$d(P, Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2 + \dots + (p_n - q_n)^2} \quad (4.1)$$

The standard Euclidean distance can be squared in order to place progressively greater weights on objects that are farther apart. In this case, the equation becomes:

$$d^2(P, Q) = (p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2 + \dots + (p_n - q_n)^2 \quad (4.2)$$

Manhattan distance is another most common used distance function to compute the differences between points in a city road grid. It determines absolute differences between coordinates of pair of objects

$$d(p, q) = \sum_{i=1}^n |p_i - q_i| \quad (4.3)$$

4.2.1.3 Cluster Center Initialization

One of the metrics which affects the quality of algorithms is the first location of cluster centers. Usually cluster centers are assigned randomly. However, in order to minimize the calculation time, some methods have been proposed along with the K-means or FCM algorithms.

A method was proposed by finding the distance of the point from mean of the dataset and then sorting and dividing them into k partitions. Then, cluster centers were obtained by calculating the mean of these partitions (Arnaldo & Bedregal, 2013).

Another study determined to find cluster centers by first finding the distance of the all the pairs, then minimum two points were selected. Distance of these points and other points in data set are calculated. These measurements have brought into new matrices which contain $\alpha \times (X_n/k)$ data points (where k is the number of clusters and α is 0.75). Mean of these matrices were assigned as cluster centers (Yuan, Meng, Zhangz, Dong, 2004).

Subtracting clustering can be also used to find new cluster centers (Yang, Zhang, Tian, 2010). In subtracting clustering, every point is assumed as cluster center. Then, density measure for every point is calculated. At the end, the points with high density are chosen as initial cluster centers.

An alternative method was introduced to find cluster centers for K-means by assuming k number of points randomly, let them to be shown as m and n . The distance between them is calculated as $(C_{dist} = EuDist(m, n))$. Then, another arbitrary point in dataset is chosen and the distance of this point to m is calculated. If this distance is bigger than C_{dist} , the new point is selected as new cluster center and the mean of the m and n is found as another cluster center point. This procedure continues until all points are proposed (Weizhi Huang, 2014). This algorithm works well in K-means. However, owing to the limitation of FCM in selecting data points as cluster centers, in final step the mean between distance of final cluster centers obtained by this method and variance of the dataset is chosen as new cluster centers for FCM.

4.2.1.4 Number of Clusters

One of the most discussed topics in clustering is determining the optimal number of clusters which fits the data. Although different methods have been proposed to solve this problem, the exact and precise clusters would be known by user. In clustering task, the validity indices are generally used to represent how good is our clustering method to proposed K number of clusters.

One of the metrics which is commonly used in order to compare results across different values of K is the within cluster sum of squared errors (WCSS). Since increasing the number of clusters will always reduce the distance to data points, this method gives an easy and quick answer.

$$WCSS = \sum_{i=1}^K \text{distance}(x, c)^2 \quad (4.4)$$

If one sees the elbow in the graphic, appropriate number of cluster could be found (Figure 4.4). But it always doesn't work properly and sometimes elbow does not appear or could be found more than once. In this case, the optimal number of cluster may be more than we calculated or the dataset has a particular pattern.

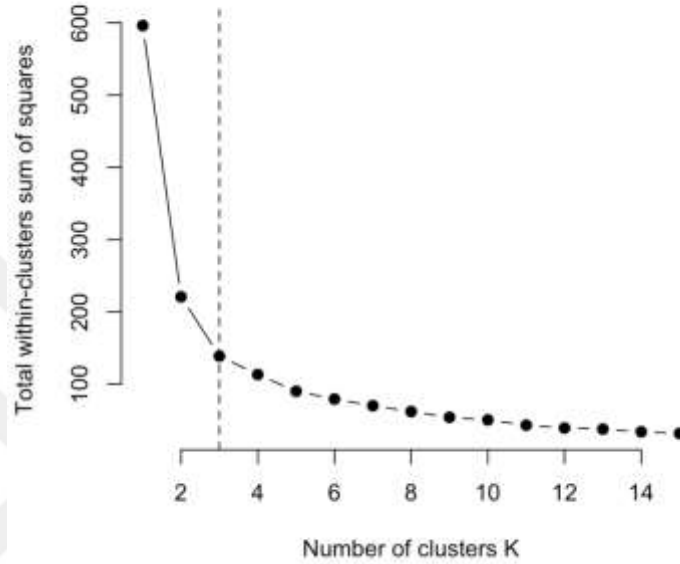


Figure 4.4 Elbow represents the optimal number of cluster

4.2.2 K-Means Clustering

One of the most famous and easiest methods in partitional clustering is K-Means which was proposed by J. B. MacQueen (1967). The K-Means algorithm partitions a collection of data points $X = \{x_1, x_2, x_3, \dots, x_n\}$ into k groups and find cluster center for each group in case that cost function or objective function is minimized (Equation 4.5). Between point x_j , $j=1 \dots, n$, and cluster center various methods such as Euclidean distance, Manhattan distance or other metrics can be used as similarity measure. The goal of K-means is to minimize the sum of the squared error over all K clusters:

$$J(k, X, C) = \sum_{i=1}^k \sum_{x_k \in G_i} U_{ij} d(x_k, C_i) \quad (4.5)$$

where C_i is cluster center of i^{th} cluster, k is number of clusters and $d(x_k, C_i)$ is Euclidean distance or other distance metric. Also, there is membership matrix which can be defined as

$$U_{ij} = \begin{cases} 1 & \text{if } d(X_j - C_i) \leq d(X_j - C_k), \text{ for } k \neq i, \\ 0 & \text{otherwise} \end{cases} \quad (4.6)$$

According to Equation 4.7, each point in dataset can be assigned to one group. Membership matrix has the following properties:

$$\sum_{i=1}^k U_{ij} = 1, \forall j = 1, \dots, n \quad (4.7)$$

$$\sum_{i=1}^k \sum_{j=1}^n U_{ij} = n \quad (4.8)$$

Therefore, it would be easy to figure out which point belongs to which group. After initializing membership matrix, new cluster center which minimizes objective function is defined:

$$C_i = \frac{1}{|G_i|} \sum_{k, X_k \in G_i} X_k \quad (4.9)$$

where $|G_i|$ is the size of G_i .

The K-means algorithm can be seen as two phases: the initialization phase, where the algorithm randomly assigns the subjects into k clusters, and the iteration phase where the algorithm computes the distance between each subject and each cluster and assigns the subject to the nearest cluster. The algorithm is inherently iterative (Table 4.1) (Gan, Ma, Wu, 2007).

Advantages of K-means include computational efficiency, fast implementation and easy mathematical background. However, K-means also has limitations such as random choice of centroid locations at the beginning of the procedure and an unknown number of clusters k .

Table 4.1 K-means algorithm

1) Determine number of clusters (K) and similarity measure
2) Define cluster centers randomly
3) Repeat;
Find membership matrix, distance matrix and new cluster centers
4) Until;
Centroids do not change, or objective function is not minimized enough.

In recent decade a lot of optimization of the K-means algorithm have been proposed. X-means is the modified K-means algorithm which uses BIC criterion to find optimal number of clusters (Pelleg & Moore, 2000). Given a range for k , $[k_{\min}, k_{\max}]$, the X-means algorithm starts with $k = k_{\min}$ and continues to add centroids when they are needed until the upper bound is reached.

K-means++ is another algorithm which optimizes the clustering by choosing appropriate cluster centers before proceeding with the standard K-means optimization iterations (Arthur & Vassilvitskii, 2007). This can boost the algorithm. The exact algorithm is as follows:

1. Choose one center randomly among the data points.
2. For each data point x , compute the distance between x and the nearest center that has already been chosen.
3. Choose one new data point at random as a new center, using a weighted probability distribution where a point x is chosen with probability proportional to $D(x)^2$.
4. Repeat Steps 2 and 3 until k centers have been chosen.

Then, these achieved cluster centers are used for standard K-means algorithm.

4.2.3 Fuzzy C-means Clustering

Fuzzy C-means (FCM), improved by James C. Bezdek in 1981, is one of most well-known and powerful fuzzy based clustering methods frequently used in pattern recognition and data mining. This method can be seen as modified K-means algorithm and sometimes called Fuzzy Hard means. FCM is a data clustering algorithm in which each data point belongs to a cluster with largest membership value. Using this method, the data can be seen in groups which facilitate to make good understanding of data.

Let us consider M-dimensional data points represented by $X_M = \{x_1, x_2, x_3, \dots, x_n\}$, FCM partitions the data into K fuzzy groups and find corresponding cluster centers. The objective function for this method is:

$$J(U, C) = \sum_{i=1}^C \sum_{j=1}^n U_{ij}^m d_{ij}^2 \quad (4.10)$$

Where C is the number of clusters, m is a weighting exponent, $d_{ij} = |x_j - c_i|$ is the Euclidean distance between cluster centers and their corresponding data points and U_{ij} is membership matrix in which its elements donate the membership of j^{th} sample. Membership values are within interval $[0, 1]$ and, like K-means, the membership matrix must satisfy the condition stated in the previous section (Equations 4.8 and 4.9).

Fuzzy C-means algorithm is as follows (see Table 4.2):

- 1) Membership matrix is created. Values between 0 and 1 are randomly assigned to every data point.
- 2) Cluster centers are determined based on membership values

$$C_i = \frac{\sum_{j=1}^n U_{ij}^m x_j}{\sum_{j=1}^n U_{ij}^m} \quad (4.11)$$

- 3) New membership values are found based on obtained cluster centers:

$$U_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ij}}{d_{kj}} \right)^{\frac{2}{m-1}}} \quad (4.12)$$

- 4) Cost function is calculated based on Equation 4.11. If the improvements are less than a certain threshold, then stop; otherwise continue.
- 5) Go to step 2.
- 6) Repeat until the satisfied conditions are reached.

Before using this algorithm, it would be necessary to consider some parameters. Exponent value is one of the issues in FCM which should be determined beforehand. If m is inappropriately selected, it affects convergence of FCM. When m is chosen 1, the algorithm approaches to fuzzy hard means and when m tends to infinity, the only solution of the FCM will be the mass center of data set. It is accepted that the range for m is between 1.5 and 2.5 and normally m is chosen 2.

Splitting dataset into different clusters in fuzzy clustering is not as obvious as in the case of K-means clustering where each object is assigned to exactly one cluster. In FCM clustering, each point belongs to each cluster, with the degree given by the membership value. Owing to this, a threshold is used to determine the assignment of an object. For a given group, if the membership value for i^{th} point is above 0.5, it would be assigned to that cluster. In addition, because different distance functions may affect the result of clustering, similarity metric must be specified beforehand (Kaur, 2014).

Table 4.2 Fuzzy C-means algorithm

- 1) Determine number of clusters (K) and initial membership value
- 2) Find cluster centers
- 3) Repeat;
 - Find new membership matrix and new cluster centers
- 4) Until;
 - Centroids do not change, or objective function is not minimized enough.

FCM and its variants realize the clustering task for a dataset by minimizing an objective function subject to certain concepts. A lot of modification and enhancement algorithms based on FCM have been proposed in recent years.

Using the concept of K-means++, Fuzzy C-means++ was introduced to utilize the seeding mechanism which improves the effectiveness and speed of Fuzzy C-means (Stetco, Zeng, Keane, 2015.). Proposed method has significant improvement in number of iteration, which means less time consuming.

An improved fuzzy partition for fuzzy regression models (IFP-FCM) is another algorithm which aims to modify the original FCM's objective function (Höppner and Klawonn, 2003). Equation 4.14 is the objective function of this model with a restriction rule in the second part of the equation:

$$J_{\text{IPF-FCM}} = \sum_{i=1}^c \sum_{j=1}^n u^2 d^2(x_j, c_i) + \sum_{j=1}^n \alpha_j \sum_{i=1}^c (U - \frac{1}{2})^2 \quad (4.13)$$

where $\alpha_j = \min\{d^2(x_j, c_i) - \eta\}$ is introduced to avoid the possible case of negative membership degree. There is a constraint in the second term on the right side of Equation 4.13 which makes the algorithm less sensitive to noise and outliers. On the other side, fuzzier exponent is a fixed value and is selected as 2. In this method, new cluster center and membership function are given in Equations 4.14 and 4.15 respectively.

$$C_i = \frac{\sum_{j=1}^n u_{ij}^2 x_j}{\sum_{j=1}^n u_{ij}^2} \quad (4.14)$$

$$U_{ij} = \frac{1}{\sum_{j=1}^c \left(\frac{d^2(x_j, c_i) - \alpha_j}{d^2(x_j, c_k) - \alpha_j} \right)} \quad (4.15)$$

Generalized Fuzzy C-Means Clustering Algorithm with Improved Fuzzy Partitions (GIFP-FCM) was proposed to eliminate some disadvantages of IFP-FCM (Zhu, Chung, Wang, 2009). These improvements let the algorithm to choose values larger

than 2 for fuzzier exponent while it is selected 2 in the IFP-FCM. Equation 4.16 represents the objective function of the method:

$$J_{GIFP-FCM} = \sum_{i=1}^c \sum_{j=1}^n U_{ij}^m d^2(x_j, c_i) + \sum_{j=1}^n \alpha_j \sum_{i=1}^c (1 - U_{ji}^{m-1}) \quad (4.16)$$

where $\alpha_j = \omega * \min\{d^2(x_j, c_i) - \eta\}$. Although ω is set as a values between 0 and 1, the best results are obtained when $\omega = [0.9, 0.99]$. So, with parameter α , GIFP-FCM rewards the biggest membership and suppress the others simultaneously. Another advantage of introducing the parameter α that it helps to build a general link between FCM and IFP-FCM. That is, when $\alpha = 0$ GIFP-FCM becomes FCM, when the fuzziness index $m = 2$ and α is approaching one, GIFP-FCM tends to become IFP-FCM with its parameter η approaching zero. Cluster center equation is the same as standard FCM (Equation 4.11) and new membership function is given below

$$U_{ij} = \frac{1}{\sum_{j=1}^c \left(\frac{d^2(x_j, c_i) - a \cdot \min\{d^2(x_j, c_s)\}}{d^2(x_j, c_k) - a \cdot \min\{d^2(x_j, c_s)\}} \right)^{\frac{1}{m-1}}} \quad (4.17)$$

Suppressed Fuzzy C-means (s-FCM) is another FCM based algorithm which aims to reduce time while preserving good accuracy and convergence speed (Fan, Zhen, Xie, 2003). This algorithm doesn't reduce the objective function. There is an additional step after membership function determination to suppress the values. In this method, new membership function is implemented as:

$$\mu_{ik} = \begin{cases} 1 - \alpha + \alpha U_{ik} & \text{if } i = \operatorname{argmax}(U_{ik}) \\ \alpha U_{ik} & \text{otherwise} \end{cases} \quad (4.18)$$

where $\mu_{ik} = (i=1 \dots c, k= 1 \dots n)$ indicates the new fuzzy memberships obtained after suppression. Equation 4.19 represents the fact that for a non-winner point, its value is decreased via multiplying by suppression coefficient α where $0 \leq \alpha \leq 1$. In

case $\alpha = 0$ and $\alpha = 1$ the algorithm behaves as K-means and FCM, respectively. One possible method to determine the proper value for α is (Li, Fan, 2014) given as

$$\alpha = \frac{1}{\log c} \left(-\frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n U_{ij} \log U_{ij} \right) \quad (4.19)$$

4.3 Ensemble Learning

In order to solve a particular pattern classification problem, the process of combining and generating a model based on several classifiers is widely used. This technique is known as Ensemble learning (Boongoen, Tossapon, Natthakan, 2018). Designing a combination of clustering algorithms is a challenging and difficult task because cluster labels are symbolic and so one must also solve a correspondence problem (Ghaemi, Sulaiman, Ibrahim, Mustapha, 2009). Generally, if several learning schemes are available, instead of choosing best-performing algorithm, it may be advantageous to use them all and combine the results. An obvious approach to making decisions more reliable is to combine the output of several different models. Several machine learning techniques do this by learning an ensemble of models and using them in combination (Gari, Clifford, Azuaje, McSharry, 2006). As illustrated in Figure 4.5, in ensemble learning, some strategies may be considered:

- a) Generated output from one method is the input to another method.
- b) Modifying the output of one method to produce the input to another method.
- c) Combining output of two independent algorithms.
- d) Using one methodology to adapt the learning process of another one.

It is important to emphasize that there is no guarantee that the combination of classifiers will always carry out better than the best individual classifier in the ensemble.

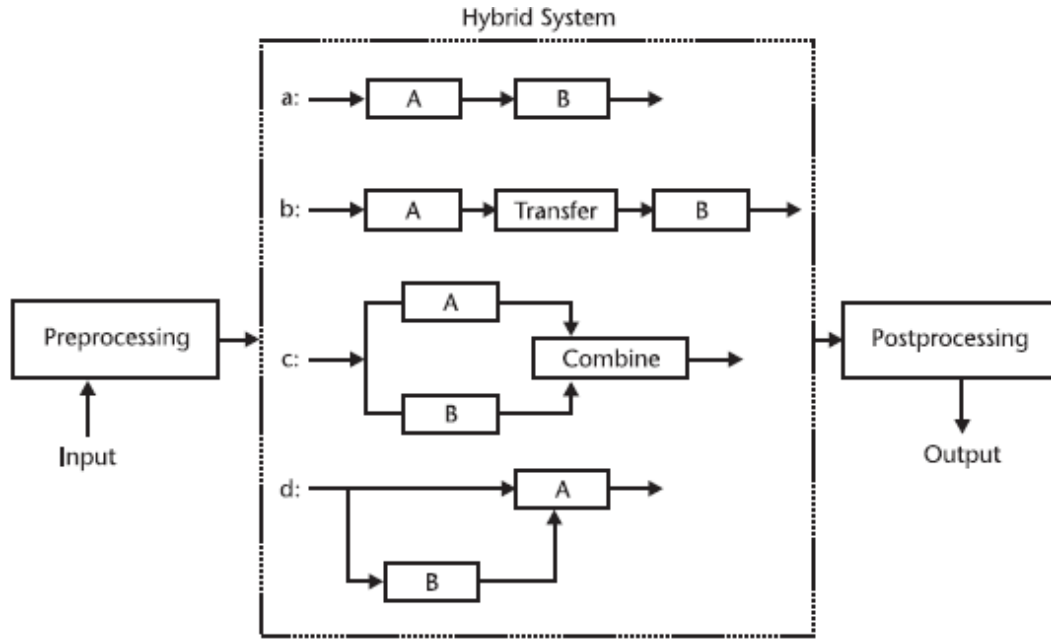


Figure 4.5 A fundamental representation of the combining clustering algorithms. A and B are different algorithms and a, b, c and d are different combinations we can form (Gari, Clifford, Azuaje, McShar-ry, 2006)

4.4 Model Assessment

In order to get better results and conduct precise analysis on dataset, it would be reasonable to split the whole dataset to train and test groups and implement clustering algorithms on these groups and then find total accuracy. This is done by cross-validation. In the following section, cross-validation technique and validity indices are discussed. In addition, accuracy measure for medical dataset are going to be explained.

4.4.1 Cross-Validation

Cross-validation is a well-established technique that can be used to obtain estimates of model parameters that are unknown (Duda, Hart, Stork, 2012). The performance of most classifiers is typically evaluated through cross-validation, which involves the determination of classification accuracy for multiple partitions of the input samples used in training. The general idea here is to divide the dataset to v groups (or folds). Then one fold is selected as test while remaining folds are used for training.

This process is repeated for all folds, hence every data point gets to be in a test set exactly once, and gets to be in a training set $v-1$ times (Figure 4.6). The average of performance values for each fold is used for the final evaluation. This process is time consuming and expensive in case that the dataset is so large.

4.4.2 Cluster Validation

The validation or assessment of the clustering results is a fundamental task in clustering analysis. This is because clustering a dataset is an unsupervised process and there are no predefined classes that can show that the clusters found by the clustering algorithms are valid. Thus, the main purpose of cluster validity methods is to find the partitioning which best fits the underlying data (Kovács, Legány, Babos, 2005).

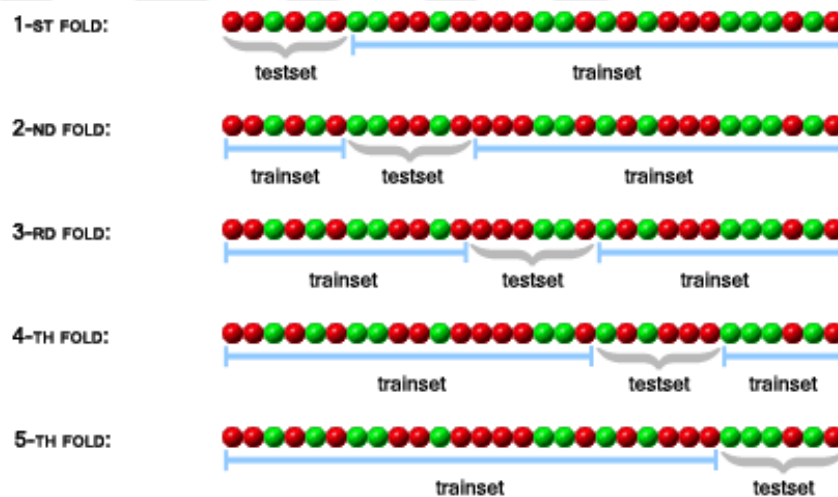


Figure 4.6 The scheme of dividing dataset to 5-fold using cross validation

As determining optimal number of clusters in K-means and FCM is one of most important issues, it is a highly nontrivial task to find the optimal number of clusters. To do this, we need some cluster validation methods. In general, there are 3 approaches:

1. Internal criteria
2. External criteria
3. Relative criteria

Internal and external approaches are based on some metrics related to dataset and cluster scheme and some user specific intuition, respectively. For internal indices, we evaluate the results using quantities, features and their correlations inherent in the data set. Finding the optimal number of clusters in unsupervised classification is usually determined based on an internal validity index. For external indices, we evaluate the results of a clustering algorithm based on known structure of a data set (or cluster labels). They are computationally expensive and need high computation power. The third approach of clustering validity is based on relative criteria. Here the basic idea is the evaluation of a clustering structure by comparing it to other clustering schemes, resulting by the same algorithm but with different parameter values. There are two criteria proposed for clustering evaluation and selection of an optimal clustering scheme (Halkidi, Batistakis, Vazirgianni, 2001):

1. Compactness
2. Separation

Compactness means that each member in a cluster should be as close as possible to each other. Variance (lower variance indicates better compactness) and distance measures are used for this criterion. However, separation measures how distinct or well-separated a cluster is from other clusters.

Although validity indices for K-means calculate the distance between points and cluster centers, in FCM usually membership values are used to evaluate the clustering results. This is because it is independent of the distance and location of data points. But recently it is widely accepted that better results are achieved by considering membership values and data set itself (Wang, Zhang, 2007).

One of the most famous and well-known methods for determining number of clusters is Silhouette method (Rousseeuw, 1987). The Silhouette index validates the clustering performance based on the pairwise difference of between and within-cluster distances. In addition, the optimal cluster number is determined by maximizing the value of this index. For each point i Silhouette Coefficient is defined as

$$S(i) = \frac{\min \{b(i) - a(i)\}}{\max \{a(i), b(i)\}} \quad (4.20)$$

where $a(i)$ is the mean distance between i^{th} point and all other points in the same class, and $b(i)$ is the mean distance between i^{th} point and all other points in the next nearest cluster. For each point, Silhouette coefficients could be found between -1 and +1. When $S(i)$ is at its largest value (+1), it is indicated that the sample is far away from the neighboring clusters. Value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicates that those samples might have been assigned to the wrong cluster.

The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually.

In addition, overall average silhouette width ($S_{(k)}$) can be considered in order to select the appropriate value for k . In this case, we calculate $S_{(k)}$ for different number of k and consider the largest value.

Some of validity methods are summarized for K-means and FCM algorithms in Tables 4.3 and 4.4, respectively.

4.4.3 Accuracy Measure

In order to evaluate the performance of classifier or to find optimal values for parameters for a classifier, accuracy measure may be calculated. For our case, there are two kinds of responses for each decision: true or false. According to these responses, there are 4 possible situations:

- True positive (TP): PAF patients are classified as PAF.
- True negative (TN): non-PAF patients are classified as non-PAF.
- False positive (FP): non-PAF patients are classified as PAF.
- False negative (FN): PAF patients are classified as non-PAF.

Table 4.3 K-means validity indices

<i>Method</i>	<i>Formula</i>	<i>Description</i>	<i>Optimal Value</i>
Silhouette	$S(i) = \frac{\min \{b(i) - a(i)\}}{\max \{a(i), b(i)\}}$	Results are between [-1 1]	The largest value
Score Function (SF)	$bcd = \frac{\sum_{i=1}^k C_i - C_{tot} \cdot n_i}{n \cdot k}$ $wcd = \sum_{i=1}^k \left(\frac{1}{n_j} \sum \ x - c_i\ \right)$ $SF = 1 - \frac{1}{e^{e^{bcd-wcd}}}$	Results are between [0 1]	The higher the value of the SF, the more suitable the number of clusters
SD Validity Index	$Scat = \frac{1}{k} \sum_{i=1}^k \frac{\sigma(v_i)}{\sigma(x)}$ $Distance = \frac{\max (V_j - V_i)}{\min (V_j - V_i)} \sum_{i=1}^k \left(\sum_{j=1}^k V_j - V_i \right)^{-1}$ $SD \text{ index} = \alpha \cdot Scat + Distance$	α is weighting factor, which is equal to Distance parameter in case of maximum number of clusters. it is indicated that there is no significant influence of α value on SD index results.	The least value

Table 4.4 Fuzzy C-means validity indices

<i>Method</i>	<i>Formula</i>	<i>Description</i>	<i>Optimal Value</i>
PC	$\frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n U_{ij}^2$		Maximum: C ([2, n-1])
PE	$-\frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n U_{ij} \log U_{ij}$		Minimum: C ([2, n-1])
MPC	$1 - \frac{C}{C-1} \left[1 - \left(\frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n U_{ij}^2 \right) \right]$	Modification of PC index	[0 1] and optimal number of cluster has reached when we find the max value
Xie and Beni	$\frac{\sum_{i=1}^c \sum_{j=1}^n U_{ij}^m \ X_j - C_i\ ^2}{n \min \ C_i - C_j\ ^2}$	The numerator indicates the compactness and dominator shows how the clusters separated	High values for separa- tion can give us good results. Minimum: [2 ≤ C ≤ n-1]

Table 4.4 continues

Kwon Index	$\frac{\sum_{i=1}^c \sum_{j=1}^n U_{ij}^2 \ X_j - C_i\ ^2 + \frac{1}{C} \sum_{i=1}^c \ C_i - Me\ ^2}{\min \ C_i - C_j\ ^2}$ $Me = \sum_{j=1}^n X_j / n$	<p>Improved XB algorithm aims to eliminate the monotonically decrease in case of increasing number of cluster</p> <p>Minimum: [$2 \leq C \leq n-1$]</p>
V_{sc} Index	$SC1 = \frac{\sum_{i=1}^c \ C_i - Me\ ^2}{C \times \sum_{i=1}^c \left(\frac{\sum_{j=1}^n U_{ij}^m \ X_j - C_i\ ^2}{\sum_{j=1}^n U_{ij}} \right)}$ $SC2 = \frac{\sum_{i=1}^c \sum_{l=i+1}^c \left(\sum_{j=1}^n (\min(U_{ij}, U_{lj}))^2 / \sum_{j=1}^n \min(U_{ij}, U_{lj}) \right)}{\sum_{j=1}^n (\max(U_{ij}, U_{lj}))^2 / \sum_{j=1}^n \max(U_{ij}, U_{lj})}$ $V_{sc} = SC_1 - SC_2$	<p>Maximum: [$2 \leq C \leq n-1$]</p>

Using these possibilities, confusion matrix is created which represents performance of the classifier (Table 4.5).

Table 4.5 The confusion matrix

		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

In this study, during classification PAF subjects are recognized as ‘Positive’ and non-PAF subjects are regarded as ‘Negative’. To analyze the performance of the classifier, using confusion matrix some parameters are utilized:

- **Sensitivity:** refers to the proportion of the performance of the classifier in recognizing True positives. In our case, it shows the ratio of the correctly classified PAF subjects to the total number of PAF subjects:

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \quad (4.21)$$

- **Specificity:** is the measure of detecting True negatives. In our case, it shows the ratio of the correctly classified non-PAF subjects to the total number of non-PAF subjects:

$$Specificity = \frac{TN}{TN + FP} \times 100 \quad (4.22)$$

- **Precision:** can be thought of as a measure of a classifiers exactness:

$$Precision = \frac{True\ Positives}{True\ Psitives + False\ Positives} \times 100 \quad (4.23)$$

- **Accuracy:** overall accuracy is calculated by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (4.24)$$

- **F-Score:** F score or F1 value conveys the balance between the precision and the recall:

$$F1 = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \quad (4.25)$$

- **ACU:** The area under (a ROC) curve is used in classification analysis in order to determine which of the used models predicts the classes best. This is done by plotting false positive rate (1-specificity) over true positive rate (sensitivity), find point in the plot and calculate area under it.

CHAPTER FIVE

METHODS AND RESULTS

Clustering and classification methods are widely used in clinical fields to determine people with health problems. Data representation is one of the most important factors that influence the performance of the clustering algorithm. If the representation (choice of features) is good, the clusters are likely to be compact and isolated and even a simple clustering algorithm will find them perfectly. Unfortunately, there is no universally good representation; the choice of representation must be guided by the domain knowledge.

Physionet is one of the mostly used and popular web-based datacenters which provides large collections of recorded physiologic signals and related open-source software. In this study, The PAF Prediction Challenge database is used (Goldberger et al., 2000). The dataset consists of two channel long-term ECG recordings sampled with 128 Hz per signal with 12-bit resolution. The RR interval series derived from these QRS occurrence times were used in this study. The database has 100 ECG record sets obtained from 98 different subjects. There are 3 types of subjects;

- 1) Records which is not PAF (non-PAF)
- 2) Records just before PAF attack (prior to PAF)
- 3) Records 45 minutes after PAF attack (distinct from PAF)

In order to analyze the data, only non-PAF and distinct from PAF records were considered. All records were extracted from longer ECG records. Our data collection is derived from HRV features obtained from normal sinus rhythm (NSR) ECG records and based on the usage of genetic algorithm in order to select best features from ECG records (Hilavin, 2016).

5.1 Methods and Parameters

This study contains different experiments but before clarifying each of them, it is necessary to consider some aspects as follows:

5.1.1 Determining the Number of Clusters

Silhouette and Elbow are considered methods in order to find optimal number of clusters because they are not expensive in calculation and easy to implement.

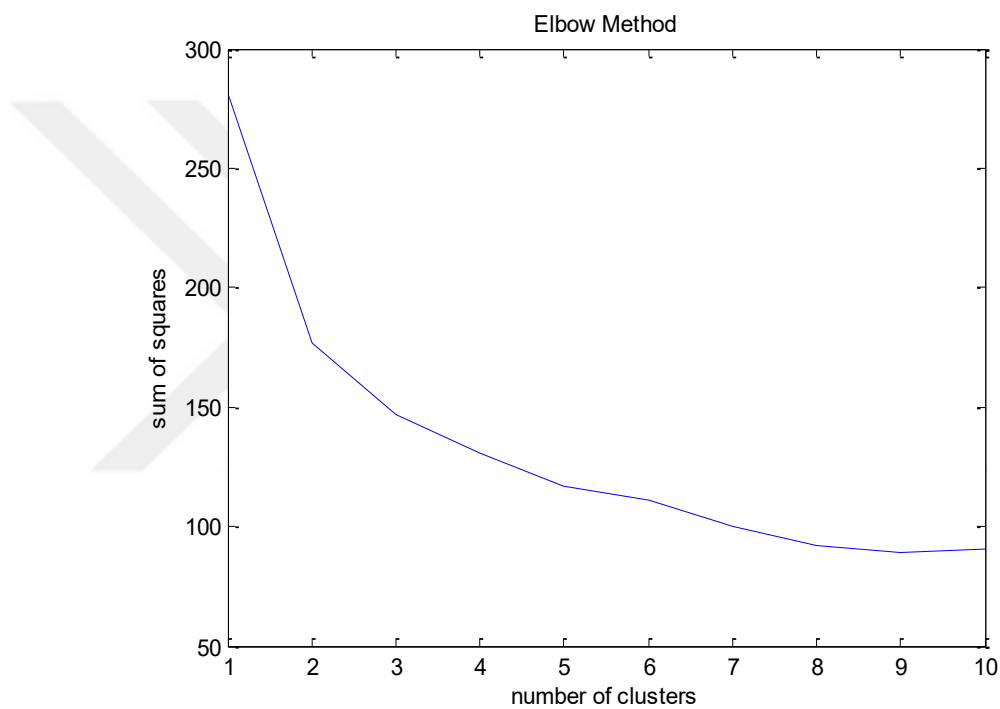


Figure 5.1 The Elbow can be distinguished at $k=2$

In Elbow method, sometimes it is not easy to detect the elbow point but in our case the elbow is seen easily (Figure 5.1). Therefore, this method determines the optimal number of clusters as 2. On the other hand, trend in Silhouette is not like the one in Elbow method. The silhouette value for each point is a measure of similarity of that point to points in its own cluster. With a comparison between different plots in Figure 5.2, in case $k=2$ most points have a high silhouette value and negative values are less than others.

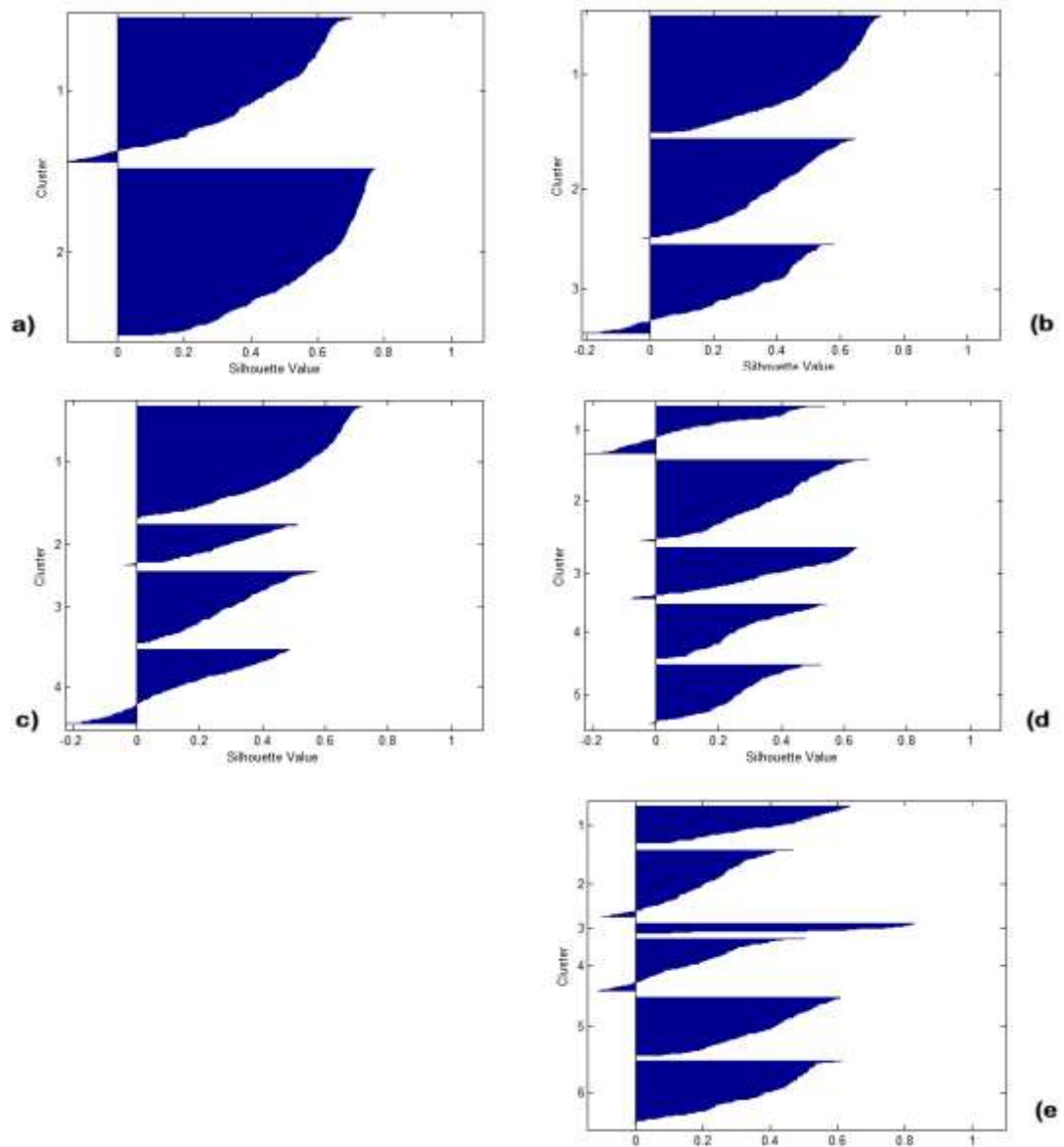


Figure 5.2 Silhouette plots for a) $k=2$ b) $k=3$ c) $k=4$ d) $k=5$ e) $k=6$

Beside Silhouette plots, overall average silhouette width may also be used for finding optimal number of clusters (see Table 5.1). The larger the values for average Silhouette width, the more suitable the case. Both Silhouette plot and overall average value indicate that the number of clusters should be chosen as 2.

Table 5.1 Average Silhouette values for different number of clusters

Number of clusters	2	3	4	5	6
Average silhouette width	0.5016	0.4021	0.2927	0.3536	0.35507

5.1.2 Cross-Validation

One of the negative predictions in machine learning models is overfitting. Overfitting occurs when the model learns every detail in training set and performs pretty well. In this situation when a new data comes, the model cannot have a good performance. One of the reasons cross-validation is used is to eliminate this problem. 4-fold cross validation is implemented to classify the whole dataset into 4 different groups. As represented in Figure 5.3, one of these groups is Test set and the others are Train set. Our dataset contains 510 non-PAF subject ECG records and 288 PAF subject ECG records. The number of PAF records in all folds are equal.

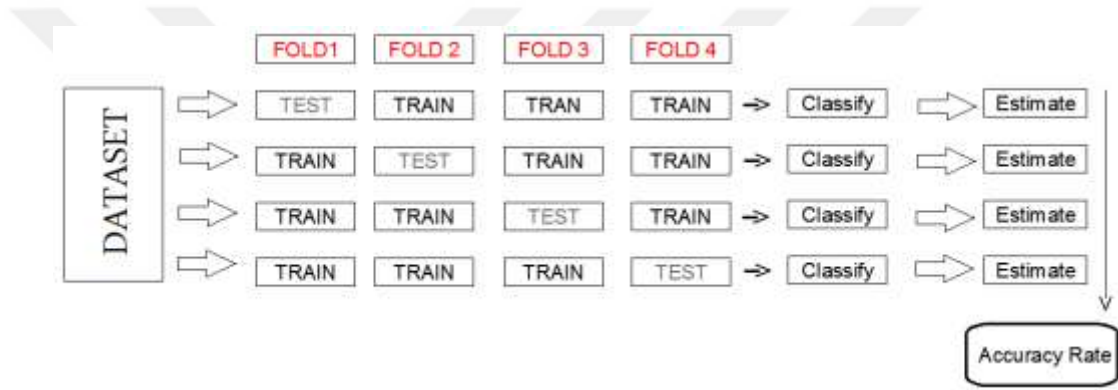


Figure 5.3 The scheme of dividing dataset to 4-folds using cross validation. Total accuracy is achieved by finding the mean of all estimations

Unsupervised classification methods are implemented to these Test and Train sets separately. As illustrated in Figure 5.3, the procedure is repeated 4 times and accuracy is calculated. Then, the average of these estimates represents our final accuracy rate.

5.2 Experiments

Different algorithms and models are employed to analyze data. To have a proper interpretation on performance of models, each algorithm is applied 20 times and mean of results is calculated. It is important to consider the point that without prior knowledge of categories those obtained label vectors cannot be directly used for the following conclusion. For instance, although the label vectors $[1, 2, 5, 3, 4, 6, 4]^T$

and $[3, 2, 5, 1, 6, 4, 6]^T$ are different in statement, they come up with the same clustering result. To combine different clustering results, the cluster label vectors should be aligned. Two different perspectives, single clustering algorithms and ensemble models, are explained in the next section.

5.2.1 Single Clustering Algorithms

First experiment is performed by applying K-means, FCM, GIFP-FCM and S-FCM algorithms (see Figure 5.4). This experiment is done by considering random initial center selection.

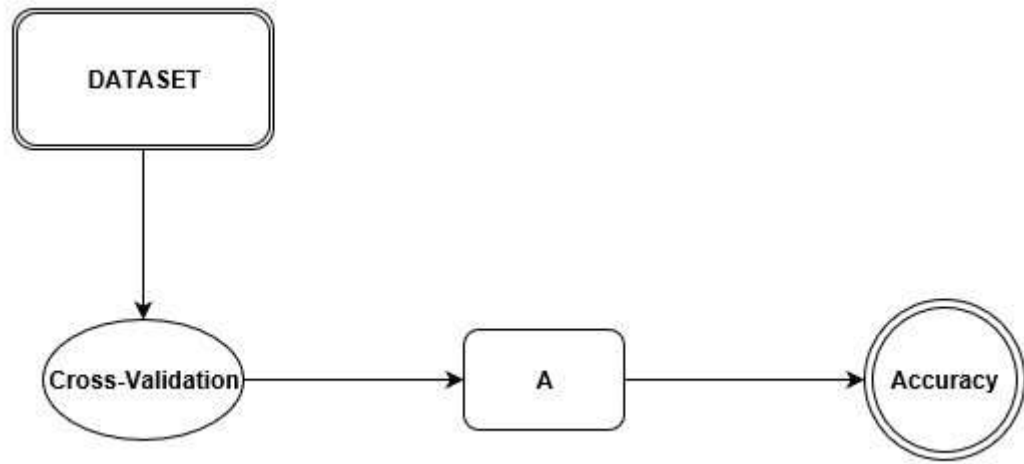


Figure 5.4 Procedure of a typical clustering analysis. Box A can be any clustering algorithm

In Table 5.2, all measurements for each algorithm are summarized.

5.2.2 Ensemble Models

To achieve better results, the combination of clustering algorithms is considered. One possible way is illustrated in Figure 5.5. Membership values of FCM algorithms are accumulated as first step and are given to K-means algorithm as second step. FCM, GIFP-FCM and S-FCM algorithms are used to generate membership values in first step. Results are given in Table 5.3.

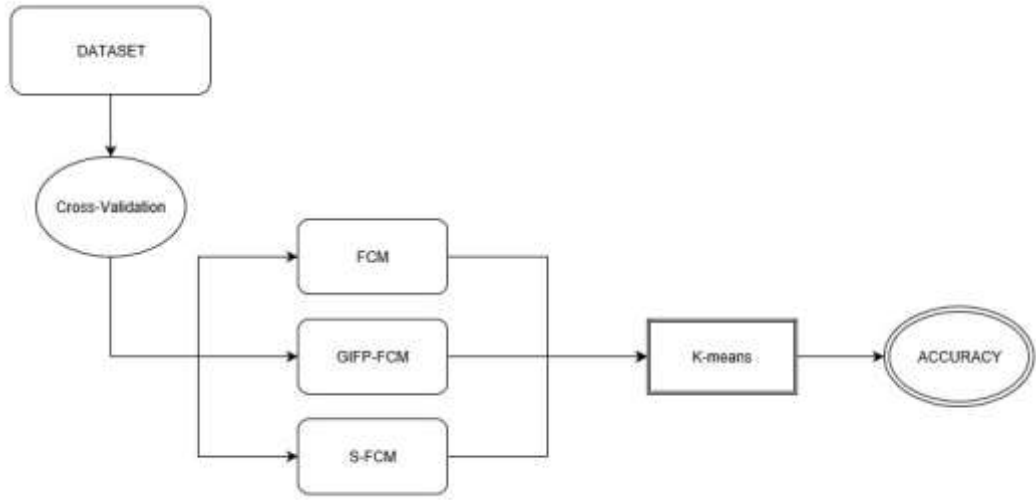


Figure 5.5 Ensemble of different algorithms based on membership values

Applying threshold to membership values (U_{ij}) is another method introduced to enhance the performance of a single clustering algorithm. After obtaining membership values (by using FCM or FCM-based algorithms) following rule is applied on membership values to achieve better understanding of data points by eliminating some portion of data.

$$U = \begin{cases} U_{ij} \geq \gamma, & PAF \\ U_{ij} < 1 - \gamma, & nonPAF \\ otherwise, & Unknown \end{cases} \quad (5.1)$$

The Equation 5.1 determines the fact that for a threshold function (γ), a simple data point can be assigned in 2 ways; belonging to one of clusters or unlabeled. To avoid any unpredicted trouble, threshold value (γ) must be set between $[0.5, 1)$. A large γ corresponds to a small amount of points. However, result is equivalent to FCM in the case that $\gamma = 0.5$. On the other hand, for small values of γ we may have small number of unassigned points. Therefore, results strongly depend on the choice of the value of γ and structure of data.

Table 5.2 Results of clustering algorithms after 20 times implementation

	Sensitivity	Specificity	Precision	F1	Accuracy	AUC
K-means	65.38	65.21	51.81	57.77	65.27	65.3
FCM	64.64	62.54	49.55	56.07	63.30	63.59
GIFP-FCM	65.87	64.75	51.56	57.80	65.15	65.31
SFCM	62.93	62.54	49.03	55.09	62.68	62.74

Table 5.3 Ensemble model results after 20 times iteration

	Sensitivity	Specificity	Precision	F1	Accuracy	AUC
K-means Ensemble (Experiment 1)	64.22	63.22	49.92	56.14	63.58	63.72

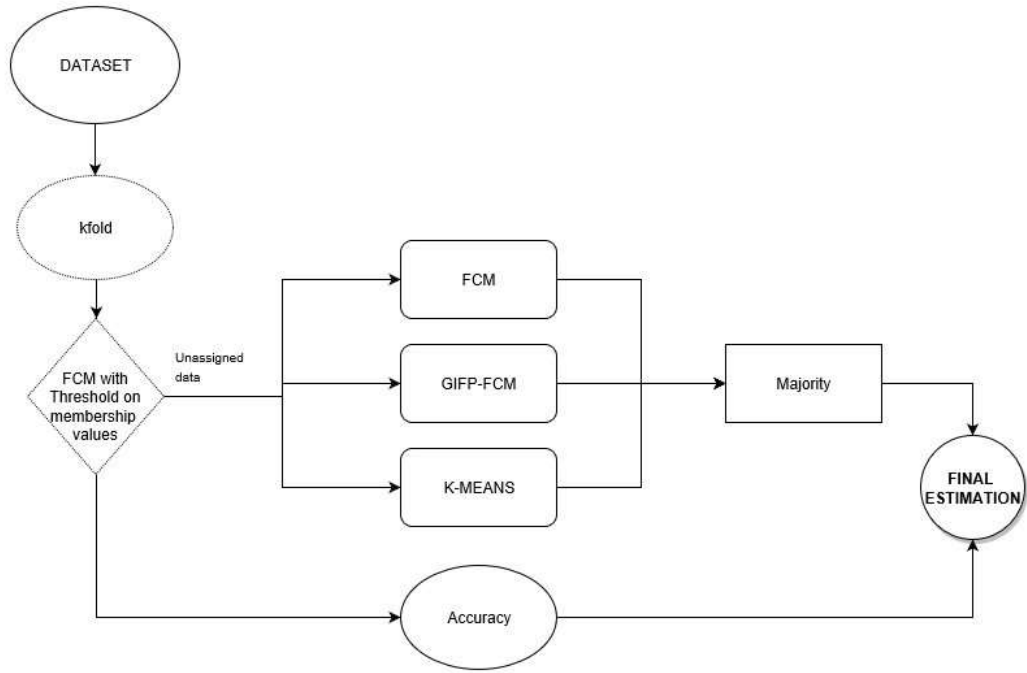


Figure 5.6 An ensemble approach based on fuzzy membership restriction rule followed by 3 different clustering methods

As shown in Figure 5.6, the new model consists of 3 steps. At first step, after implementing a standard FCM algorithm, restriction on membership values (Equation 5.1) is implemented at each fold and then new data (unlabeled data points) are derived. In second step, obtained data is given to K-means, FCM and GIPF-FCM algorithms in order to be partitioned. Thus, for an unlabeled data point we obtain 3 assessments. To have better decision, results of these 3 algorithms are accumulated in a matrix on which majority voting is performed. Whether it is a crisp or a fuzzy one, voting is a formal way of combining the opinions of several voters into a single consolidated decision. The ensemble output is assigned to the class with the maximum number of votes among all classes. For better estimation, various threshold values have been considered. The final evaluation of the model is given in Table 5.4, where threshold values are indicated in percentage.

Figure 5.7 illustrates the comprehensive approach of previous model. Although in the second step of this model different algorithms may be utilized, it should be highlighted that in the first step FCM algorithm should be performed in order to obtain

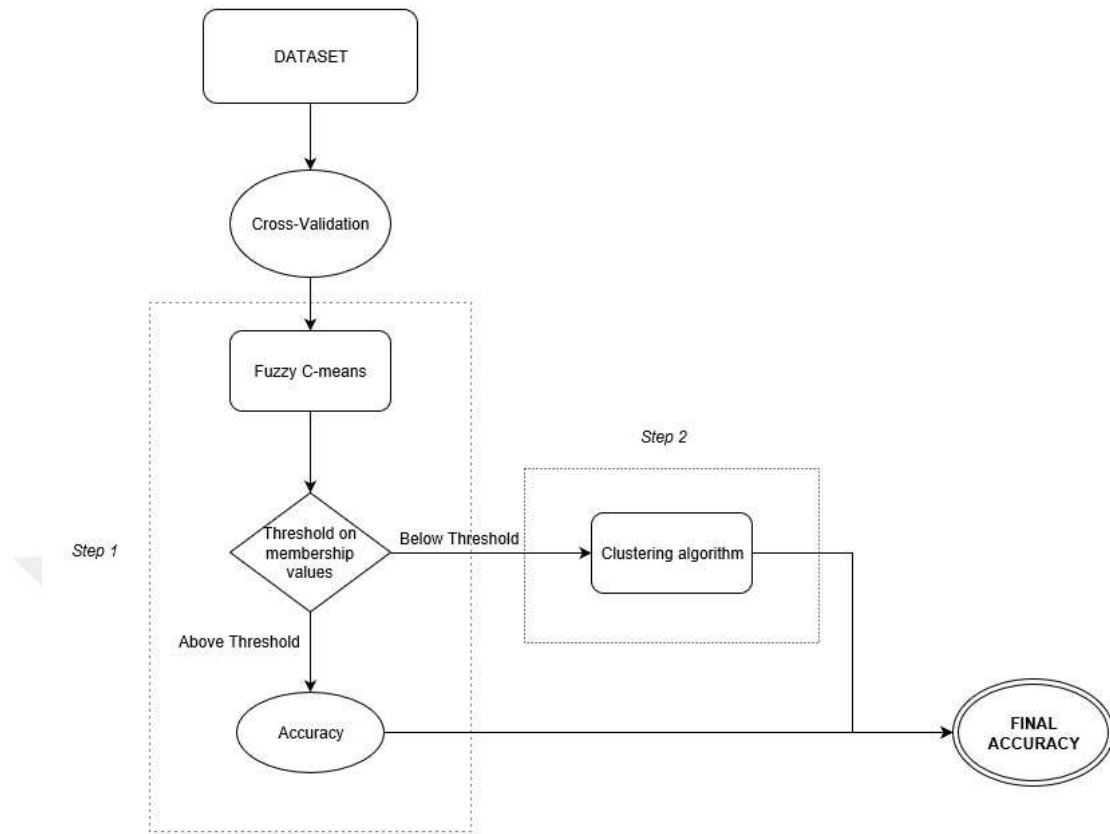


Figure 5.7 Comprehensive ensemble clustering approach

membership degrees and apply threshold on them. Previous model uses this technique in which 3 different algorithms are utilized for the second step. The advantage of using different algorithms is that it may be useful in eliminating the weak results of some algorithms. Therefore, combination of algorithms in the second step can contribute to better clustering results. On the other hand, this combination not only significantly increases the mathematical calculations, but also due to the probability of weak performance of some clustering algorithms, the results may not be satisfying. Thus, in the second step the right algorithm should be applied. In the next experiment, instead of using different algorithms in the second step, just K-means algorithm is performed which has simplified the model. Final results are summarized in the Table 5.5.

Table 5.4 Results of ensemble model by considering different threshold values. Threshold values are presented in percentage. For each step, accuracy is calculated

		Sensitivity	Specificity	Precision	F1	Accuracy	AUC
Ensemble Model (Experiment 2)	55%	54.22	50.20	40.34	46.25	54.45	54.22
	60%	45.66	43.53	36.55	40.58	45.66	43.53
	65%	43.47	58.04	37.02	39.97	52.53	50.76
	70%	46.89	51.96	41.03	43.70	56.35	49.43
	75%	42.67	51.57	39.18	40.75	55.18	47.12
	80%	53.13	49.80	45.40	48.85	59.67	51.46
	85%	62.83	61.96	49.29	55.22	62.99	62.40
	90%	73.87	68.04	55.80	63.57	69.42	70.96
	95%	74.25	68.24	57.07	64.53	70.55	71.24

Table 5.5 Final performance of ensemble model with restriction on membership values followed by K-means

		Sensitivity	Specificity	Precision	F1	Accuracy	AUC
Fuzzy K-means ensemble Model (Experiment 3)	55%	55.64	55.62	47.51	47.51	55.64	55.62
	60%	49.24	57.64	44.11	44.11	54.61	53.44
	65%	41.63	57.55	38.57	38.57	51.80	41.63
	70%	39.05	61.15	37.84	37.84	39.05	61.15
	75%	37.24	62.12	36.65	36.65	53.14	49.68
	80%	48.39	66.60	46.87	46.87	60.03	57.49
	85%	64.79	64.87	57.17	57.17	64.84	64.83
	90%	72.99	67.81	63.48	63.48	69.68	70.40
	95%	73.75	69.02	64.52	64.5	70.73	73.75

In addition, the number of unclassified data points achieved after using equation 5.1 contains some information. As can be seen at Table 5.6, with increasing the threshold value, number of unassigned data points increases as well, which indicates presence of small number of data points with high membership degree. Thus, a good examination of distribution of clusters can be obtained by considering this criterion.

Table 5.6 Amount of number of unassigned data points for various threshold values

Threshold	55%	60%	65%	70%	75%	80%	85%	90%	95%
Number of unassigned data points	44	104	171	224	316	425	588	742	795

CHAPTER SIX

CONCLUSION

Atrial fibrillation (AF) is the most common arrhythmia type. In this arrhythmia atria cannot completely push the blood to ventricles and consequently clot formation can occur which may lead to serious health problems or stroke. In most cases, AF starts with short episodes and progresses to longer and developed form over a time. Paroxysmal AF is a type of AF in which the episodes last 48 hours up to 7 days and terminate by themselves. If paroxysmal AF continues, it is converted into Persistence AF.

The objective of this thesis is to analyze the data obtained from arrhythmia-free ECG records of PAF and non-PAF subjects using different clustering algorithms. Performance of clustering algorithms is data dependent. Thus, choosing right clustering technique for a given dataset is a research challenge. Before implementing clustering algorithms, the dataset is divided into Test and Train sets. Chapter 5 deals with searching methods in order to successfully split the dataset. The analysis methods are comprised of two perspectives; single clustering algorithms and ensemble methods.

First experiment represents the usage of single clustering algorithms. Four different algorithms are applied and their performances are compared in terms of discrimination. Referring to the results, all algorithms perform very close to each other but in case of accuracy, K-means is slightly better than others.

In order to split dataset better than previous experiment, second experiment aims to enhance results of one clustering algorithm by combining the outcomes of soft and hard techniques. In this model, outcome of FCM (membership values) is given to K-means algorithm. Results determine that using K-means in the second step is not effective enough.

Finally, in the last experiment an ensemble model is introduced by using cut-off membership values and, in consequence, new data is extracted. But it should be

considered that without any prior knowledge about the data, this procedure could run into problems. 70%, 80%, 90% and 95% thresholds for membership values are considered. Two points of view are taken into account. Firstly, at the second step of the model, a combination of 3 different clustering algorithms followed by voting method is performed. Accuracy for given threshold values are %55.51, %60.66, %69.52 and %70.41, respectively. Higher accuracy values are achieved by applying 95% threshold. Secondly, instead of using three different algorithms in Step 2, which complicates our model, K-means algorithm is performed. Results are obtained as %56.62, %63.46, %69.82 and %70.71, respectively. Therefore secondly introduced model is slightly better than first one in which using one strong algorithm reduces calculation time and increases the accuracy of model. Sensitivity and Specificity values for given threshold (95%) are found as 73.70% and 69.02%, respectively. Consequently, ensemble models definitely represent better performance in terms of Accuracy, Sensitivity and Specificity compared to a single clustering method.

Another interesting result that comes from using proposed ensemble model is the number of unlabeled data points concluded from the first step. As discussed before, depending on threshold value some data points are labeled and remaining ones are not assigned. For above mentioned threshold values, the number of labeled data points are 574, 373, 56 and 3, respectively. This demonstrates that large threshold values result in more unlabeled data points. In other words, for our dataset, the number of data points away from cluster center (points with low membership value) is more than the number of data points in the central region. As a consequence, the dataset is not well-separated and distribution of two clusters are close to each other.

The executed methods in this thesis require further developments which take special efforts in spite of the efficiency and adequate consistency. Referring to different experiments, which include conventional and hybrid approaches, the obtained results show that distinguishing two groups in the given dataset as PAF and non-PAF subjects using only HRV features obtained from normal sinus rhythm (NSR) ECG records is not an easy task. One possible interpretation of the results indicates the complex structure of dataset. In other words, there is an intersection

among these two groups. Since in case of PAF subjects, records are obtained about 45 minutes away from PAF attacks, structural point of view led us to have a conclusion on existence of similarity between two data groups. Owing to this significant overlapping between PAF and non-PAF subjects, a sophisticated classifier should be employed in order to successfully identify each group.



REFERENCES

- Arnaldo, H. A., & Bedregal, B. R. C. (2013). A new way to obtain the initial centroid clusters in Fuzzy C-Means algorithm. In *Proceedings of the 2nd Workshop-School on Theoretical Computer Science*, 139-144.
- Arthur. D., Vassilvitskii.S. (2007). K-means++: The advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027-1035.
- Bezdeck, J. C. (1974). Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3(3), 58–73.
- Bezdeck, J.C., Ehrlich, R., Full, W. (1984). FCM: The fuzzy C-means clustering algorithm. *Computers and Geoscience*, 10(2–3), 191-203.
- Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press: New York.
- Boongoen, T., & Iam-On, N. (2018). Cluster ensembles: A survey of approaches with recent extensions and applications. *Computer Science Review*, 28, 1-25.
- Bora, D. J., Gupta, D., & Kumar, A. (2014). A comparative study between fuzzy clustering algorithm and hard clustering algorithm. *International Journal of Computer Trends and Technology (IJCTT)*, 10(2), 108-113.
- Cebeci, Z., Yildiz, F. (2015). Comparison of K-Means and fuzzy C-means algorithms on different cluster structures. *Journal of Agricultural Informatics*, 6(3), 13-23.
- Clifford, G. D., Azuaje, F., McSharry, P. (2006). *Advanced Methods and Tools for ECG Data Analysis*. Norwood, MA, USA: Artech House.
- Dave, R. N. (1996). Validating fuzzy partition obtained through c-shells clustering. *Pattern Recognition Letters*, 17, 613–623.
- Davies, A., & Scott, A. (2014). *Starting to Read ECGs: A Comprehensive Guide to Theory and Practice*. London: Springer-Verlag.

- Dayan, G. (2006). Arrhythmia classification with SOM. M.Sc Thesis, Dokuz Eylül University, Izmir.
- Donoso, F. I., Figueroa, R. L., Lecannelier, E. A., Pino, E. J., & Rojas, A. J. (2013). Clustering of atrial fibrillation based on surface ECG measurements. In *Engineering in Medicine and Biology Society (EMBC), 35th Annual International Conference of the IEEE*, 4203-4206.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern Classification*. New York: John Wiley & Sons.
- Fan, J. L., Zhen, W. Z., & Xie, W. X. (2003). Suppressed fuzzy C-means clustering algorithm. *Pattern Recognition Letters*, 24(9), 1607-1612.
- Fix, E., Hodges, J. L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *USAF School of Aviation Medicine*.
- Gacek, A., & Pedrycz, W. (Eds.). (2011). *ECG signal processing, classification and interpretation: A comprehensive framework of computational intelligence*. London: Springer-Verlag.
- Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: Theory, algorithms, and applications*. Philadelphia: American Statistical Association and the Society for Industrial and Applied Mathematics.
- Ghaemi, R., Sulaiman, M. N., Ibrahim, H., & Mustapha, N. (2009). A survey: Clustering ensembles techniques. *World Academy of Science, Engineering and Technology*, 50, 636-645.
- Ghosh, S., & Dubey, S. K. (2013). Comparative analysis of K-means and fuzzy C-means algorithms. *International Journal of Advanced Computer Science and Applications*, 4(4), 35-39.
- Gokana, V., Phua, C. T., & Lissorgues, G. (2014). Automatic detection of atrial fibrillation using RR interval from ECG signals. In *The 15th International Conference on Biomedical Engineering*, 215-218.

- 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.
- Grabusts, P., & Borisov, A. (2002). Using grid-clustering methods in data classification. In *International Conference on Parallel Computing in Electrical Engineering*, 425-426.
- Green, S. G., Welsh, M. A., & Dehler, G. E. (2003). Advocacy, performance, and threshold influences on decisions to terminate new product development. *Academy of Management Journal*, 46(4), 419-434.
- Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). Clustering algorithms and validity measures. In *Proceedings. Thirteenth International Conference on Scientific and Statistical Database Management*, 3-22.
- Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques, *Journal of Intelligent Information Systems*. 17(2-3), 107-145.
- Halkidi, M., Vazirgiannis, M., & Batistakis, Y. (2000). Quality scheme assessment in the clustering process. *Principles of Data Mining and Knowledge Discovery*, 265 - 276.
- Hilavin, İ. (2016). *Development of a system to diagnose Paroxysmal Atrial Fibrillation patients from arrhythmia free ECG records*. PhD Thesis, Dokuz Eylül University, Izmir.
- Höppner, F., Klawonn, F. (2003). Improved fuzzy partitions for fuzzy regression models. *International Journal of Approximate Reasoning*, 32(2-3), 85–102.
- Jang, J. S. R., Sun, C. T., Mizutani, E. (1997). Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence. USA: Prentice-Hall.
- Ka-Chun Wong. (2015). A Short Survey on Data Clustering Algorithms. *2nd International Conference on Soft Computing and Machine Intelligence*. 64-68.

- Kaur, D. (2014). A comparative study of various distance measures for software fault prediction. *International Journal of Computer Trends and Technology (IJCTT)*, 17(3): 117-120.
- Kikillus, N., Hammer. G., Wieland, S. and Bolz, A. (2007). Algorithm for Identifying Patients with Paroxysmal Atrial Fibrillation without Appearance on the ECG. *29th Annual International Conference Engineering in Medicine and Biology Society*, 275-8.
- Kirchhof. P., Benussi. S., Kotecha. D., Ahlsson. A., Atar. D., Casadei. B., et al. (2016). 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. *European Heart Journal*, 37(38), 2893–2962.
- Kovács, K., Legány, C., Babos, A. (2005). Cluster validity measurement techniques. In *Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, 388-393.
- Kwon, S. H. (1998). Cluster validity index for fuzzy clustering. *Electronics Letters*, 34(22), 2176-2177.
- Li, J., Fan, J. (2014). Parameter selection for suppressed fuzzy C-means clustering algorithm based on fuzzy partition entropy. *11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 82-87.
- Long, X., Fonseca, P., Haakma, R., Aarts, R. M., & Foussier, J. (2012). Time-frequency analysis of heart rate variability for sleep and wake classification. In *Bioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference on*, 85-90.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*, 1(14), 281-297.
- Najarian, K., & Splinter, R. (2012). *Biomedical Signal and Image Processing*. (Second Edition). NW: CRC Press.

- Nixon, M., Aguado, A. (2008). *Feature Extraction & Image Processing*. (2nd Edition). San Diego: Academic Press.
- Panda, S., Sahu, S., Jena, P., & Chattopadhyay, S. (2012). Comparing Fuzzy-C means and K-means clustering techniques: A comprehensive study. *Advances in Computer Science, Engineering & Applications*, 451-460.
- Pelleg, D., Moore, A. W. (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *In Proceedings of Proceedings of the 17th International Conference on Machine Learning, 1*, 727-734.
- Pourbabae, B., & Lucas, C. (2008). Automatic detection and prediction of paroxysmal atrial fibrillation based on analyzing ECG signal feature classification methods. In *Biomedical Engineering Conference (CIBEC 2008)*, 1-4.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.
- Saitta, S., Raphael, B., Smith, L. F. C. (2007). Bounded Index for Cluster Validity. In *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition*, 174-187.
- Samoilov, A. A., Telishev, D. V., & Pyanov, I. V. (2018). Application of heart rate variability methods to the TT-intervals and ST-segments durations. In *Young Researchers in Electrical and Electronic Engineering (EIConRus)*, 1928-1931.
- Seyd, P. A., Ahamed, V. T., Jacob, J., & Joseph, P. (2008). Time and frequency domain analysis of heart rate variability and their correlations in diabetes mellitus. *International Journal of Biological and Life Sciences*, 4(1), 24-27.
- Shedthi, B. S., Shetty, S., & Siddappa, M. (2017). Implementation and comparison of K-means and fuzzy C-means algorithms for agricultural data. In *Inventive Communication and Computational Technologies (ICICCT)*, 105-108.

- Stetco, A., Zeng, X., Keane, J. (2015). Fuzzy C-means++: Fuzzy C-means with effective seeding initialization, *Expert Systems with Applications*, 42(21), 7541–7548.
- Task Force of the European Society of Cardiology. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93, 1043-1065.
- Wang, W., Zhang, Y. (2007). On fuzzy cluster validity indices. *Fuzzy Sets and Systems*, 158(19), 2095-2117.
- Wong, K. C. (2015). A short survey on data clustering algorithms. *Second International Conference on Soft Computing and Machine Intelligence (ISCMI)*, 64-68.
- Xhyheri, B., Manfrini, O., Mazzolini, M., Pizzi, C., & Bugiardini, R. (2012). Heart rate variability today. *Progress in Cardiovascular Diseases*, 55(3), 321-331.
- Xie, X. L., Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8), 841–847.
- Yang, Q., Zhang, D., Tian, F. (2010). An initialization method for fuzzy C-means algorithm using subtractive clustering. *In 3rd International Conference on Intelligent Networks and Intelligent Systems (ICINIS)*, 393–396.
- Yuan, F., Meng, Z. H., Zhang, H. X., Dong, C. R. (2004). A new algorithm to get the initial centroids. *In Proceedings of International Conference on Machine Learning and Cybernetics*, 2, 1191-1193.
- Zahid, N., Limouri, M., Essaid, A. (1999). A new cluster-validity for fuzzy clustering. *Pattern Recognition*, 32 (7) 1089–1097.
- Zhu, L., Chung, F. L., Wang, S. (2009). Generalized fuzzy C-means clustering algorithm with improved fuzzy partitions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(3), 578-591.
- Zhu, M., Wang, W., Huang, H. (2014). Improved initial cluster center selection in K-means clustering. *Engineering Computations*, 31(8). 1661 – 1667.