DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

PAF SCREENING FROM SINUS RHYTHM ECG RECORDS BY ENSEMBLE LEARNING

by

Fırat BİLGİN

December, 2017

İZMİR

PAF SCREENING FROM SINUS RHYTHM ECG RECORDS BY ENSEMBLE LEARNING

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

by
Fırat BİLGİN

December, 2017
İZMİR

M. Sc. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"CLASSIFICATION OF BIOMEDICAL SIGNALS USING COMMITTE MACHINES NEURAL NETWORKS"** completed by **FIRAT BİLGİN** 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

Yard.Doç.Dr. Nalan ÖZKURT

Jury Member

Yard.Doç.Dr. Güleser K. Demir

Jury Member

Prof. Dr. Kadriye ERTEKİN

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to thank my thesis advisor Prof. Dr. Mehmet KUNTALP for giving me an opportunity to work with him. I gained invaluable experience thanks to his advice and guidance.

I also would like to thank my parents for their support and encouragement.

I also would like to thank to Council of Higher Education for supporting me in the framework of ÖYP.

I wish to thank to staff of the Graduate School of Natural and Applied Sciences of Dokuz Eylül University for their valuable support.

Fırat BİLGİN

PAF SCREENING FROM SINUS RHYTHM ECG RECORDS BY ENSEMBLE LEARNING

ABSTRACT

Signal processing has always been of great interest. As the computers can do more complex processes, some different and new mathematical formulations are adapted to signal processing. Terms such as machine learning have been accepted by the literature and some combining methods have been developed. Ensemble learning is a combining method of using a combination of different experts to get better results in pattern classification. Briefly, ensemble learning is a method whereby different classifiers work together. In this study, ensemble learning was used for the aim of paroxysmal atrial fibrillation (PAF) screening, i.e. finding whether a person is PAF patient or not from his/her ectopic-free electrocardiogram (ECG) records. Both hierarchical and parallel structures of ensemble learning were tried. Dataset used consists of ECG records from both PAF patients and non-PAF subjects. To train experts, k–fold cross validation and bootstrap sampling methods were used and their performances were compared. The best results were obtained by using the hierarchical structure of ensemble learning.

Keywords: Signal processing, bagging, bootstrap aggregation, ensemble learning, committee machines, k fold cross validation, PAF screening

BİRLEŞİK ÖĞRENME KULLANILARAK SİNÜS RİTİM EKG KAYITLARINDAN PAF TARAMASI

ÖZ

Sinyal işleme her zaman büyük bir ilgiye sahip olmuştur. Bilgisayarlar daha karmaşık işlemler yapabildikçe, daha farklı ve yeni matematiksel formüller de sinyal işleme konusuna uyarlanmıştır. Makine öğrenmesi gibi terimler literatürce kabul görmüş ve bazı birleştirme yöntemleri geliştirilmiştir. Birleşik öğrenme örüntü tanımada daha iyi sonuçlar almak için farklı sınıflandırıcıların kombinasyonlarını kullanan bir birleştirme metodudur. Özetle, birleşik öğrenme birçok sınıflandırıcının birlikte çalışabildiği bir yöntemdir. Bu çalışmada paroksismal atriyal fibrilasyon (PAF) tarama amacıyla, kişinin ektopiksiz elektrokardiyogram (EKG) kayıtlarına göre PAF hastası olup olmadığını bulmak için birleşik öğrenme kullanılmıştır. Kullanılan veri kümesi PAF hastası ve PAF hastası olmayan kişilerin ECG kayıtlarından oluşmaktadır. Hem hiyerarşik hem de paralel yapıda birleşik öğrenme yapıları denenmiştir. Uzmanlar eğitilirken k katlamalı çapraz doğrulama ve *bootstrap* örnekleme metotları kullanılmış ve performansları karşılaştırılmıştır. En başarılı sonuçlar hiyerarşik yapı ile elde edilmiştir.

Anahtar Sözcükler: Sinyal işleme, birleşik öğrenme, komite makinaları, k katlamalı çapraz doğrulama, PAF tanıma

CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

1.1 Objective of the Thesis

Atrial fibrillation (AF) is an irregular heartbeat (arrhythmia) that can lead to stroke, heart failure and other heart-related symptoms. Atrial fibrillation is the most common sustained disorder of cardiac rhythm, which is often associated with a high risk of morbidity and mortality from heart failure, stroke and thromboembolic complications. According to the duration of episodes, types of AF are categorized as paroxysmal AF, persistent AF, longstanding persistent AF, permanent AF, nonvalvular AF (January et. al., 2014).

Paroxysmal AF (PAF) is an episode of uncoordinated movement of the atria that occurs occasionally and then stops. Episodes can last as short as minutes before returning to normal (sinus) rhythm (Marcin, 2013). The PAF usually leads to persistent AF in several years. Therefore, it is important to diagnose it in a person with PAF. It is easily done by recording the electrocardiogram (ECG) of as person during a PAF episode. However, since the PAF episodes could be short, it is usually very hard to record the ECG during a PAF attack. This is due to the fact that the subject who has this arrhythmic event could not have sufficient time to go to a health clinic and get an ECG record. On the other hand, when the ECG of a PAF patient is recorded during non-episodic intervals, it is very hard to diagnose the disease from this normal sinus rhythm (NSR) ECG signal. Therefore, it would be very helpful to construct an automatic computer based system that would diagnose PAF disease from episode-free ECG records; i.e. from NSR ECG records.

There have been various studies conducted for the purpose of PAF detection from normal sinus rhythm (NSR) ECG records based on different features of ECG signal. Indeed, there was a challenge about PAF prediction held in 2001. The challenge was organized with the help of Physionet which is a databank for the researchers. And the participants tried to predict the PAF disease according to given datasets (train and test sets) (Moody et. al., 2001). According to that challenge, Schreier et. al. achieved 82% accuracy and that was the best accuracy. Their approach was based on premature P-waves and the accuracy rate of them indicated that abnormal P-waves may herald or even trigger PAF (Schreier et. al., 2001). Zong et.al. also participated that challenge and their accuracy rate was 80%. Firstly, they examined the ECG records visually, then they used a previously developed automated arrhythmia detection algorithm which identifies beat types (normal, atrial premature complex (APC), ventricular premature complex, etc.) as well as rhythm types. After examining the detected arrhythmia patterns, they found that the number and timing of the detected APCs appeared to be of significant value in terms of predicting imminent PAF episodes (Zong et. al., 2001). In years after that PAF prediction challenge, there were also studies conducted on that dataset. Ros et al. used dataset taken from Physionet PAF Prediction Challenge Database (AFPDB) and they used 22 parameters extracted from P-wave analysis (Ros et. al., 2004). In 2016, I. Hilavin worked on the same dataset, and she applied genetic algorithm on the features from obtained by analyzing heart rate variabilities (Hilavin, 2016). Another study in PAF screening, Martinez et. al obtained ECG records of 46 PAF patients and 53 healthy subjects. Then they tried to calculate the variability of P-wave. They found that using a decision tree with P-wave area and P-wave arc length achieved 95.42% accuracy to discriminate ECG segments of healthy subjects and patients suffering from PAF (Martinez et. al., 2012).

Recently new ways of pattern classification have been developed. Ensemble learning is one of these new ways. It is a method of using a combination of different experts (classifiers) to get better results in pattern classification tasks (Yu et. al., 2008; Breiman, 1996; Dietterich, 2000). Briefly, ensemble learning is a method whereby different classifiers work together. Ensemble is constructed as a structure with the aim of compensating for the errors provoked by single classifiers. In this thesis study, several PAF screening systems were constructed based on ensemble learning using MATLAB software.

There are both parallel and hierarchical ensemble structures in literature while signal processing (Polikar, 2006). In this thesis, several experiments with both types of these structures were conducted and their results were compared. Results of these classifiers were combined by a gating network which is based on either averaging or majority voting. The data used in this study consists of electrocardiogram (ECG) signals recorded from subjects; one group were previously diagnosed with PAF, other group consists of subjects without PAF disease.

The features used in this study consist of different heart rate variability (HRV) features such as mean RR, std RR, high frequency (HF) peak power, relative HF power extracted from the ECG records. HRV has been used extensively to assess autonomic control of the heart under various physiological and pathological conditions. The analysis of HRV is based on analysis of RR intervals which are series of time intervals between heartbeats. Various features have been used to analyze HRV. For example, a simple time domain analysis of HRV, such as the mean, standard deviation, and root mean square of successive RR interval differences have been widely employed in quantification of the overall variability of the heart rate. Frequency domain analysis of HRV and non-linear analysis of HRV can be also done (Lee et. al., 2008).

1.2 Organization of the Thesis

The thesis is organized as five chapters. Chapter 1 is the introduction part. In this part, the objective of this work and organization of the thesis were given. Some of the studies finding a place in literature and proposed works of previous studies were mentioned.

Chapter 2 is related to physiological background. There is mostly theoretical information about human body signals, electrocardiogram, PAF disease and some medical terms.

In Chapter 3, it is mentioned about methods used in this work. Structures of ensemble learning are presented. The classifiers and their algorithms are described. Flowcharts of performed experiments are presented. General processes are mentioned and performance evaluation is explained.

Chapter 4, the experiments conducted and the results obtained from these experiments are presented. These performance results are given in tables.

In Chapter 5, the results obtained in this study are discussed and a conclusion is made. Possible future works related to this study are also proposed.

CHAPTER TWO
PHYSIOLOGICAL BACKGROUND

2.1 Electrocardiogram (ECG)

The heart is a hollow muscular tube that consists of four chambers; two upper chambers called atria and two lower chambers called ventricles. These chambers are organized in a way that right atrium co-works with right ventricle to get CO_2 rich blood from body and pump it to lungs for the purpose of cleaning whereas left atrium cooperates with left ventricle to get O_2 rich blood from lungs and pump it to the body. As the blood leaves each chamber of the heart, it passes through a valve. The heart valves enable that blood flows in only one direction.

The output of the heart per minute (cardiac output) is the vital event required to maintain blood flow on the regular basis. In addition to blood volume and contractile strength, the heart must continue both relaxation and contraction to perform well. This system of perfection is based on a series of electrophysiological events occurred within the cardiac tissues that can be observed using a device, which is known as an electrocardiogram (ECG). An ECG signal describes heart’s electrical activity. As the heart continue to beat, these beats cause small voltage differences (Becker, 2006). These differences provide us some information such as heart rate, rhythm, and morphology. In general, ECG is recorded by attaching a set of electrodes on body surface such as chest, neck, arms, and legs.

A typical ECG wave consists of a P wave, a QRS complex, and a T wave. Figure 2.1 shows the basic shape of a healthy ECG signal. The P wave reflects the sequential depolarization of the right and left atria. It usually has positive polarity, and its duration is less than 120 milliseconds. The spectral characteristic of a normal P wave is usually considered to be low frequency, below 10–15 Hz. The QRS complex corresponds to depolarization of the right and left ventricles. It lasts for about 70– 110 milliseconds in a normal heartbeat, and has the largest amplitude of the ECG waveforms.

Due to its steep slopes, the frequency content of the QRS complex is considerably higher than that of the other ECG waves, and is mostly concentrated in the interval of 10–40 Hz. The T wave reflects ventricular repolarization and extends about 300 milliseconds after the QRS complex. The position of the T wave is strongly dependent on heart rate, becoming narrower and closer to the QRS complex at rapid rates (Wang et. al., 2008).

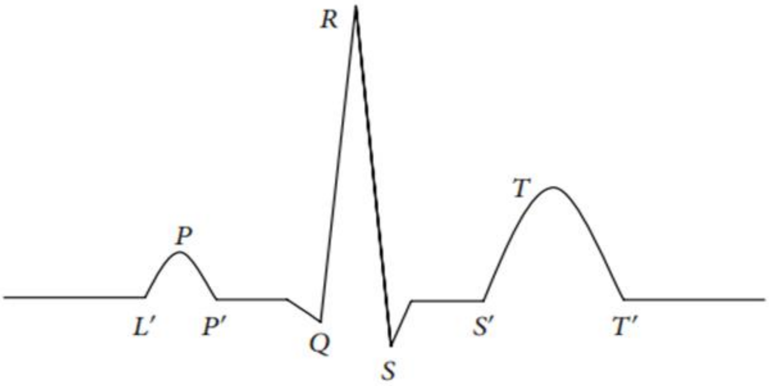
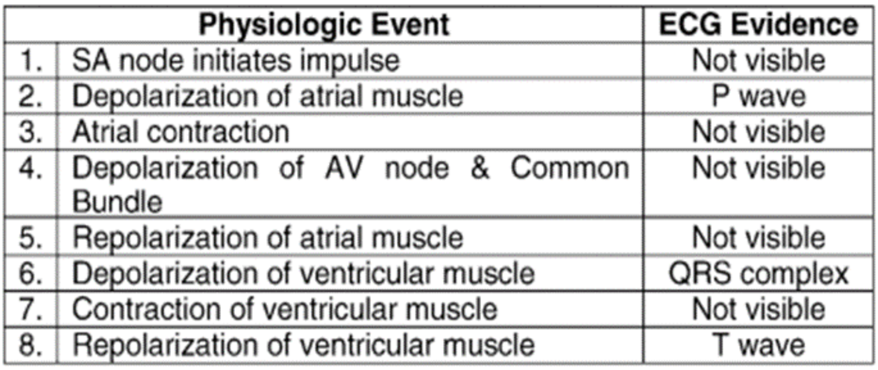


Figure 2.1 Basic shape of an ECG heartbeat signal (Wang et. al., 2008)

Table 2.1 Summary of events of a cardiac cycle (Becker, 2006)



	Physiologic Event	ECG Evidence
1.	SA node initiates impulse	Not visible
2.	Depolarization of atrial muscle	P wave
3.	Atrial contraction	Not visible
4.	Depolarization of AV node & Common Bundle	Not visible
5.	Repolarization of atrial muscle	Not visible
6.	Depolarization of ventricular muscle	QRS complex
7.	Contraction of ventricular muscle	Not visible
8.	Repolarization of ventricular muscle	T wave

Electrocardiogram (ECG) which gives tips about the rhythm and function of the heart is an important guide for cardiologists to diagnose several heart diseases. A medical doctor may fail to diagnose the arrhythmias due to the dynamic nature of ECG signals. A doctor could interpret an ECG signal based on its morphological shape and other parameters such as RR interval, PP interval, and QT interval. The task of determining fiducial points and computation of parameter is a tedious job for doctors. Hence, there is a need for computer aided diagnosis system which can achieve a higher recognition accuracy (Thomas et. al., 2015).

2.2 Arrhythmias & Paroxysmal Atrial Fibrillation (PAF)

Arrhythmia can be defined as the any deviation of heart's rhythm from normal operation. The result of arrhythmias may change from nothing to death. During an arrhythmia, the heart can beat too fast, too slow or irregularly. Arrhythmias can be classified according to the underlying mechanism or the origin of the arrhythmia. Three underlying mechanism of arrhythmias are abnormal impulse initiation, abnormalities of impulse propagation and combination of both (Gertsch, 2003). Arrhythmias can also be identified according to where they occur in the heart as supraventricular or ventricular arrhythmias. Supraventricular arrhythmias include arrhythmias caused by atrial tissue.

Atrial fibrillation (AF) is an irregular heartbeat (arrhythmia) that can cause to stroke, heart failure and other heart-related symptoms. Atrial fibrillation is the commonest sustained disorder of cardiac rhythm, which is often associated with a high risk of morbidity and mortality from heart failure, stroke and thromboembolic complications. Normally, the heart contracts and relaxes regularly. In atrial fibrillation, the upper chambers of the heart (the atria) beat irregularly instead of beating effectively to carry blood into the ventricles. If a clot breaks off, enters the bloodstream and lodges in an artery leading to the brain, a stroke results. About 15–20 percent of people who have strokes have this heart arrhythmia (American Heart Association, 2017). And the most common symptom of atrial fibrillation is fatigue. Increasing age, hypertension, obesity, smoking are such factors which trigger AF.

This diagnose is easily done by recording the ECG of a person during a PAF episode. For example, it is known that the P wave reflects atrial depolarization. So, a distortion in P wave can be the onset of the PAF disease (Thong et.al., 2004). In Figure 2.2, an ECG record which owns to a healthy person can be seen while in Figure 2.3, an ECG record of a PAF patient is given. When it is looked in Figure 2.3, it is seen that there is no P wave. So, this subject can be labelled as PAF patient.

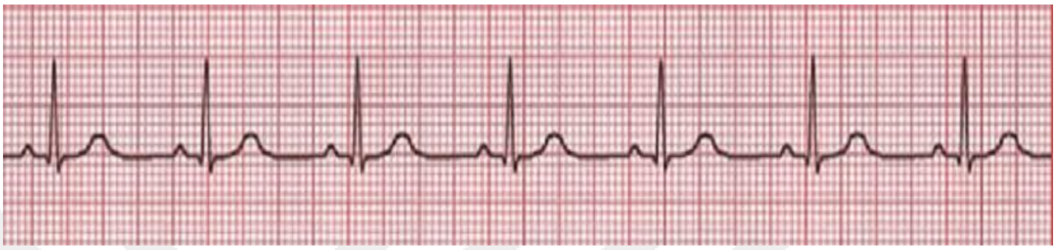


Figure 2.2 Normal (sinus) rhythm (EpoMedicine, 2016)

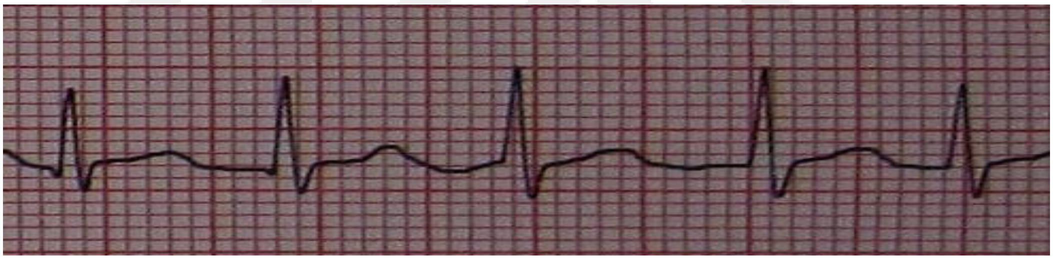


Figure 2.3 ECG record of a PAF person (Anumonwo et. al., 2014)

Although the detection of the PAF disease seems easy by looking at ECG records, it is very hard in real life. This is because PAF episodes could last very short. Thus, it is usually very hard to record the ECG signal during a PAF attack. When the ECG of a PAF patient is recorded during non-arrhythmic intervals, it is very hard to diagnose the disease. Therefore, it would be very helpful to construct an automatic computer based system that would diagnose PAF disease from episode-free ECG records.

CHAPTER THREE
METHODS

3.1 HRV Analysis and HRV Features

Heart rate variability analysis (HRV) is generally used for evaluating the effect of autonomic nervous system (ANS) on the functioning of the cardiovascular system. This effect occurs by adaptively changing the heart rate by regulation of the sinoatrial (SA) node. ANS is divided into sympathetic and parasympathetic branches and their influences on heart rate (HR) and HRV are quite well understood. Roughly speaking, sympathetic activity tends to increase HR and decrease HRV, whereas parasympathetic affects in the other way (Tarvainen et. al., 2014; Berntson et. al. 1997).

HRV is the evaluation of the fluctuations in the time intervals between heart beats, known as RR intervals. The importance of HRV is that it can reveal information about the autonomic nervous function, sympathetic-parasympathetic balance and cardiovascular health (Berntson et. al., 1997; Camm et. al., 1996; Malik et. al., 1996)**.**

HRV analysis can be done based on time domain, frequency domain and non-linear methods. Time domain analysis is the statistical examination of the fluctuations in RR intervals and commonly used because of their easy calculation. Statistical analysis and geometrical analysis are two branches of time domain HRV analysis. Geometrical methods require long-term RR intervals. And short- term RR intervals are used in this thesis. The features extracted from time domain value are mean RR intervals (mean RR), standard deviation of RR intervals (std RR), standard deviation of instantaneous heart rate, root mean square of successive differences between RR intervals etc. (Malik et. al., 1995). Mean RR and std RR are used as features in this work. Frequency domain methods decompose the total variation of the RR interval series into different frequency components, which can be considered as markers of different physiological effects (Berntson et. al., 1997; Nattel et. al., 2014). Heart has a nonlinear nature and analyzing the nonlinear properties of the RR intervals may reveal some information about the complex and nonlinear nature of these physiological mechanisms (Huikuri et. al., 2003).

Data used in this study were selected heart rate variability (HRV) features taken from Hilavin's Ph.D. thesis (Hilavin, 2016). These HRV features were obtained from ECG records used for PAF Screening & PAF Detection Challenges provided by PhysioBank, which is a comprehensive archive of well-grouped digital recordings of physiological signals for use by the biomedical researchers. It currently includes databases of multiparameter cardiopulmonary, neural, and other biomedical signals from healthy subjects and from patients with a variety of conditions with major public health implications, including life-threatening arrhythmias, congestive heart failure, sleep apnea, neurological disorders, and aging. PhysioNet is an on-line forum for the dissemination and exchange of recorded biomedical signals and open-source software for analyzing them. It provides facilities for the cooperative analysis of data and the evaluation of proposed new algorithms (Goldberger et. al., 2000).

Data used in the study consists of the following eight HRV features selected by a genetic algorithm out of 33 HRV features (Hilavin, 2016).

Mean RR (s): Arithmetic mean value of all RR intervals. For a series with length N, the mean is calculated as:

$$\mathrm{Mean\ RR} = \overline{\mathrm{RR}} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{RR}_{\mathrm{i}} \qquad (3.1)$$

Std RR (s): Standard deviation of RR intervals which reflect overall variation and defined as:

$$\mathrm{SDRR} = \sqrt{\frac{1}{\mathrm{N}-1}\sum_{\mathrm{j}=1}^{\mathrm{N}}\left(\mathrm{RR}_{\mathrm{j}} - \overline{\mathrm{RR}}\right)^2} \tag{3.2}$$

HF peak (Hz): HF band peak frequency. This band shows parasympathetic activity and is frequently called the respiratory band. During inhalation, the cardiorespiratory center inhibits vagal outflow, resulting in speeding up heart rate. Conversely, during exhalation, vagal outflow is restored, resulting in slowing heart rate. The magnitude of the oscillation is variable, but in healthy people, it can be increased by slow, deep breathing (HearthMath Institute Research Staff, 1993).

HF power prc (%): Relative power of HF band.

$$\text{HF power prc} = \frac{\text{HF power}}{\text{Total power}}\text{x}100\% \tag{3.3}$$

LF power prc (%): Relative power of LF band:

$$\text{LF power prc} = \frac{\text{LF power}}{\text{Total power}}\text{x}100\% \tag{3.4}$$

A typical heart rate variability records can be seen in Figure 3.1. That record was obtained from a healthy person during 15 minutes in resting conditions. (HearthMath Institute Research Staff, 1993).

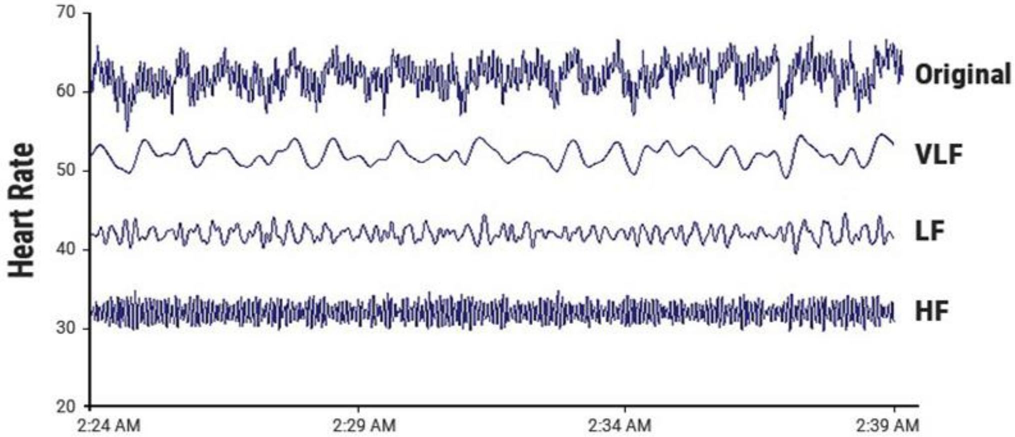


Figure 3.1 Typical HRV record (HearthMath Institute Research Staff, 1993)

Sample Entropy (SampEn): Approximate entropy(ApEn) is a method used to describe regularity in a time series such as heart rate time series. SampEn is precisely the negative natural logarithm of the conditional probability that two sequences similar for *m* points remain similar at the next point, where self-matches are not included in calculating the probability. Thus a lower value of SampEn also indicates more self-similarity in the time series. In addition to eliminating self-matches, the SampEn algorithm is simpler than the ApEn algorithm, requiring approximately one-half as much time to calculate. SampEn is largely independent of record length and displays relative consistency under circumstances where ApEn does not (Richman et. al., 2000).

Detrended Fluctuation Analysis (DFA): Detrended fluctuation analysis (DFA), which is a well-established method for the detection of long-range correlations in time series. detrended fluctuation analysis (DFA) has been established as an important tool for the detection of long-range (auto-) correlations in time series with non-stationarities (Kantelhardt et. al., 2001).

SD1 (Standard Deviation 1): Poincaré Heart Rate Variability (HRV) plot is a graph in which each RR interval is plotted against next RR interval (a type of delay map). SD1 is standard deviation of points perpendicular to the axis of line of identity. SD2 is standard deviation of points along the axis of line of identity.

The axis of both standard deviations, SD1 and SD2, can be seen in the Poincaré plot which is given in Figure 3.2.

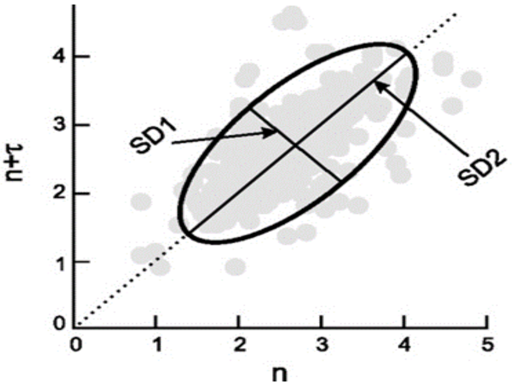
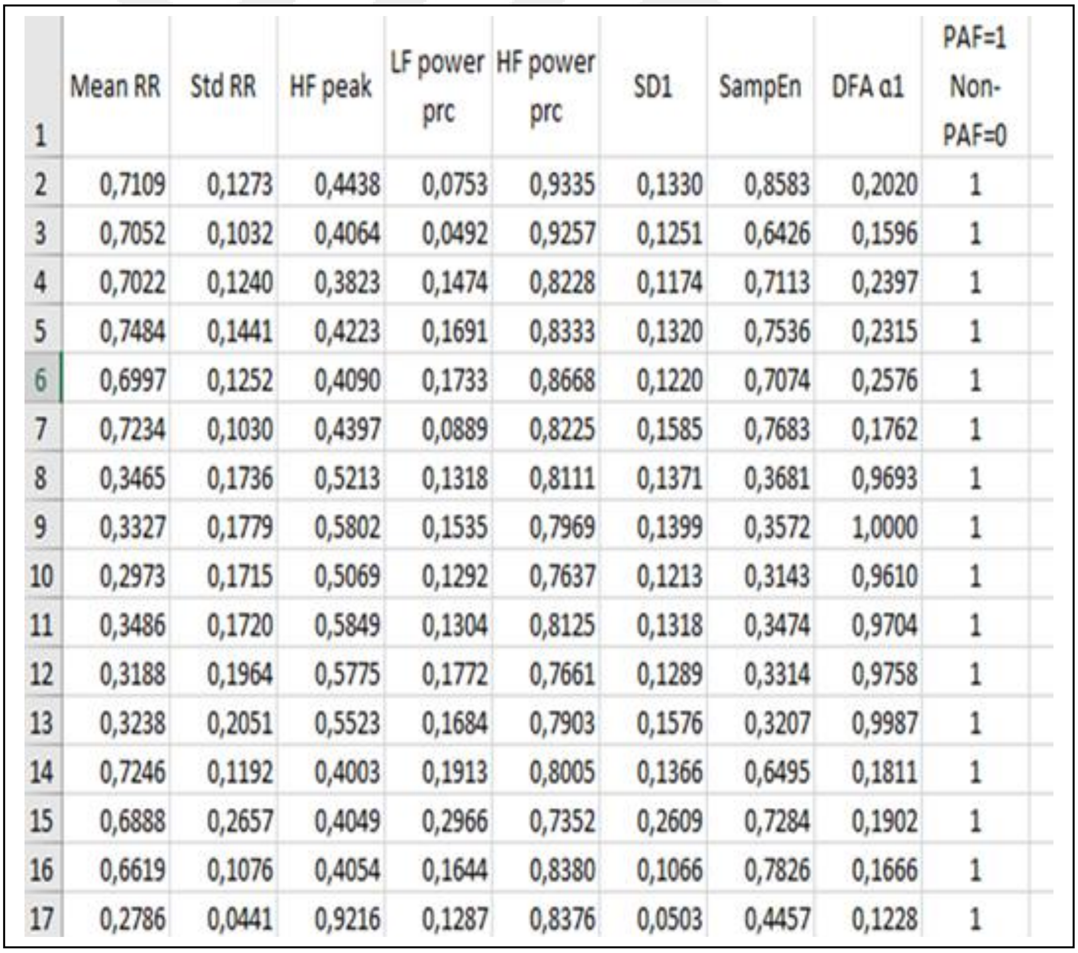


Figure 3.2 Poincaré Heart Rate Variability (HRV) plot



1	Mean RR	Std RR	HF peak	LF power prc	HF power prc	SD1	SampEn	DFA α1	PAF=1 Non-PAF=0
2	0,7109	0,1273	0,4438	0,0753	0,9335	0,1330	0,8583	0,2020	1
3	0,7052	0,1032	0,4064	0,0492	0,9257	0,1251	0,6426	0,1596	1
4	0,7022	0,1240	0,3823	0,1474	0,8228	0,1174	0,7113	0,2397	1
5	0,7484	0,1441	0,4223	0,1691	0,8333	0,1320	0,7536	0,2315	1
6	0,6997	0,1252	0,4090	0,1733	0,8668	0,1220	0,7074	0,2576	1
7	0,7234	0,1030	0,4397	0,0889	0,8225	0,1585	0,7683	0,1762	1
8	0,3465	0,1736	0,5213	0,1318	0,8111	0,1371	0,3681	0,9693	1
9	0,3327	0,1779	0,5802	0,1535	0,7969	0,1399	0,3572	1,0000	1
10	0,2973	0,1715	0,5069	0,1292	0,7637	0,1213	0,3143	0,9610	1
11	0,3486	0,1720	0,5849	0,1304	0,8125	0,1318	0,3474	0,9704	1
12	0,3188	0,1964	0,5775	0,1772	0,7661	0,1289	0,3314	0,9758	1
13	0,3238	0,2051	0,5523	0,1684	0,7903	0,1576	0,3207	0,9987	1
14	0,7246	0,1192	0,4003	0,1913	0,8005	0,1366	0,6495	0,1811	1
15	0,6888	0,2657	0,4049	0,2966	0,7352	0,2609	0,7284	0,1902	1
16	0,6619	0,1076	0,4054	0,1644	0,8380	0,1066	0,7826	0,1666	1
17	0,2786	0,0441	0,9216	0,1287	0,8376	0,0503	0,4457	0,1228	1

Figure 3.3 A partition of the data

Dimensions of the data are 800x8. It means there are 800 people and 8 features. 290 of the people are PAF patient, remaining 510 are non-PAF. A section of the data which includes the features used in the study can be seen in Figure 3.3.

3.2 Ensemble Learning

Ensemble learning is a method whereby more accurate predictions can be obtained. Main idea underlying the ensemble learning is working with more than one classifier and increasing the performance of the system. Normally, an ensemble structure consists of classifiers (experts) and a combiner. Among the ensemble classifiers, neural networks (NNs), support vector machines, fuzzy systems are mostly used in pattern recognition. These methods have different advantages and disadvantages with respect to each other in solving various problems (Duda et. al, 1973). K-means, kNN algorithm, artificial neural network, Naive Bayes algorithms and support vector machines (SVM) were the classifiers used in the study.

Both experiment-based studies and specific machine learning applications prove that although a given classification method could outperform all others for a particular problem or for a specific subset of the input data, it is not possible to find a single method achieving the best results on the overall problem domain. As for the advantages of ensemble learning, it provides better overall performance, could reuse existing pattern classification expertise, has heterogeneity, etc.. Heterogeneity supplies some advantages that expert classifiers do not need to be of the same type and different features can be used for different classifiers. There are also some disadvantages of ensemble learning. Most important deficiency is that the method needs high computational process (Dietterich, 2000). As a consequence, it can be said that the ensemble tries to improve the accuracy and the reliability of the overall classification system. Moreover, if a classifier fails, the system can compensate for this error (Kotsiantis, 2011). This idea, compensating failure of a classifier, forms the basis of the ensemble learning.

When an ensemble is constructed, there are some requirements that should be satisfied to improve the success rate. First, the ensemble members should be diverse or complementary (Yu et. al., 2008). Diversity can be described as different reactions given by different classifiers when the input changes. Second, the classifiers should be independent from each other. Benefit of these two criteria is that uncorrelated errors of individual classifiers can be compensated by the combined effect. According to Breiman’s work, the most significant element is the uncertainty of the prediction method. If a small distortion of the learning set can cause significant changes in the predictor constructed, then ensemble model can improve the accuracy because each new learning set has created a new hypothesis. In addition, bootstrap replicates as much as created hypotheses are needed to train the experts. As the dimension of data grows, the need for bootstrap replicates also increases (Breiman, 1996).

There are two types of constructing ensemble systems: static and dynamic (adaptive). In static combination, ensemble structures can be designed as hierarchical or parallel. These methods were mentioned in T. Dietterich’s work under the headings such as manipulating the training examples, manipulating the input features, and manipulating the output targets (Dietterich, 2000). The ensemble is constructed as a structure to combine highly accurate classifiers instead of the less accurate ones. T. Dietterich also mentioned that using local classifiers could be interesting (Dietterich, 2000).

There are different ensemble structures in literature (Avnimelech et. al., 1999; Zheng et. al.,2010). Some of them are constructed as parallel and hierarchical structures. In the hierarchical combination, fast and straightforward sub-problems are preferred at first step, then other sub–problems which need more intensive calculations are tried to be resolved in later stages. Classifiers could be placed in a sequential or tree-structured shape. These structures of ensemble learning could be seen in Figure 3.4 and Figure 3.5. But not all the structures of ensemble learning have to follow this hierarchy. There are also studies about ensemble learning designed as parallel structure.

Much of ensemble learning in the literature fall into this type of ensemble learning (Polikar, 2006). Parallel structure of ensemble learning can be seen in Figure 3.6.

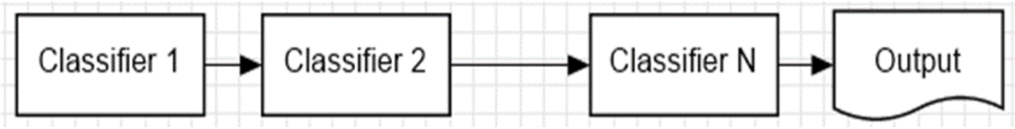


Figure 3.4 Hierarchical structure of ensemble learning.

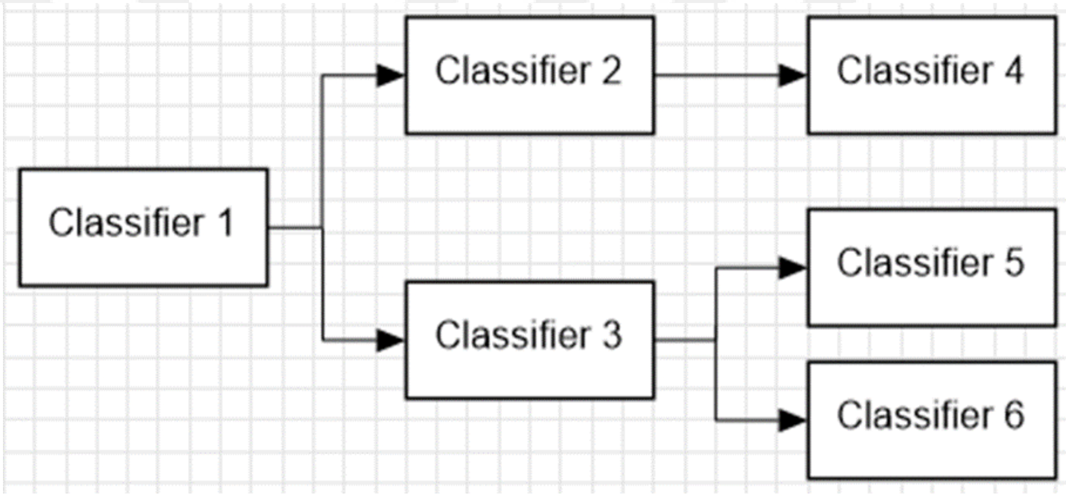


Figure 3.5 Hierarchical structure of ensemble learning tree trend

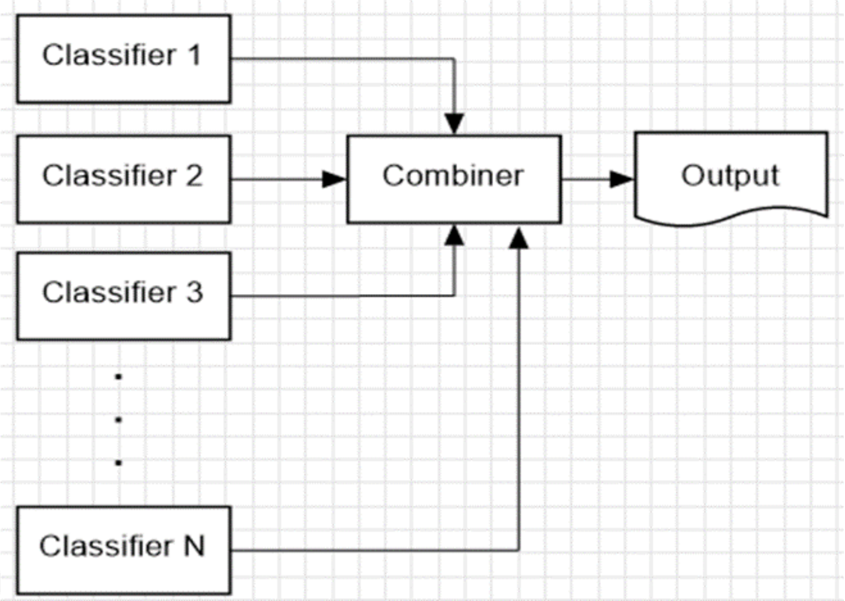


Figure 3.6 Parallel structure of ensemble learning

In much of the studies about ensemble learning, combining process is called as gating. Gating networks has the same mission with the combiner seen in Figure 3.6 and processes done by gating network could be voting, weighting or averaging. Some of these networks can be seen in Figure 3.7.

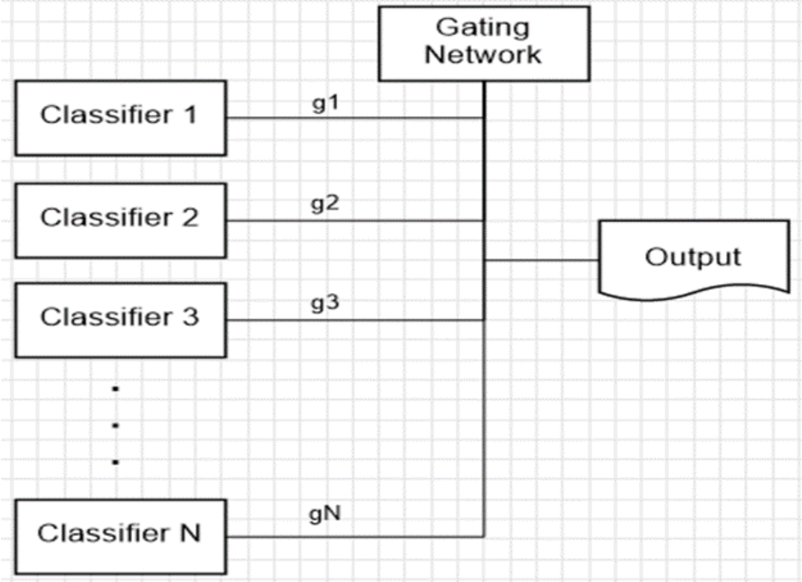


Figure 3.7 Gating network

In a voting network, each classifier has one vote. Then final decision is given based on the majority of votes. In a weighting network, different classifiers have different weight values. When an input is given to the system, a classifier tries to classify this data and the output of the classifier is multiplied by its weight value. The outputs of the classifiers are then summed together. When the final decision is to be made, it is looked whether summed value is greater than the threshold value or not. Deciding a suitable threshold value is another process.

Before deciding which gating type will be used, it can be more fundamental to look at training examples. In this study several cross validation methods were used to separate the training and test examples from the entire data. One of these cross validation methods was bootstrap sampling. There was no need to use bootstrap replicate as much as in Breiman's study, because there was enough data to train classifiers. But the number of subspaces still can be increased. K fold cross validation was also tried in addition to bootstraps. Different experiments were carried out to decide which subsampling or validation method is more successful to train the experts. Details about these experiments will be given in next chapters.

When an ensemble is constructed, using uncorrelated classifiers increase the success of the system. As the correlated classifiers are used, same mistakes can be repeated by different classifiers. There are similarities between ensemble learning method and phenomena of daily life. For example, before deciding to enter a serious surgery we want to diagnosed by a few doctors. And according to the comments of the doctors, ultimate decision is given.

When the final decision of each expert is obtained, gating or combining processes must be done. Majority of votes and averaging are among the most widely used combination methods. According to Polikar, it is important to point out two issues here: first, in the context of ensemble systems, there are many ways of combining ensemble members, of which averaging the classifier outputs is only one method. Second, combining the classifier outputs does not necessarily lead to a classification performance that is guaranteed to be better than the best classifier in the ensemble.

Rather, it reduces our likelihood of choosing a classifier with a poor performance. After all, if we knew a priori which classifier would perform the best, we would only use that classifier and would not need to use an ensemble (Polikar, 2012).

Bagging and Boosting algorithms are the most used and the most known ensemble algorithms. Both of these algorithms fall under the headings of manipulating the training examples mentioned in T. Dietterich's work (Dietterich, 2000). It is known that bagging can significantly reduce the variance, and therefore, it is better to be applied to learners who suffered from large variance, e.g., unstable learners such as decision trees or neural networks. Boosting can considerably reduce the bias in addition to depressing the variance, and hence, on weak learners, boosting is generally more effective (Stan et. al., 2015).

3.3 Classifiers

There are different types of classifiers used as experts in ensemble structures. These experts are k-means algorithm, kNN algorithm, artificial neural network, Naive Bayes algorithm and support vector machines.

3.3.1 k- means Algorithm

Lloyd's algorithm, often referred as k-means algorithm, is the simplest and most commonly used classifier. This algorithm starts with choosing *k* centers, centroid, randomly, these centroids are generally chosen uniformly according to features of data, mostly means of data which share same attributions. Each data are assigned to the nearest centroid, then each centroid is recalculated with assigned new data. These two steps are repeated until prediction rule met. The k-means algorithm is the most widely used partitional classifying algorithm. Its popularity can be associate with several reasons. First, it is conceptually simple and easy to implement. Nearly all data mining software includes an implementation of it. Second, it is versatile, i.e., almost every aspect of the algorithm (initialization, distance function, termination criterion, etc.) can be modified.

Finally, it is independent that how distribution of data is spread in space. On the other hand, k-means has several significant disadvantages. First, it requires the number of clusters, k, to be specified a priori. The value of this parameter can be determined automatically by means of various cluster validity measures (Celebi et. al., 2013).

3.3.2 K nearest neighbor Algorithm

K-nearest-neighbor (kNN) classification is one of the most fundamental and simple classification methods and should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data. K-nearest-neighbor classification was developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine. In an unpublished US Air Force School of Aviation Medicine report in 1951, Fix and Hodges (Fix et. al., 1989) introduced a non-parametric method for pattern classification that has known the k-nearest neighbor rule (Peterson, 2009).

In the classification or discrimination problem with two populations, denoted by X and Y, one wishes to classify an observation z to either X or Y using only training data. The kth-nearest neighbor classification rule is based on simple rule. If neighborhood between z and X is greater than the neighborhood between z and Y, z is assigned to the class X, otherwise it assigned to the class Y. The first study of this method was undertaken by Fix and Hodges,1951 (Fix et. al., 1989). Since then there have been many studies into the method’s statistical properties and optimal choice of k (Hall et. al., 2008).

3.3.3 Artificial Neural Network

An artificial neural network (ANN) is a group of simple units with adjustable weight. ANNs are inspired from human neurological system and are composed of neuronlike units connected together through input and output paths that have adjustable weight. Each neuron takes an input and produces an output signal proportional to its weight.

$$y_i = f\left(\sum_{i=1}^{N} x_i w_i\right) \tag{3.5}$$

where x_i is the input, w_i are the weights, *f(.)* is the activation function and y_i is the output of the i^{th} unit. Different functions can be applied as an activation function but mostly a sigmoid function is used.

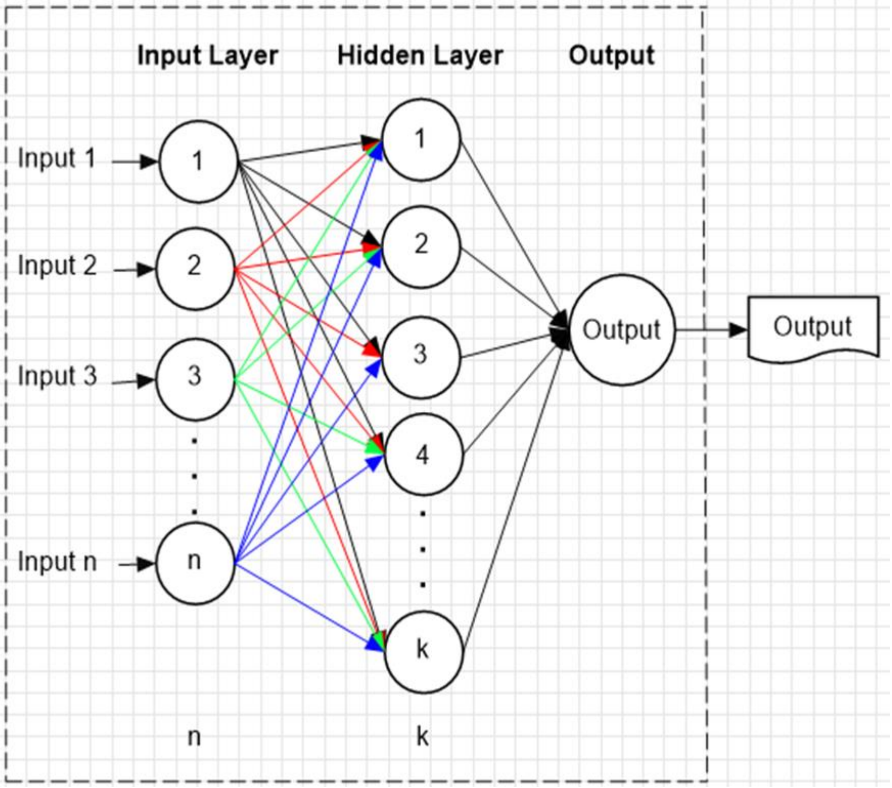


Figure 3.8 Basic scheme of a neural network

Multi-layer perceptrons (MLPs) are the mostly used ANN structures. As the name implies, a MLP consists of successive layers, each of which includes a different number of processing units. The units in the first layer receive inputs from the outside and are fully connected to units in the hidden layer. The units in the hidden layer are connected to output layer units. The units in the output layer produce an output.

A training phase for ANN means that the values of the connection weights are adjusted. Then the network can produce a correct output for new data entered from outside. The proper weights are determined under the control of a training algorithm. When the best weight is adjusted, the network can be tested on the sample data.

There will be an error function, since entered new data are probably different from the training data. As the error rate decreases, it can be said that the network is well trained.

Commonly used training functions are summarized:

Gradient Descent with Adaptive Learning Rate (GDALR): In plain gradient descent, the learning rate is held fixed during the training phase. However, changing the learning rate during the training process is a method that could increase the performance of the network (Yu et. al., 2002).

Scale Conjugate Gradient (SCG): This train function seems to perform well both pattern recognition and function approximation problems. The train function is almost as fast as the Levenberg – Marquardt algorithm (trainlm) on function approximation problems.

Step size scaling mechanism is used which avoids a time consuming line search per learning iteration. This mechanism makes the algorithm faster than any other second order algorithms. The scale conjugate gradient (trainscg) function requires more iteration to converge than the other conjugate gradient algorithms, but the number of computations in each iteration is significantly reduced because no line search is performed.

Resilience backpropagation (RP): This training algorithm eliminates the effects of the magnitudes of the partial derivatives. In this sign of the derivative is used to determine the direction of the weight update and the magnitude of the derivative have no effect on the weight update. This function does not perform well on function approximation problems. Performance of the function decreases as the error goal is reduced. The memory requirements for this algorithm are a little bit smaller than other training functions.

Broyden – Fletcher – Goldfarb – Shanno Algorithm (BFG): This algorithm approximates Newton's method, a class of hill-climbing optimization techniques that seeks a stationary point of a function. For such problems, a necessary condition for optimality is that the gradient be zero. This algorithm requires more storage and computation than the conjugate gradient methods, but it converges in fewer iterations. BFGS have good performance even for non-smooth optimizations and an efficient training function for smaller networks.

Levenberg-Marquart (LM) Algorithm: Levenberg-Marquart (trainlm) function is a network training function that updates weight and bias values according to Levenberg-Marquardt optimization. Trainlm function is often the fastest backpropagation algorithm in the Matlab toolbox, and is highly recommended as a first-choice supervised algorithm, although it does require more memory than other algorithms.

There are no significant differences between the correct classification percentage for Scale Conjugate Gradient and Levenberg - Marquart function, and, they are in acceptable range. The convergence speed of Levenberg - Marquart and Scale Conjugate Gradient are higher than other training functions. Considering the sample size of input patterns, Levenberg - Marquart suits to larger data set. It converges in less number of iterations and in lesser time than the other training functions (Sharma et. al., 2014). In this study, more correct predictive and faster train function is important instead of memory efficient train functions. So, Levenberg - Marquart train function was used in the designed network.

In this study, neural networks were created to determine the people as PAF patient or non-PAF. According to this information about training and learning functions, Levenberg- Marquart Algorithm was selected to use. Created network was a type of multilayer which number of neurons in hidden layer was equal to ten. And it has two neurons in output layer. Because of determining the output layer as two, the output layer creates two values in this structure of neural network. The reason why the output layer has two neurons is to benefit from their weight value while averaging their results. Scheme of the network is given in Figure 3.9. Different parameters were tried. And results of this different parameters will be given in next chapters.

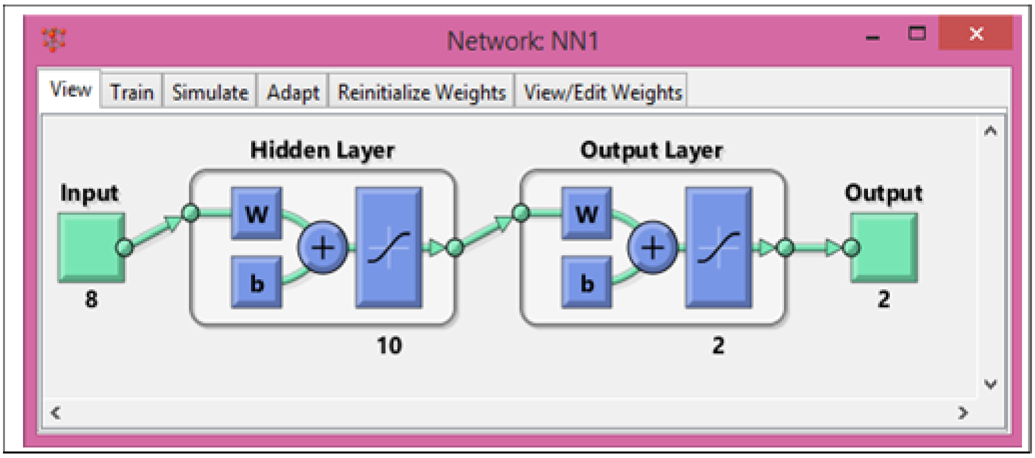


Figure 3.9 Flowchart of artificial neural network used in this study

It is shown that there are eight input, since dimension of data is equal to eight. Then there is a hidden layer which weights of network adjust. There is also an output layer which is the last layer before the output is created.

3.3.4 Naïve Bayes Algorithm

Naïve Bayes Algorithm has gained bad reputation (Lewis, 1998), and has earned the dubious distinction of placing near last in numerous classification papers (Yang et. al., 1999). But, it is frequently used because it is fast and easy to implement. More successful algorithms can tend to be slower and can need more computational cost. Bayesian classifiers assign the most likely class to a given example described by its feature vector. Learning such classifiers can be simplified by assuming that features are independent.

Basic formulation of probability is given in following formula 3.6.

$$\mathrm{P(X/C)} = \prod_{i=1}^{n} x_i \ x \ c \qquad (3.6)$$

Where x_i is a feature vector and c is a class. Despite this unrealistic assumption, the resulting classifier known as Naive Bayes is remarkably successful in practice, often competing with much more sophisticated techniques. Naive Bayes has been used effectively in many applications such as text classification, medical systems, and data mining.

The success of Naive Bayes in the presence of feature dependencies can be explained as follows: optimality in terms of classification error is not necessarily related to the quality of the fit to a probability distribution (i.e., the appropriateness of the independence assumption). Rather, an optimal classifier is obtained as long as both the actual and estimated distributions agree on the most-probable class. For example, it is proved that Naive Bayes optimality for some problems classes that have a high degree of feature dependencies, such as disjunctive and conjunctive concepts (Rish, 2001).

3.3.5 Support Vector Machines (SVM)

The algorithm, support vector machines (SVM), is based on the idea that two classes can always be separated from each other via a hyperplane. But, the problem is more than one hyperplane may be drawn to separate the classes. So, the goal in support vector machines is to find the separating hyperplane with the largest margin; as the margin is greaten, the generalization of the classifier becomes better (Bosher et. al., 1992; Cortes et. al., 1995). Support vector machines are helpful in text recognition, classification of images and biological applications (Chen et. al., 2001; Gaonkar, 2013; Cuingnet et. al., 2011).

The equations of this algorithm are derived from using vector properties. Let's consider a space where there are points which are wanted to be classified. And let's show these points as symbols of "+" and "-". Figure 3.10 is given as an example for that space.

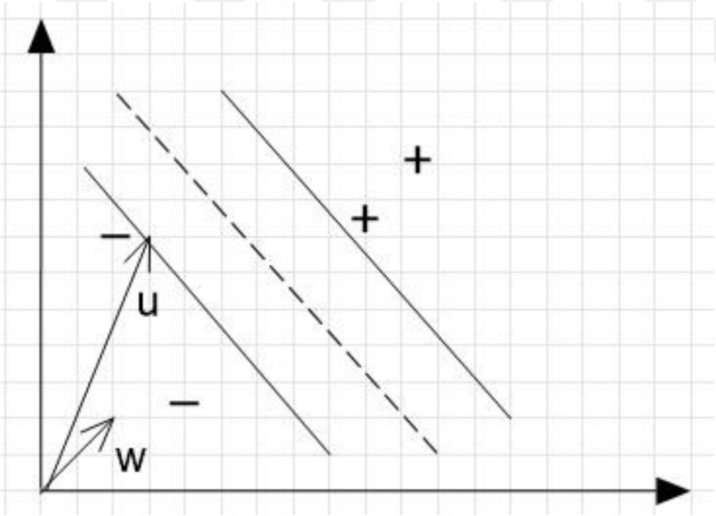


Figure 3.10 A space where there are two different classes

The vector, $\vec{w}$, is the vector which is chosen as perpendicular to the median of the dashed line. The vector, $\vec{u}$, is the vector which we do not know its place whether it is in positive side of the dashed line or vice versa. To find its place, projection of the unknown vector which is same direction with the vector $\vec{w}$ must be calculated. So the equation about place of vector $\vec{u}$ can be written as :

$$\vec{\mathrm{w}} \cdot \vec{\mathrm{u}} \geq c \tag{3.7}$$

In equation 3.7, $\vec{u}$ and $\vec{w}$ are vectors and c is a constant. That dot product in the equation gives us the projection onto $\vec{w}$. As the projection greatens, it can be said that the sample is positive. This equation can be arranged like the equation 3.8.

$$\vec{\mathrm{w}} \cdot \vec{\mathrm{u}} + b \geq 0 \tag{3.8}$$

The equation 3.8 gives us the decision rule for the positive samples where the b is a constant. The problem in equation 3.8 is that place of $\vec{\mathrm{u}}$ and $\vec{\mathrm{w}}$ are not known. $\vec{\mathrm{w}}$ has to be perpendicular to dashed line in Figure 3.10. But, there may be so many vector which can be drawn as perpendicular to that line. Because, length of that vector is not specified. For overcoming from this problem, let’s examine the unknown vector, $\vec{\mathrm{u}}$, first. Assume that $\vec{\mathrm{u}}$, is a known vector, it is a positive or negative sample and let’s show that vector with the symbol of $\overrightarrow{\mathrm{x}_+}$ and $\overrightarrow{\mathrm{x}_-}$. In this situation the equations can be given as equation 3.9 and 3.10:

$$\vec{\mathrm{w}} \cdot \overrightarrow{x_+} + b \geq 1 \tag{3.9}$$

$$\vec{\mathrm{w}} \cdot \overrightarrow{x_-} + b \leq -1 \tag{3.10}$$

The equation 3.9 is the new decision rule for the positive samples. It is greater than one because it is known that the sample is a positive sample. In equation 3.10, the equation is arranged as less than minus one or equal to minus one due to the same reason in equation 3.9. Some arrangements can make these equations more convenient mathematically. For doing this, let’s describe a variable y_i ,where y_i is +1 for the positive samples, -1 for the negative ones. So, the equations, 3.9 and 3.10, can be arranged.

$$\mathrm{y_i}(\vec{\mathrm{x}}_\mathrm{i}\vec{\mathrm{w}}+\mathrm{b})-1\geq 0 \tag{3.11}$$

The equation , 3.11, gives us the decision rule for the $\vec{\mathrm{x}}_\mathrm{i}$ where the $\vec{\mathrm{x}}_\mathrm{i}$ is in a gutter alongside of the dashed line in Figure 3.10. Decision rule for the samples is obtained. But, the width of the gutter is still unknown. So, another problem is about the width of the gutter. Because, in early stages of the topic, there was a sentence about margins (width of the gutter) which specifies the goal in support vector machines. When we look at Figure 3.11, it may give us an idea to understand about calculations of the width.

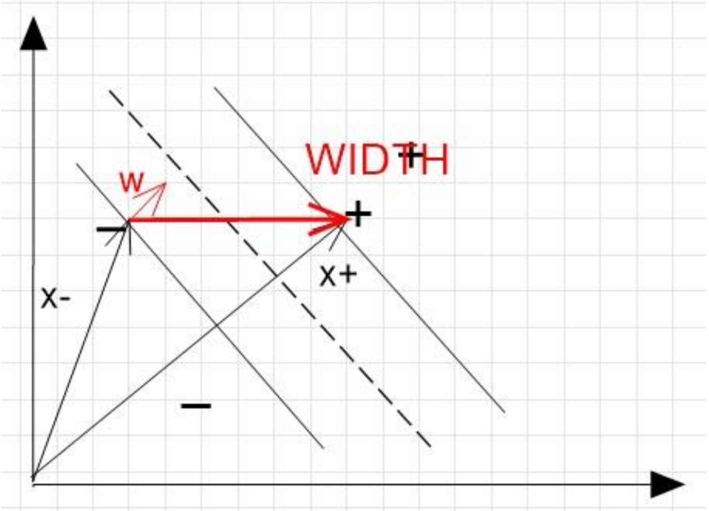


Figure 3.11 Vector representation of the width of the gutter

In Figure 3.11, there are three vectors. Two of them represent the vectors $\overrightarrow{x_+}$ and $\overrightarrow{x_-}$. And the other one, $\vec{w},$ is an unit vector which is perpendicular to dashed line. According to the Figure 3.11, an equation about the width can be extracted:

$$Width = (\vec{x}_{+} - \vec{x}_{-}) \cdot \frac{\vec{w}}{\|w\|} \tag{3.12}$$

When we look at the equation 3.12, the arrangements can be done. From equation 3.11, $\vec{x}_{+}.\vec{w}$ and $\vec{x}_{-}\overrightarrow{.w}$ can be obtained. When the equation is rearranged:

$$Width = \frac{2}{\|w\|} \tag{3.13}$$

According to main goal in support vector machines, it is wanted to maximize the width between lines which separate the classes. According to equation 3.13, a new equation can be written for maximizing the width:

$$MAX\,\frac{2}{\|w\|} \;\rightarrow\; MAX\,\frac{1}{\|w\|} \rightarrow MIN\,\|w\| \;\rightarrow\; MIN\,\frac{1}{2}\|w\|^{2} \tag{3.14}$$

According to the equation 3.11 and 3.14, it will be tried to find an extremum of a function. The method of Lagrange multipliers is a good method to solve this problem. Lagrange method comes up in the beginning of 1800s and this method named after the Italian mathematician Joseph-Louis Lagrange (Bertsekas, 1999). According to the Lagrange multipliers, following equation can be written:

$$L = \frac{1}{2}\|w\|^{2} - \sum \alpha_i [y_i(\vec{w} \cdot \vec{x}_i + b) - 1] \tag{3.15}$$

When it is tried to solve equation 3.15, following equations can be written:

$$\frac{dL}{dw} = \vec{w} - \sum \alpha_i y_i \vec{x}_i = 0 \tag{3.16}$$

From equation 3.16, $\vec{w}$ can be written as:

$$\vec{w} = \sum \alpha_i y_i \vec{x}_i \tag{3.17}$$

While interpreting the equation 3.17, this equation tells us that the vector, $\vec{w},$ is the linear summation of the some vectors included in the space.

$$\frac{dL}{db} = -\sum \alpha_i y_i = 0 \quad \rightarrow \quad \sum \alpha_i y_i = 0 \tag{3.18}$$

The vector, $\vec{w}$, can be written into equation 3.15. In this situation:

$$L = \frac{1}{2} \sum \alpha_i y_i \vec{x}_i \sum \alpha_j y_j \vec{x}_j - \sum \alpha_i y_i \vec{x}_i \left(\sum \alpha_j y_j \vec{x}_j \right) - \sum \alpha_i y_i b + \sum \alpha_i \tag{3.19}$$

And when Eq. 3.17 and Eq. 3.18 are written into Eq. 3.19, following equation is obtained:

$$L = \sum \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y i y_j x_i \cdot x_j \tag{3.19}$$

What equation 3.19 says is the optimization depends on the dot product of pairs of samples. So, these equations are quietly efficient in linearly separable sampling sets. But, in linearly inseparable examples there may be problems. For overcoming this, a transformation process is needed. The reason why there is another process can be understood better when we look in Figure 3.12.

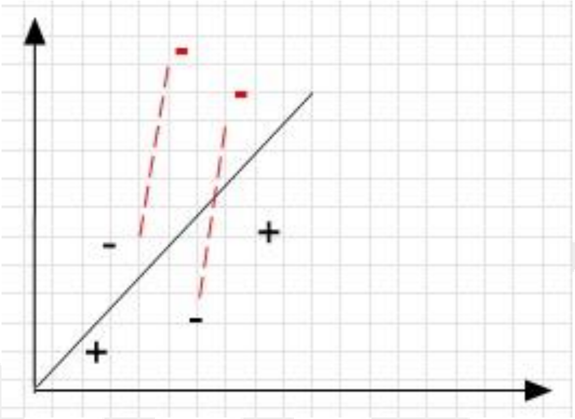


Figure 3.12 A space which consists of linearly inseparable samples

When it is looked in Figure 3.12, there is red dashed lines are the lines which show the transformation process. Some of the samples are transformed into another space, then classifying process continues. The black solid line is the line which classify the examples from each other. From the Eq. 3.19, it was found that optimization depends only on the dot products of two vectors. After the transformation process, dot product is also essential for us. Transformation process can be shown with the symbol of phi, "ϕ". And transformation function is described as *"k"*. In the last situation, the equation can be written as:

$$k(\vec{x_i}, \vec{x_j}) = \phi(\vec{x_i}) \cdot \phi(\vec{x_j}) \tag{3.20}$$

From equation 3.20, the kernel function, *“k”*, provides the dot product of two vectors in another space. And there is no need to know transformation function. Most popular kernel functions are linear kernel function and exponential function (Min et. al., 2005). In equations 3.21 and 3.22 these kernels can be seen (MIT OpenCourseWare, 2014).

$$(\vec{u} \cdot \vec{w} + 1)^{n} \tag{3.21}$$

$$e^{-\left(\frac{\|x_i - x_j\|}{\sigma}\right)} \tag{3.22}$$

3.4 Resampling Methods

Data was composed of 800 people records which some of them are patient and some of them not. Both k - Fold Cross Validation (k - Fold CV) and Bagging algorithm were used. According to McLachlan, 10-fold cross-validation is commonly used, but in general, *k* remains an unfixed parameter in k- fold CV (McLachlan, 2005). In bootstrap samplings, each sampling has a training rate with nearly 63%. So, minimizing the difference between training rates of both type of cross validation method will be fine for the experiments. If k is chosen as five instead of ten which is common, then the training rate will be 80%. Therefore, k is chosen as five in this study, while using k – Fold CV. Since k was chosen as five, creating five Bootstrap replicates for Bagging algorithm was determined. In general, at the time choosing the parameter, important point is that separated training set must have enough number of examples for training the classifier. Because, accurate results can not be expected from a classifier which is trained badly. After dividing by five, training data consist of 640 people, and the test data consists of 160 people. In each part which contains 160 people, there are 58 people with PAF disease, and 102 people who are healthy people. It was also created five bootstrap replicates to perform bagging algorithm. For bootstrap replicates, selection rate of them were 64.53%, 61.09%, 65.63%, 62.34%, 60% respectively.

3.4.1 k – Fold Cross Validation

In k – fold cross validation, data is divided into parts – *folds*. *k-1* folds are allocated as training data, and one different fold is specified as test data. This process continues until each fold is tested. Advantage of this method is that all observations are used for both training and test. Another advantage of this method is averaging the results of each iteration. In conventional method, there is one result obtained from test data. But, it is hard to determine this test data represents the whole data good. Or it is unclear the training examples are whether enough for training a classifier. By means of k fold cross validation, the challenges caused by conventional method may be fixed.

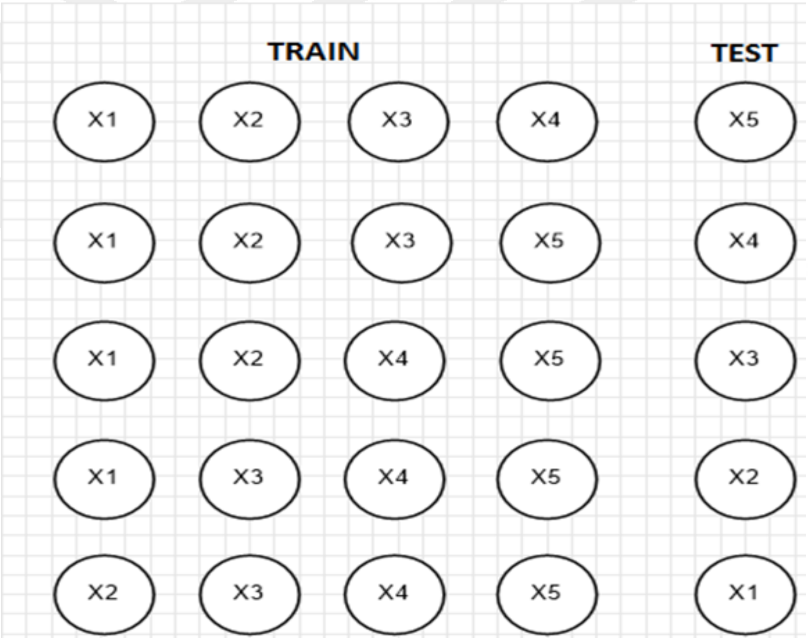


Figure 3.13 k – Fold Cross Validation Scheme. Where, X1, X2, X3, X4 and X5 are data whose dimensions are 160x8

In this work, four parts of data were separated for training and one another was separated for test. This test data was changed in every iteration. In each iteration, gold standardization values - specificity, selectivity, accuracy, positive predictive value, negative predictive value - were obtained. Then average of these values were determined as success of the system. In this work, the data were divided into five subsamples and one of them was reserved for testing. So that 80% of the data was allocated for training in each iteration.

3.4.2 Bootstrap Subsamplings

Another method to divide data into regions is Bootstrap. A "bootstrap" data set is one created by randomly selecting n points from the training set D, with replacement. (Since D itself contains n points, there is nearly always duplication of individual points in a bootstrap data set.) In bootstrap estimation, this selection process is independently repeated B times to yield B bootstrap data sets, which are treated as independent sets (Duda et. al, 1973).

For a given bootstrap sample, a randomly selected instance in the training set has probability $1-(1-1/m)^m$ of being selected at least once. For large m, this is about $1 - 1/e = 63.2\%$, which means that each bootstrap sample contains only about 63.2% unique instances from the training set. This perturbation causes different hypotheses to be built if the classifier is unstable (e.g., neural networks, decision trees) and the performance can improve if the results of unstable classifiers are good and not correlated; however, Bagging may slightly degrade the performance of stable algorithms (e.g., k-nearest neighbor) because effectively smaller training sets are used for training each classifier (Breiman, 1996). In this paper, both the classifiers ANN and kNN were used. So, there is a chance to see difference of these classifiers, when the input changes.

Table 3.1 An Example of Bootstrap Replicates Derived from Original Data

Original Data	1	2	3	4	5	6	7	8	9	10
Iteration - 1	1	3	7	2	3	8	8	7	5	6
Iteration - 2	6	6	3	4	2	10	8	9	10	6
Iteration - 3	4	3	7	7	9	1	3	10	5	4

As it is seen from the Table 3.1, randomly chosen data could be unordered or it may repeat. Five subsamples from the original data as bootstrap subsample were created in this study. And selection rate of each subsample were given in the beginning of the chapter.

Two validation methods were used to train artificial neural networks and classifiers. One of them, Bootstrap, had a selection rate approximately 63%. And the other one had the training data whose selection rate is 80% of the entire data set. As the test – training rate and dividing method changed, it is expected to get different percentage of success.

3.4.3 Bagging (Bootstrap Aggregation)

Bagging — a name derived from “bootstrap aggregation” — uses multiple versions of a training set, each created samples from D with replacement. Each of these bootstrap data sets is used to train a different component classifier and the final classification decision is based on the average or vote of each component classifier. Traditionally the component classifiers are of the same general form - i.e., all hidden Markov models, or all neural networks, or all decision trees - merely the final parameter values differ among them due to their different sets of training patterns (Duda et. al., 1973).

A classifier/learning algorithm combination is called unstable if small changes in the training data lead to relatively large changes in accuracy. In general, bagging improves recognition for unstable classifiers since it effectively averages over such discontinuities. The decision rule in bagging - mostly a simple vote among the component classifiers - is the most elementary method of pooling or integrating the outputs of the component classifiers (Breiman, 1996).

3.5 Performance Evaluation

Success of the experts were evaluated by the following measures: specificity, sensitivity, accuracy, positive predictive value (PPV) and negative predictive value (NPV). Before using these measures, the terms must be understood such as true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

True Positive (TP): If a person is predicted as patient and the person has the disease, this prediction is labelled as TP.

True Negative (TN): If a person is predicted as healthy and the person is not patient, this prediction is labelled as TN.

False Positive (FP): If a person is predicted as patient but the person has not the disease, this prediction is labelled as FP.

False Negative (FN): If a person is predicted as healthy, but the person is patient, this prediction is called as FN.

Table 3.2 Table of TP, TN, FP, FN (Confusion Matrix)

		Predicted Results	
		Patient (1)	**Healthy (0)**
Actual	**Patient (1)**	TP	FN
	Healthy (0)	FP	TN

These terms (TP, TN, FP, FN) are counted. After counting, the standards mentioned above specificity, selectivity, accuracy, PPV and NPV can be used. But first, let mention about these standards meaning.

Specifity: Specificity is the probability that a test will label as 'healthy' among people who have not the disease. Briefly, specificity indicates how well the test predicts one category. Formula of specifity is given as:

$$\textit{Specificity} = \frac{TN}{TN+FP}\; x\,100 \tag{3.7}$$

Sensitivity: Sensitivity is the probability that a test will label as 'patient' among people who have the disease. Sensitivity indicates how well the test predicts the other category which is not considered by specificity. Formula of sensitivity is given as:

$$\textit{Sensitivity} = \frac{TP}{TP+FN}\; x\,100 \tag{3.8}$$

Accuracy: Accuracy shows us how well the test predicts both category which consist of patient and healthy people. As the accuracy increase, we can say the test is well trained. Formula of accuracy:

$$\textit{Accuracy} = \frac{TN+\;TP}{TN+TP+FP+FN}\; x\,100 \tag{3.9}$$

Positive Predictive Value (PPV): It means the probability that the disease is present when the test is positive. It can be called correctness of positive values. Formula of PPV:

$$\textit{Positive Predictive Value} = \frac{TP}{TP + FP}\ x\ 100 \tag{3.10}$$

Negative Predictive Value (NPV): It means the probability that the disease is not present when the test is negative. It can be called correctness of negative values. Formula of NPV:

$$\textit{Negative Predictive Value} = \frac{TN}{TN + FN}\ x\ 100 \tag{3.11}$$

Matlab R2015b was used in this study. The matrix-based Matlab language is a software to apply computational mathematics. Compared to other programming languages, matrix computations and pattern recognition processes are easier in Matlab (Duin, 2000).

CHAPTER FOUR
RESULTS

In this thesis study, both hierarchical and parallel structures were used. The dataset was divided into two subsets: training and test datasets. The experts were trained on training sets to correctly classify whether the subject is PAF patient or not. Then, these experts were evaluated on the test data and their performances were obtained. In order to compare different ensembles, different experiments were carried out. While interpreting the overall success of the systems, the accuracy values were taken into account.

In parallel structures, only artificial neural networks were selected as the experts whereas, in hierarchical ones, different types of classifiers were used. To find the best parameters to use in experts, various tests were carried out based on both k-fold CV and bootstrap techniques. The results for the ANN classifier for different number of hidden neurons are given in Table 4.1 and Table 4.2. Based on these results, ten hidden neurons were selected for the ANN classifier.

Table 4.1 Results of ANN with different number of neurons obtained from k - Fold Cross Validation

	Number of neurons	**6**	**8**	**10**	**12**	**14**
ANN	Specificity	95.49	94.00	89.19	79.44	89.70
	Sensitivity	96.58	81.42	93.88	84.42	92.05
	Accuracy	83.52	88.01	**90.63**	88.60	89.75
	PPV	88.65	96.00	79.31	94.67	85.17
	NPV	91.30	93.99	97.06	94.01	92.85

Table 4.2 Results of ANN with different number of neurons obtained from Bootstrap subsamplings

ANN	**Number of neurons**	**6**	**8**	**10**	**12**	**14**
	Specificity	95.20	92.55	93.72	88.76	89.69
	Sensitivity	86.19	90.08	93.89	88.48	83.08
	Accuracy	92.43	91.34	**93.82**	92.10	92.15
	PPV	90.80	85.79	88.65	91.14	87.49
	NPV	86.51	84.99	96.68	89.69	88.39

4.1 Ensemble Experiments

4.1.1 Experiment 1

In the first experiment, a parallel structure of ensemble learning was designed. Artificial neural networks were used as the experts. Five different neural network structures were created for this experiment. Training data were obtained from bootstrap samplings. Since creating unique training sets are expected by the bootstrap samplings, different results could be obtained from each neural network expert. The results of each neural network expert were combined by both voting and averaging processes. When the results are evaluated, gold standardization values -true positive (TP), true negative (TN), false positive (FP), false negative (FN)- were calculated. According to these values, specificity, sensitivity accuracy, PPV and NPV values were determined. Flowchart of this experiment is given in Figure 4.1. The results of this experiment could be seen in Table 4.3.

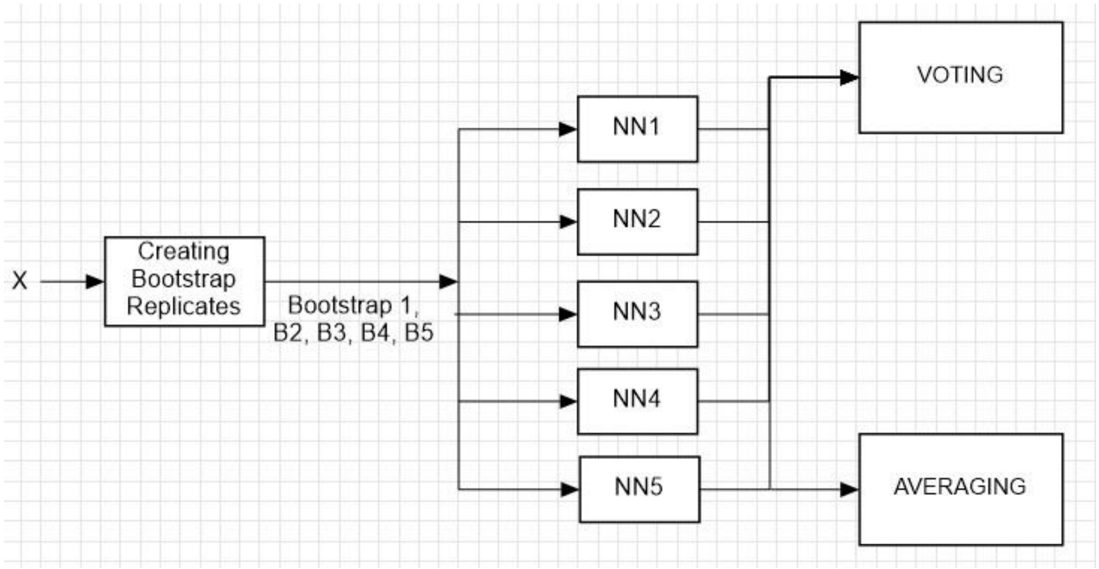


Figure 4.1 Flowchart of the Experiment 1

In Figure 4.1, X represents the data, and Bootstrap 1, B2, B3, B4, and B5 are the replicates derived from the original data via bootstrap subsampling. NN1, NN2, NN3, NN4, and NN5 represent the neural network experts.

Table 4.3 Results of Experiment 1

Bagging - ANN		
NN1	Specificity	88.24
	Sensitivity	79.65
	Accuracy	85.13
	PPV	81.03
	NPV	88.24
NN2	Specificity	88.04
	Sensitivity	81.95
	Accuracy	85.88
	PPV	80.35
	NPV	89.03
NN3	Specificity	81.59
	Sensitivity	85.83
	Accuracy	81.68
	PPV	76.54
	NPV	90.41

Table 4.3 continues

NN4	Specificity	89.81
	Sensitivity	78.97
	Accuracy	85.88
	PPV	81.86
	NPV	88.27
NN5	Specificity	86.67
	Sensitivity	78.97
	Accuracy	83.88
	PPV	83.58
	NPV	87.45
Averaging	Specificity	91.38
	Sensitivity	84.26
	Accuracy	88.75
	PPV	85.27
	NPV	91.04
Voting	Specificity	91.18
	Sensitivity	84.14
	Accuracy	88.63
	PPV	84.98
	NPV	91.02

According to the results of Table 4.3., NN2 and NN4 have the same accuracy rate with 85.88%. However, the accuracy rate of voting and averaging is greater than this rate. So, it can be said that there is an improvement in results when the ensemble structure is used.

When results of averaging and voting are compared, it is seen that there is no significant difference between them. Normally, the difference between averaging and voting is expected to be more considerable. But in this experiment, it is hard to see it. This may be caused by using artificial neural networks as the experts. The output layer in the neural network structure had two output values, one of them was closer to one and the other one was closer to zero. Briefly, summation of these two value was equal to one. These values can be explained as the probability of being one (patient) or being zero (healthy). But, in this experiment, these probabilities happened to be so close to 1 and 0 values so that no difference occurs between averaging and voting. In future works, different classifiers such as fuzzy c - means can be used and thus differences can be observed between these two techniques.

4.1.2 Experiment 2

In Experiment 2, the structure of the ensemble and the classifier are the same as in Experiment 1. Contrary to the first experiment, five-fold cross validation is used while training the networks. And different results are expected to come up. Then these results are combined with averaging or voting. The flowchart of this experiment is given in Figure 4.2.

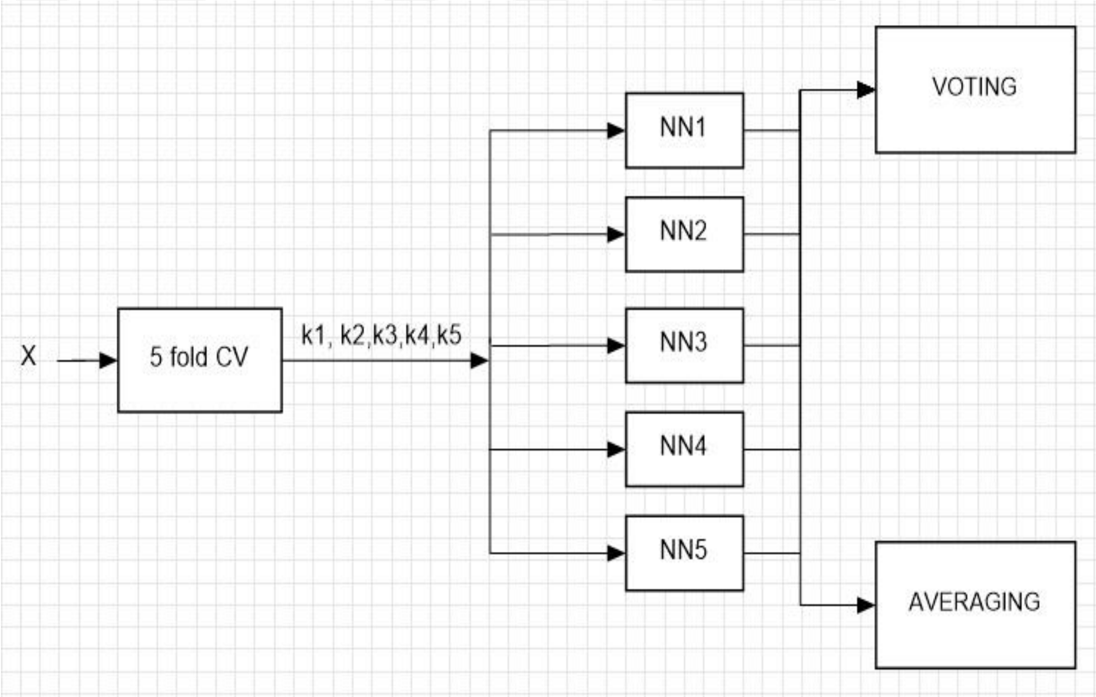


Figure 4.2 Flowchart of the Experiment 2

Table 4.4 Results of Experiment 2

k fold - ANN		
NN1	Specificity	87.65
	Sensitivity	85.17
	Accuracy	86.75
	PPV	82.48
	NPV	91.31
NN2	Specificity	86.06
	Sensitivity	87.93
	Accuracy	86.75
	PPV	81.46
	NPV	92.57
NN3	Specificity	88.04
	Sensitivity	89.66
	Accuracy	88.63
	PPV	85.48
	NPV	94.22
NN4	Specificity	86.28
	Sensitivity	86.90
	Accuracy	86.50
	PPV	79.63
	NPV	92.29
NN5	Specificity	85.47
	Sensitivity	87.64
	Accuracy	86.13
	PPV	80.74
	NPV	92.08
Averaging	Specificity	89.20
	Sensitivity	89.31
	Accuracy	**89.25**
	PPV	85.04
	NPV	93.69
Voting	Specificity	89.59
	Sensitivity	89.33
	Accuracy	**89.5**
	PPV	85.36
	NPV	93.70

As can be seen from Table 4.4, among the other neural networks NN3 has the best accuracy rate with 88.63%. Similar to the first experiment, accuracy rates of both voting and averaging are greater than single neural networks' performance in this experiment, too. However, contrary to first experiment, accuracy rate of voting is slightly greater than accuracy rate of averaging.

4.1.3 Experiment 3

In the first two experiments parallel structures of ensemble learning were used. Approaches to the ensembles are mostly under this category in literature. There are also studies done with the hierarchical structures in literature (Yu et. al., 2003; Yu et. al., 2009). In a hierarchical structure, each classifier produces its output according to a sequence. Efficient classifiers are used in first stages; then more accurate but complex classifiers are used in later stages.

Experiment 3 was designed as hierarchical structure. Main aim of this experiment was to divide the data into sub problems. Then these sub problems were tried to be classified with classifiers which could get best accuracy rate on that sub problem. Division process was done by using k means algorithm which is a clustering algorithm. From this aspect, this experiment can remind us divide and conquer method. Flowchart of this experiment is given in Figure 4.3.

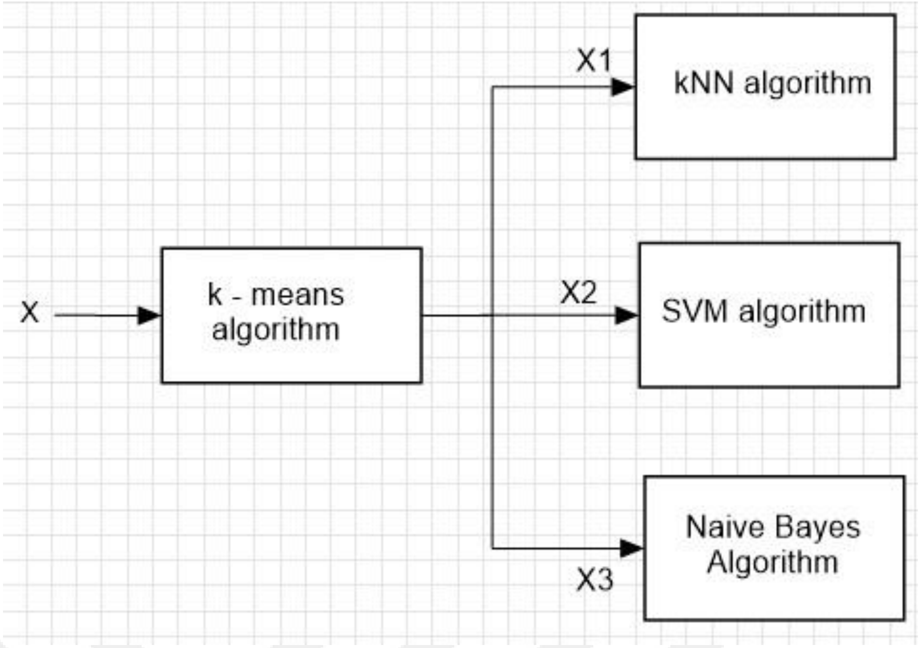


Figure 4.3 Flowchart of hierarchical structure of ensemble learning where X represents the entire data. X_1, X_2, X_3 is obtained from the X by dividing

While dividing, the data were expected to separate three regions, since k was selected as three, in k means algorithm. The flowchart, given in Figure 4.3, is the ultimate flowchart of the experiment. It is called ultimate, because all combinations were tried before deciding which classifier will be used on sub problems. As a result of these trials, kNN was assigned on the first sub problem, SVM was assigned on second sub problem, and Naive Bayes algorithm was assigned on last sub problem. According to sub problems, results of each classifier is given in Table 4.5. In this example, accuracy values of the classifiers were enough to determine when a classifier will be assigned.

Table 4.5 Results of Classifiers

	X1 Accuracy	**X2** Accuracy	**X3** Accuracy
ANN	82.46	84.51	34.38
KNN	**94.74**	25.35	71.88
NB	89.47	88.73	**100.00**
SVM	89.47	**90.14**	75.00

According to selected classifiers, specificity, selectivity, accuracy, positive predictive value and negative predictive value of final result were calculated. And these values are given in Table 4.6.

Table 4.6 Results of Experiment 3

Final Result	
Specificity	91.18
Sensitivity	98.28
Accuracy	**93.75**
PPV	86.36
NPV	98.94

According to the Table 4.6, the accuracy rate of this experiment is equal to 93.75% and this rate is the best accuracy rate when compared with the parallel structures. There is information about negative predictive value in Table 4.6. The rate of negative predictive value is nearly 99%. When it is thought there are 102 healthy people in test data, the experiment is quietly successful for detecting the healthy people. This rate is also the best negative predictive value obtained in this study. According to this result, there is one healthy person misclassified in Experiment 3.

As it can be seen from Table 4.6, best accuracy rate is equal to 93.75% among all experiments done in this study. It means that if there are 100 people, we can label them as PAF-patient or not with the correctness of 93.75%. In each iteration, there are 160 people divided for the test. So, 150 of 160 people are labelled correctly. According to the same table positive predictive value is equal to 86.36. So it means that nearly 50 of 58 people which are patient are labeled as patient. But remaining 8 patient people are labeled as non-PAF. Negative predictive value is equal to 98.94. It means that there are 102 healthy people. 101 of them are labeled as healthy. But one people which are not PAF-patient were labeled as PAF patient.

CHAPTER FIVE
CONCLUSION

Although it is very easy to diagnose PAF disease from the ECG recordings taken during the arrhythmic event, it is not easy to acquire the ECG signal during a PAF episode. This is due to the fact that PAF episodes could come to an end in minutes. Therefore, there would be no sufficient time for the subject to go to a health clinic and take an ECG record during the episode. Thus it would be very helpful if a computer based system that would be able to diagnose PAF disease from arrhythmia-free ECG records could be developed. The goal of this thesis is PAF screening of subjects based on their arrhythmia-free ECG records.

In this thesis, several different ensemble learning structures was developed to create a PAF screening system, which would be expected to be more effective than the ones based on only a single classifier. Both hierarchical and parallel types of ensemble learning were designed and their results were compared. The hierarchical structure of ensemble learning model was more successful than the parallel structure with the accuracy value of 93.75%. On the other hand, an accuracy rate of 89.5% was obtained by using the parallel structure. This result shows us that better results could be obtained as long as better hierarchical structures are constructed.

When an ensemble is constructed, it is expected that there would be an improvement in results. In this study, with parallel ensemble structures, highly successful results were obtained with the accuracy rate of 89.5%. In the meantime, no single neural network classifier was able to achieve that success rate. This result proves that the ensemble structures are very successful to compensate for the errors rooted from using just one classifier. For compensating the errors, both averaging and voting methods were tried in the combiner. In bagging algorithm, the results of averaging were more successful than voting. On contrary to this conclusion, voting was more successful when k fold cross validation was used. There was little difference between averaging and voting for both cross validation methods.

Different ensembles using k–fold cross validation and bootstrap resampling methods were compared. A rise in performance was seen when k-fold cross validation was used instead of bootstrap replicates. This conclusion was contrast to Breiman's study, which states that better results can be obtained by using bootstrap sampling as each bootstrap replicate creates a unique hypothesis and these unique hypotheses make the neural networks train better. The reason for this contradiction could be that in this study neural networks could not have been sufficiently trained. This could be caused by using small sized dataset. In addition to size of the dataset, bootstrap replicates contain examples which are approximately 63.2% of the entire dataset. When k – fold cross validation was used for training, better results were obtained. This may be rooted from the training rate of this type of cross validation was selected as 80%.

When results of the system designed in this study are compared with similar works in the literature, it can be concluded that the performances obtained are relatively high. Martinez et. al obtained ECG records of 46 PAF patients and 53 healthy subjects. Then they calculated the variability of many morphological features of P-wave. They found that using a decision tree with P-wave area and they achieved 95.42% accuracy to discriminate ECG segments of healthy subjects and patients suffering from PAF (Martinez et. al., 2012). Ros et al. used only train set of Physionet PAF Prediction Challenge Database (AFPDB) and they obtained 92% correct classification rate using 22 parameters obtained from P-wave analysis (Ros et. al., 2004). In 2016, I. Hilavin worked on this dataset, then she obtained accuracy with the rate of 95% by using the best performing classifier, i.e. support vector machine (SVM) (Hilavin, 2016). Martinez's study was more successful than the works of Hilavin and Ros. Results of this study were close to Martinez's accuracy rate by using hierarchical structure of ensemble learning with an accuracy value of 93.75%.

Based on the results of this study, it is hoped that this system can be effectively used for PAF screening purposes. As future work, the system can be improved to give sufficient results even when the ECG record is obtained from mobile measurement devices in relatively noisy conditions. In such a case, an interface can also be designed and added so that the system can be easily used by the subjects themselves in their home environments.

REFERENCES

American Heart Association, (2017), *What is Atrial Fibrillation?,* Retrieved April 2017. http://www.heart.org/HEARTORG/Conditions/Arrhythmia/AboutArrhythmia/What-is-Atrial-Fibrillation-AFib-or-AF_UCM_423748_Article.jsp#.WS7D8OvyiUk

Anumonwo, J. M., & Kalifa, J. (2014). Risk factors and genetics of atrial fibrillation. *Cardiology clinics*, *32* (4), 485-494.

Avnimelech, R., & Intrator, N. (1999). Boosted mixture of experts: an ensemble learning scheme. *Neural computation*, *11* (2), 483-497.

Becker, D. E. (2006). Fundamentals of electrocardiography interpretation. *Anesthesia progress*, *53* (2), 53-64.

Berntson, G. G., Thomas Bigger, J., Eckberg, D. L., Grossman, P., Kaufmann, P. G., Malik, M., ... & DER MOLEN, M. W. (1997). Heart rate variability: origins, methods, and interpretive caveats. *Psychophysiology*, *34* (6), 623-648., DOI: 10.1111/j.1469-8986.1997.tb02140.x

Bertsekas, D. P. (1999). *Nonlinear programming*, *3*, 275, Belmont: Athena scientific.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992), A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (144-152). ACM.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24* (2), 123-140. DOI: 10.1023/A:1018054314350

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

Chen, D., Bourlard, H., & Thiran, J. P. (2001). Text identification in complex background using SVM. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* IEEE.

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert systems with applications*, *40* (1), 200-210., DOI: 10.1016/j.eswa.2012.07.021.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20* (3), 273-297.

Cuingnet, R., Rosso, C., Chupin, M., Lehéricy, S., Dormont, D., Benali, H., ... & Colliot, O. (2011). Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome. *Medical image analysis*, *15* (5), 729-737.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Multiple classifier systems*, *1857*, 1-15., DOI: 10.1007/3-540-45014-9_1.

Duda R. O. ,& Hart P. E,& Stork D. G., (1973), Pattern Classification, Wiley – Interscience, *9* (25)

Duin, R. P. W. (2000). Prtools version 3.0: A matlab toolbox for pattern recognition. In *Proc. of SPIE*.

Epomedicine, Basics of ECG- Interpretation of waves and intervals, http://epomedicine.com/medical-students/ecg-interpretation-waves-intervals/

Fix E.,& Hodges J. L., (1989), An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation Commentary on Fix and Hodges (1951), International Statistical Review, *57* (3), 233-238, DOI: 10.2307/1403796.

Gaonkar, B., & Davatzikos, C. (2013). Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. *Neuroimage*, *78*, 270-283.

Gertsch, M. (2003). *The ECG: a two-step approach to diagnosis*. Springer Science & Business Media.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., …& Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. *101* (23), 215–220. DOI: 10.1161/01.CIR.101.23.e215

Hall P.,& Park B. U.,& Samworth R. J., (2008), Choice of neighbor order in nearest-neighbor classification, The Annals of Statistics, *36* (5), 2135-2152, DOI:10.1214/07-AOS537

HearthMath Institute Research Staff, (1993), Chapter 3: Heart Rate Variability, Science of the Heart, HearthMath Institute.

Hilavin I., (2016), Development of a System to Diagnose Atrial Fibrillation Patients from Arrhythmia Free ECG Records, Unpublished.

Huikuri, H. V., Mäkikallio, T. H., & Perkiömäki, J. (2003). Measurement of Heart Rate Variability by Methods Based on Nonlinear Dynamics. *Journal of Electrocardiology*, *36*(SUPPL.), 95–99. DOI: 10.1016/j.jelectrocard.2003.09.021

January, C. T., Wann, L. S., Alpert, J. S., Calkins, H., Cleveland, J. C., Cigarroa, J. E., ... & Murray, K. T. (2014). 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation. *Circulation*,

Kantelhardt, J. W., Koscielny-Bunde, E., Rego, H. H., Havlin, S., & Bunde, A. (2001). Detecting long-range correlations with detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, *295* (3), 441-454. DOI: 10.1016/S0378-4371(01)00144-3

Kotsiantis, S. (2011). Combining bagging, boosting, rotation forest and random subspace methods. *Artificial Intelligence Review*, *35* (3), 223-240. DOI: 10.1007/s10462-010-9192-8

Lee, H. G., Noh, K. Y., & Ryu, K. H. (2008). A data mining approach for coronary heart disease prediction using HRV features and carotid arterial wall thickness. In *BioMedical Engineering and Informatics*, *2008. BMEI 2008*, *1*,200-206, IEEE

Lewis D. D., (1998), Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval, European Conference on Machine Learning ECML 1998: Machine Learning: ECML-98, 4-15, DOI: 10.1007/BFb0026666

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, *1* (14), 281-297.

Malik, M., & Camm, A. J. (1995). *Heart rate variability*. NY: Armonk, Futura Pub. Co. Inc.

Marcin J., (2013), *What Is Paroxysmal Atrial Fibrillation?*, Retrieved April 2017, http://www.healthline.com/health/living-with-atrial-fibrillation/paroxysmal#AtrialFibrillation1

Martínez, A., Alcaraz, R., & Rieta, J. J. (2012). Study on the P-wave feature time course as early predictors of paroxysmal atrial fibrillation. *Physiological measurement*, *33* (12), 1959. DOI: 10.1088/0967-3334/33/12/1959

McLachlan, G., Do, K. A., & Ambroise, C. (2005). *Analyzing microarray gene expression data*, *422*, John Wiley & Sons.

Min, J. H., & Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert systems with applications*, *28*(4), 603-614.

MIT OpenCourseWare, P. Winston, *Learning: Support Vector Machines*, retrieved January, 2014, https://www.youtube.com/watch?v=_PwhiWxHK8o

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

Nattel, S., & Harada, M. (2014). Atrial remodeling and atrial fibrillation: recent advances and translational perspectives. *Journal of the American College of Cardiology, 63 (*22), 2335-2345. DOI: 10.1016/j.jacc.2014.02.555

Peterson L. E., (2009), K-nearest neighbor, *4* (2), pp.1883, DOI:10.4249/scholarpedia.1883.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, *6* (3), 21-45.

Polikar, R. (2012). Ensemble learning. In *Ensemble machine learning* 1-34. Springer US. DOI:10.1007/978-1-4419-9326-7_1

Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, *278* 6, 2039-2049.

Rish I., (2001), An Empirical Study of the Naïve Bayes Classifier, IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, *3,* 41-46. IBM New York.

Ros, E., Mota, S., Fernández, F. J., Toro, F. J., & Bernier, J. L. (2004). ECG Characterization of paroxysmal atrial fibrillation: parameter extraction and automatic diagnosis algorithm. *Computers in biology and medicine*, *34* (8), 679-696. DOI: 10.1016/j.compbiomed.2003.10.002

Schreier, G., Kastner, P., & Marko, W. (2001). An automatic ECG processing algorithm to identify patients prone to paroxysmal atrial fibrillation. In *Computers in Cardiology 2001*, 133-135, IEEE

Sharma B.,& Venugopalan K., (2014), Comparison of Neural Network Training Functions for Hematoma Classification in Brain CT Images, IOSR Journal of Computer Engineering (IOSR-JCE), *16* (1), 31-35, DOI: 10.9790/0661-16123135

Stan Z. L., A. K. Jain, (2015), Ear Biometrics, *Encyclopedia of Biometrics, Springer*, 363-368, DOI: 10.1007/978-1-4899-7488-4.

Tarvainen, M. P., Niskanen, J. P., Lipponen, J. A., Ranta-Aho, P. O., & Karjalainen, P. A. (2014). Kubios HRV–heart rate variability analysis software. *Computer methods and programs in biomedicine*, *113* (1), 210-220. DOI: 10.1016/j.cmpb.2013.07.024

Thomas, M., Das, M. K., & Ari, S. (2015). Automatic ECG arrhythmia classification using dual tree complex wavelet based features. *AEU-International Journal of Electronics and Communications*, *69* (4), 715-721, DOI: 10.1016/j.aeue.2014.12.013

Thong, T., McNames, J., Aboy, M., & Goldstein, B. (2004). Prediction of paroxysmal atrial fibrillation by analysis of atrial premature complexes. *IEEE Transactions on Biomedical Engineering*, *51* (4), 561-569.

Wang, Y., Agrafioti, F., Hatzinakos, D., & Plataniotis, K. N. (2008). Analysis of human electrocardiogram for biometric recognition. *EURASIP journal on Advances in Signal Processing*, *2008*, DOI: 10.1155/2008/148658.

Yang Y., Liu X., (1999), A re-examination of text categorization methods, SIGIR '99 Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 42-49, DOI: 10.1145/312624.312647

Yu C.,& Liu B., (2002), A Backpropagation Algorithm with Adaptive Learning Rate and Momentum Coefficient, Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on, IEEE, DOI: 10.1109/IJCNN.2002.1007668

Yu K. , Anton S., Volker T., (2003), Collaborative ensemble learning: combining collaborative and content-based information filtering via hierarchical bayes, Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence 616 - 623.

Yu, L., Wang, S., & Lai, K. K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. *Expert systems with applications*, *34* (2), 1434-1444.

Yu S., Shan S., Chen X., Gao W., (2009), Hierarchical Ensemble of Global and Local Classifiers for Face Recognition, IEEE Transactions on Image Processing, *18* (8), 1885 – 1896, DOI: 10.1109/TIP.2009.2021737

Zheng, L., Li, T., & Ding, C. (2010, December). Hierarchical ensemble clustering. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, 1199-1204. IEEE.

Zong, W., Mukkamala, R., & Mark, R. G. (2001). A methodology for predicting paroxysmal atrial fibrillation based on ECG arrhythmia feature analysis. In *Computers in Cardiology 2001,* 125-128, IEEE