ECG Signal Analysis using Principal Component Analysis

A Thesis submitted in partial fulfilment of the
Requirement for the award of degree of

Master of Engineering
in
Electronic Instrumentation and Control

Submitted by
ASHISH TUTEJA
(Roll No. 801051031)

Under the Esteemed Guidance
of
Mr. M. D. Singh
Assistant Professor

Department of Electrical and Instrumentation Engineering

Thapar University
(Established under the section 3 of UGC act, 1956)

Patiala, 147004, Punjab, India

July 2012

CERTIFICATE

I hereby declare that the Thesis entitled **ECG Signal Analysis Using Principal Component Analysis** is an authentic record of my own work carried out as the requirements for the award of the degree of M.E. (Electronic Instrumentation and Control Engineering) at Thapar University, Patiala, under the guidance of **Mr. M.D. Singh**, Assistant Professor, EIED.

The matter presented in this Thesis has not been submitted for the award of any other degree of this or any other university.

ASHISH TUTEJA
(Roll no-801051031)

Date 13/07/2019

It is certified that the above statement made by the student is correct to the best of my knowledge and belief.

Mr. M.D. Singh
Assistant Professor
Department of Electrical and Instrumentation Engineering

Countersigned by:

Dr. Smarajit Ghosh
Professor & Head, EIED
Thapar University, Patiala

Dr. S.K. Mohapatra
Dean, Academic Affairs
Thapar University, Patiala

ACKNOWLEDGEMENT

The real spirit of achieving a goal is through the way of excellence and austere discipline. I would have never succeeded in completing my task without the cooperation, encouragement and help provided to me by various personalities.

With deep sense of gratitude I express my sincere thanks to my esteemed and worthy supervisor, my mentor, **Mr. M. D. Singh**, Department of Electrical and Instrumentation Engineering, Thapar University, Patiala for his valuable guidance in carrying out this work under his effective supervision, encouragement, enlightenment and cooperation. Most of the novel ideas and solutions found in this thesis are the result of our numerous stimulating discussions. His feedback and editorial comments were also invaluable for writing of this thesis.

I shall be failing in my duties if I do not express my deep sense of gratitude towards **Dr. Smarajit Ghosh**, Professor and Head of the Department of Electrical & Instrumentation Engineering, Thapar University Patiala who has been a constant source of inspiration for me throughout this work.

I am also thankful to all the staff members of the Department for their full cooperation and help. This acknowledgement would be incomplete if I do not mention the emotional support and blessings provided by my friends. I had a pleasant enjoyable and fruitful company with them.

My greatest thanks are to all who wished me success especially my parents and friends, whose support and care makes me stay on earth.

PLACE: Thapar University, Patiala
DATE:

ASHISH TUTEJA
801051031

ABSTRACT

Electrocardiogram (ECG) monitoring is the most important and efficient way of preventing heart attack and many other cardiac abnormalities. The analysis of ECG has become an important topic of medical research. Many algorithms have been presented in the literature for ECG signal analysis using different algorithms. In this work, we have proposed a method for Pattern Recognition and Classification of Atrial Premature Beat (APB), Left Bundle Branch Block Beat (LBBB), Paced Beat (PB), Right Bundle Branch Block Beat (RBBB) and Ventricular Premature Beat (VPB). We extract 450 signals from original database. We employ Principal Component Analysis (PCA) to extract the principal characteristics of the data. Then Classification Analysis is done using three best classifiers, namely, Support Vector Machines (SVM), k-Nearest Neighbour (k-NN) and the BayesNet with the Accuracy 99.34%, 99.34% and 98.54% respectively. This work gave better results as compared to other related work presented in the literature.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1
INTRODUCTION

1.1 Overview

Cardiac diseases are the main cause of deaths around the world as many people die because of sudden heart attack. Also, a large number of people die because of the delay or errors in diagnosing their cardiac diseases. Electrocardiogram (ECG) signal has been intensively used by cardiac specialists to effectively diagnose cardiovascular diseases. Electrocardiography (ECG) can detect and record abnormal heart rhythms enough to diagnose specific forms of heart disease. A typical ECG signal contains special waves such as P, T waves as well as QRS complex. Cardiologists investigate each of these waves, complexes and other features such as RR interval, PR interval, PR segment, ST interval and ST segment etc. to diagnose various types of abnormal cardiac symptoms. So, fast and effective monitoring and interpretation of ECG is important. The shape of ECG conveys very important hidden information in its structure. The amplitude and duration of each wave in ECG signals are often used for the manual analysis. Thus, the volume of the data being enormous and the manual analysis is tedious and very time-consuming task. Naturally, the possibility of the analyst missing vital information is high. Therefore, medical diagnostics can be performed using computer-based analysis and classification techniques. The subtle changes in amplitude and duration of ECG cannot be deciphered precisely by the naked eye, hence imposing the need for a computer assisted diagnosis tool.

1.2 Motivation

Many ECG processing techniques which deal mainly with ECG pattern recognition [1,2], parameter extraction, spectrotemporal techniques late potential characterization [3], arrhythmia detection [4] and noise removal [5] have been presented in the literature. Our work is related to ECG pattern recognition and classification. Accurate detection of beats helps determine different types of arrhythmia which are relevant to diagnose heart disease. Automatic assessment of arrhythmia for patients is widely studied. Thus, automatic classification of heartbeat is useful and helps medical professionals for fast and accurate diagnosis of heart beat. Utilization of pattern classifier techniques can improve ECG arrhythmia diagnoses. Our proposed work has

two phases: Pattern Recognition and Pattern Classification. Pattern recognition (or data reduction) is done with Principal Component Analysis (PCA).

1.3 Objective and Scope

The objective of this dissertation is to develop an algorithm for recognizing and classifying atrial premature, left bundle branch block beat, right bundle branch block beat, Paced Beat and ventricular premature beat. In order to do so, we extract 450 signals from the original MIT-BIH database. Each signal representing a single and complete heart beat. We extract the principal characteristics of the signal by means of the Principal Component Analysis (PCA) technique. And then Classification analysis is done using different classifiers. It's very important to detect arrhythmia and our work provides a method for arrhythmia detection and machine pre-diagnosis.

1.4 Organisation of the Dissertation

The reminder of this thesis is organized as follows: Chapter 2 contains the literature review of the work related to our proposed work. Chapter 3 presents the basic concepts involving Electrocardiography (ECG), which gives us an insight about basics of ECG, how it's done and its significance. Chapter 4 is devoted to the Methods and Methodologies used in our work, which involves introduction to Principal Component Analysis (PCA), the significance of PCA and how to apply PCA. The weka 3.6 software is explained along with the three classifiers used. Chapter 5 presents our proposed work in detail, including the use of database, PCA results, WEKA results and related discussion. Then we have the conclusion and references.

Chapter 2
LITERATURE REVIEW

Hao Zhang *et. al.* [6] proposed a method for Arrhythmia detection and classification. An algorithm is developed for recognizing and classifying normal beat, left bundle branch block beat, right bundle branch block beat and premature ventricular contraction (PVC).More than 6000 signals were extracted from the original database, each signal representing a single and complete heart beat. The principal characteristics of the signal are extracted by means of the Principal Component Analysis (PCA) technique. Support Vector Machine (SVM) has a major predominance over other classification methods in complicated problems. SVM method is applied to classify the ECG data into the 4 categories of heart diseases. Experiments were conducted using MIT-BIH arrhythmia database.

B. Anuradha *et. al.* [7] presented a paper where a study of the nonlinear dynamics of electrocardiogram (ECG) signals for arrhythmia characterization is considered. The statistical analysis of the calculated features indicated that the features differ significantly between normal heart rhythm and the different arrhythmia types and hence, were rather useful in ECG arrhythmia detection. The discrimination of ECG signals using non-linear dynamic parameters is very important in the cardiac disease therapy and chaos control for arrhythmia defibrillation in the cardiac system. The four non-linear parameters considered for cardiac arrhythmia classification of the ECG signals were Spectral entropy, Poincaré plot geometry, Largest Lyapunov exponent and Detrended fluctuation analysis which are extracted from heart rate signals. Linguistic variables (fuzzy sets) are used to describe ECG features. All the ECG data required for this work is used from the MIT-BIH dataset.

R. Ganesh Kumar *et. al.* [8] presented an ECG classification method for arrhythmic beat classification using RR interval. The methodology is based on discrete cosine transform (DCT) conversion of RR interval. The RR interval of the beat was extracted from the ECG and used as feature. DCT conversion of RR interval is applied and the beats are classified using random tree. The experiments were conducted using MIT-BIH arrhythmia database.

Maedeh Kiani Sarkaleh *et. al.* [9] proposed an expert system for ElectroCardioGram (ECG) arrhythmia classification. Discrete wavelet transform is used for processing ECG recordings, and extracting some features, and the Multi-Layer Perceptron (MLP)

neural network performed the classification task. Two types of arrhythmias can be detected by the proposed system. Some recordings were taken from MIT-BIH arrhythmias database for training and testing the neural network based classifier.

Abhinav Vishwa *et. al.* [10] proposed an automated Artificial Neural Network (ANN) based classification system for cardiac arrhythmia using multi-channel ECG recordings. Neural network model with back propagation algorithm is used to classify arrhythmia cases into normal and abnormal classes. Networks models were trained and tested for MIT-BIH arrhythmia. The different structures of ANN were trained by mixture of arrhythmic and non arrhythmic data patient.

Roshan Joy Martis *et. al.* [11] presented a method in which automatic classification of five types of ECG beats of MIT-BIH arrhythmia database is done. Three approaches are adopted, the first one uses principal components of segmented ECG beats, the second approach uses principal components of error signals of linear prediction model, whereas the third approach uses principal components of Discrete Wavelet Transform (DWT) coefficients as features. These approaches were then independently classified using feed forward neural network (NN) and Least Square- Support Vector Machine (LS-SVM). MIT-BIH arrhythmia database has been used in this work.

Ataollah Ebrahim Zadeh *et. al.* [12] presented a paper which investigates the design of an efficient system for recognition of the premature ventricular contraction from the normal beats and other heart diseases. The system includes three main modules: denoising module, feature extraction module and classifier module. In the denoising module, the stationary wavelet transform for noise reduction of the electrocardiogram signals is proposed. The Feature extraction module proposes a proper combination of the morphological-based features and timing interval-based features. As the classifier, several supervised classifiers are investigated; they are: a number of multi-layer perceptron neural networks with different number of layers and training algorithms, support vector machines with different kernel types, radial basis function and probabilistic neural networks. Data is obtained from MIT-BIH database.

Sung-Nien Yu *et. al.* [13] proposed an Electrocardiogram (ECG) beat classification system based on wavelet transformation and probabilistic neural network (PNN) is proposed to discriminate six ECG beat types. The ECG beat signals are first decomposed into components in different Subbands using discrete wavelet transformation. Three sets of statistical features of the decomposed signals as well as the AC power and the instantaneous RR interval of the original signal were exploited to

characterize the ECG signals. A PNN is adopted to classify the feature vectors. MIT-BIH arrhythmia database is used for analysis and recognition.

Peng Li *et. al.* [14] proposed a mixed Support Vector Machine (SVM)-based hierarchical learning approach to detect abnormal ECG beats. A global bi-class support vector classifier is first trained using ECG beats from different patients in a database. Then, a local novelty detector SVM is trained using only normal ECG beats from a specific patient. The fusion of the global and local classifiers significantly improved the classification. MIT-BIH database was used.

S. S. Mehta *et. al.* [15] proposed an Entropy based method for the detection of QRS complexes (cardiac beat) in the single lead Electrocardiogram (ECG). Digital filtering techniques are used to remove noise and base line wander in the ECG signal. Entropy criterion is used to enhance the QRS complexes. Support Vector Machine (SVM) is used as a classifier to delineate QRS and non-QRS regions. The performance of the algorithm is evaluated against the standard CSE ECG database.

Mi Hye Song *et. al.* [16] proposed an algorithm for arrhythmia classification, associated with the reduction of feature dimensions by linear discriminant analysis (LDA) and a support vector machine (SVM) based classifier. The features of LDA and PCA were classified using SVM and the results were compared. For a cross-validation procedure, this SVM classifier was compared with Multilayer Perceptrons (MLP) and Fuzzy Inference System (FIS) classifiers. For collecting arrhythmia data, the ECG data from the MIT/BIH Arrhythmia Database is used.

Dayong Gao *et. al.* [17] presented a system for detection of cardiac arrhythmias within ECG signals, based on a Bayesian Artificial Neural Network (ANN) classifier. The Bayesian (or Probabilistic) ANN Classifier is built by the use of a logistic regression model and the back propagation algorithm based on a Bayesian framework. The UCI Arrythmia dataset is used.

Dipti Patra *et. al.* [18] proposed a scheme to integrate fuzzy c-means (FCM) clustering, principal component analysis (PCA) and neural networks (NN) for ECG beat classification. The PCA is used to decompose ECG signals into weighted sum of basic components that are statistically mutual independent. In addition, FCM clustering is among considerable techniques for data reduction. A back propagation neural network (BPNN) is employed as classifier. A comparative study of performance of four structures such as FCM-NN, PCA-NN, FCM-ICA-NN, and FCM-PCA-NN are investigated. MIT-BIH arrhythmias database is used for experiments.

Ye Wenyu *et. al.* [19] proposed a method for clustering of QRS complexes. The method integrates principal component analysis (PCA) with self-organizing maps neural network (SOM's). The QRS complex feature is extracted based on PCA and the unsupervised SOM is employed to cluster the data. The characteristics and the behaviour of the proposed method applying different SOM's architectures are studied. The method is tested with **MIT-BIH** database.

Francisco Castells *et. al.* [20] presented a paper which reviews the current status of principal component analysis in the area of ECG signal processing. The fundamentals of PCA are briefly described and several applications are reviewed where PCA techniques have been successfully applied including data compression, ST-T segment analysis for the detection of myocardial ischemia and abnormalities in ventricular repolarization, extraction of atrial fibrillatory waves for detailed characterization of atrial fibrillation, and analysis of body surface potential maps.

Mehmet Korürek *et. al.* [21] presented a paper in which Ant Colony Optimization (ACO) based clustering analysis of ECG arrhythmias, using both time domain and discrete wavelet transform (DWT) based frequency domain features is proposed. PCA is applied on wavelet coefficients them in order to decrease their number to the number of time domain features. Different types of feature sets, namely, time domain feature set, frequency domain feature set and the mixture of them, are tried and the classification results are compared. Different types of feature sets are tried and the classification results are compared. These are: time domain feature set, frequency domain feature set and the mixture of them. A neural network algorithm is developed in parallel to verify and measure the ACO classifier's success.

Chapter 3
ELECTROCARDIOGRAPHY (ECG)

3.1 Introduction

The heart can be called as a Muscular Pump that continuously pumps blood throughout life. If the heart stops beating for a few minutes, death is inevitable. So it's very important to know the status of the heart to prevent any unwanted situation.ECG is a non-invasive, trans-thoracic recording of the electrical activity of the heart over a period of time via skin electrodes.

The ECG display indicates the overall rhythm of the heart and weaknesses in the different parts of heart muscle. It is one of the best ways to measure and diagnose abnormal rhythms of the heart. It is used to measure the rate and regularity of heartbeats as well as the size and position of the chambers, the presence of any damage to the heart.

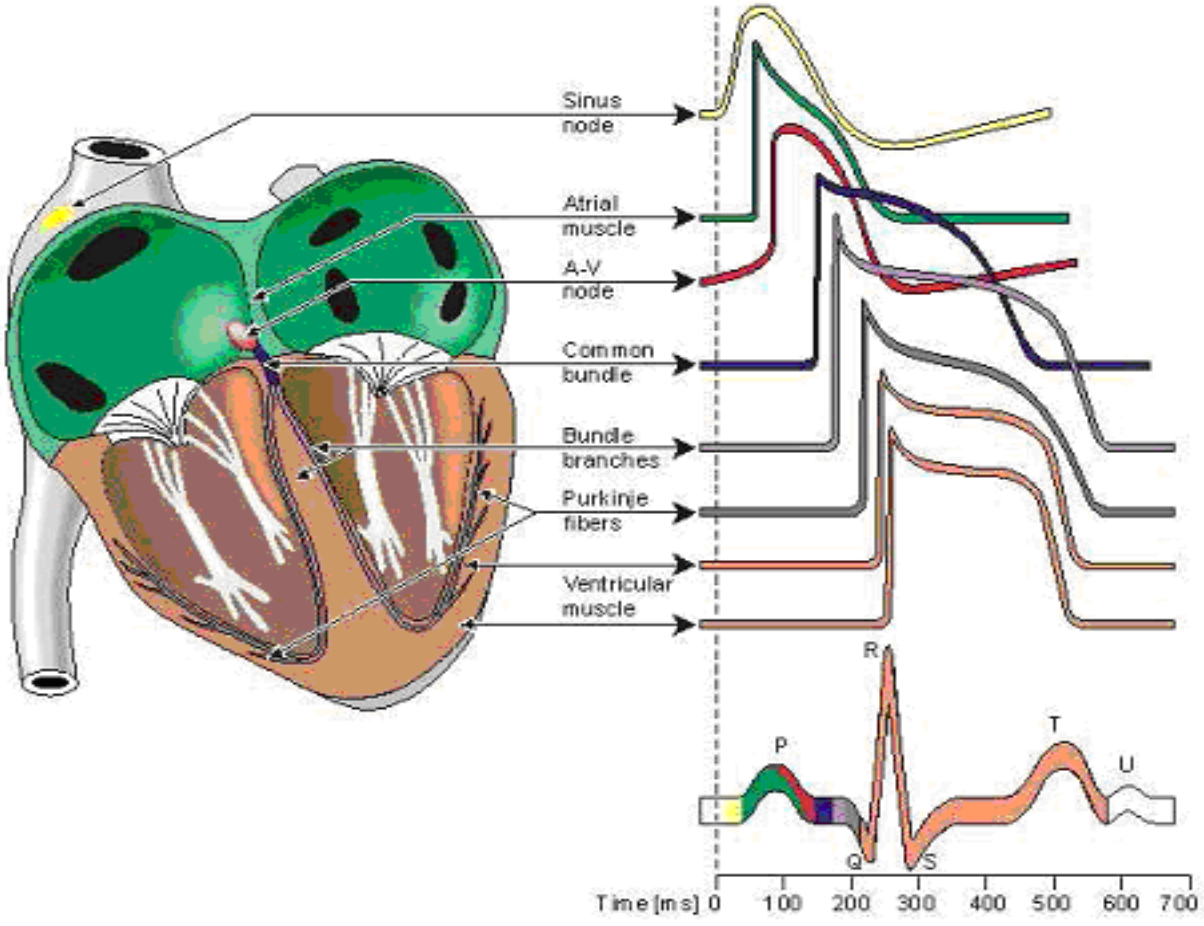


Figure 3.1 ECG waveform characteristics and their corresponding positions in heart. [7]

By selectively placing electrodes on the skin, the electrical waves passing through various parts of the heart can be measured. An ECG displays the voltage that is taken from the pairs of these electrodes. Each pair is supposed to measure a specific set of muscles from different direction. The electric wave has a magnitude and direction and thus can be represented as a vector which keeps on changing both in magnitude and direction within one cardiac cycle or heartbeat. An ECG uses pair of electrodes called leads and these leads pick up components of this vector from different directions. It is the best way to measure and diagnose the abnormal rhythms.

3.1.1 Leads

The term "lead" in electrocardiography is used to refer to two different things. Generally, the word lead may be used to refer to the electrical cable attaching the electrodes to the ECG recorder. It may be acceptable to refer to the "left arm lead" as the electrode (and its cable) that should be attached at or near the left arm. There are usually ten of these electrodes in a standard "12-lead" ECG. Alternatively, the word lead may refer (mostly in ECG) to the tracing of the voltage difference between two of the electrodes and is what is actually produced by the ECG recorder. Each will have a specific name. For example "Lead I" (lead one) is the voltage between the right arm electrode and the left arm electrode, whereas "Lead II" (lead two) is the voltage between the right limb and the feet. Twelve of these types of leads form a "12-lead" ECG.

Placement of Electrodes: A 12- lead ECG uses ten electrodes. The electrodes usually consist of a conducting gel, embedded in the middle of a self-adhesive pad onto which cables clip. Sometimes the gel also forms the adhesive. They are labeled and placed on the patient's body as shown in figure 2. It shows the proper placement of limb electrodes, colour coded as recommended by American Heart Association.

Limb Leads: Leads I, II and III are called limb leads in both 5 and 12- lead configuration. The electrodes are located on limbs- one on each arm and one on left leg.Lead I is the voltage between the (positive) left arm (LA) electrode and right arm (RA) electrode:

$$I = LA - RA. \qquad 3.1$$

Similarly,

$$II = LL - RA. \qquad 3.2$$

And

$$III = LL - LA. \qquad 3.3$$

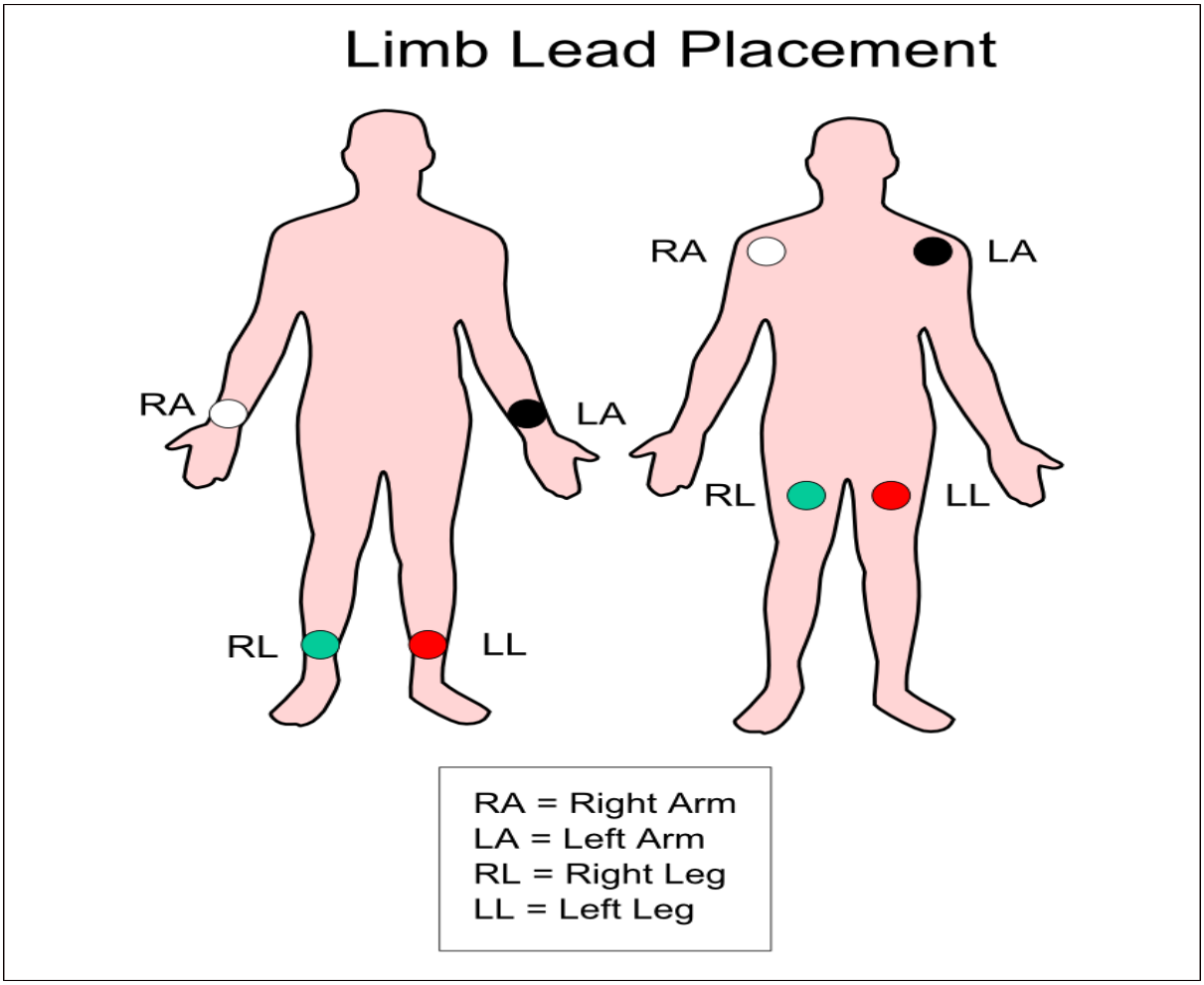


Figure 3.2 Placement of Electrodes

Unipolar & Bipolar Leads: There are two types of leads: unipolar and bipolar. Bipolar leads have one positive and one negative pole. In a 12-lead ECG, the limb leads (I, II and III) are bipolar leads. Unipolar leads also have two poles, as a voltage is measured; however, the negative pole is a composite pole (Wilson's central terminal, or WCT) made up of signals from lots of other electrodes. In a 12-lead ECG, all leads besides the limb leads are unipolar (aVR, aVL, aVF, V_1, V_2, V_3, V_4, V_5, and V_6).

Wilson's central terminal V_W is produced by connecting the electrodes, RA; LA; and LL, together, via a simple resistive network, to give an average potential across the body, which approximates the potential at infinity (i.e. zero):

$$V_W = 1/3\ (RA + LA + LL) \qquad 3.4$$

3.1.2 Waves and Intervals

A typical ECG tracing of the cardiac cycle (heartbeat) consists of a P wave, a QRS complex and a T Wave. The baseline voltage of the electrocardiogram is known as the *isoelectric line*. Typically the isoelectric line is measured as the portion of the tracing following the T wave and preceding the next P wave.

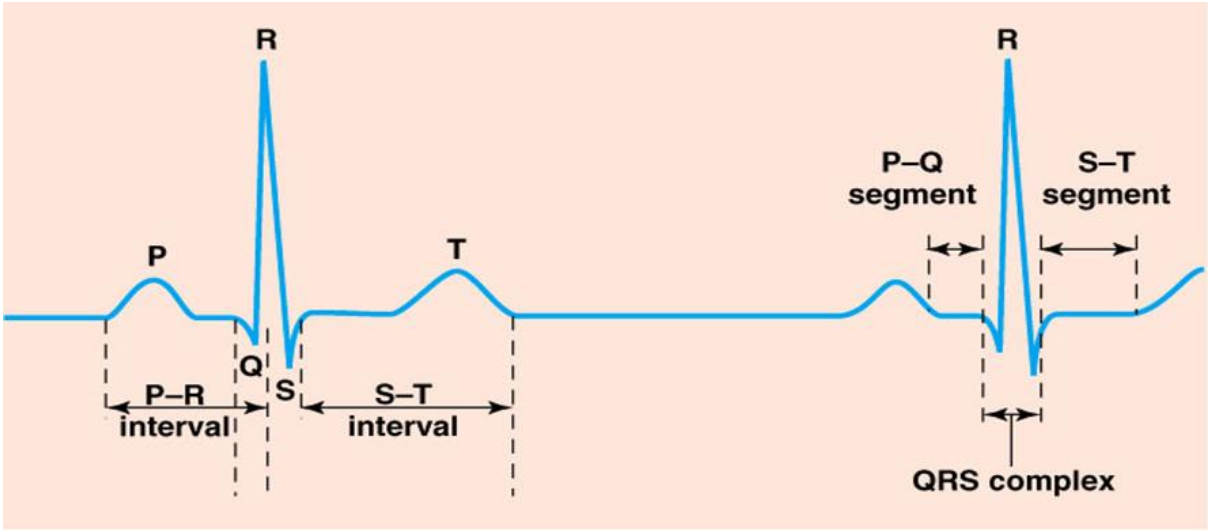


Figure 3.3. The ECG signal and its different components.

The normal conduction pathway is as follows: SA node > atrial muscle > AV node > bundle of His > Left and Right Bundle Branches > Ventricular muscle. Following is the table showing the some ECG waves and the normal time duration.

Table 3.1 ECG Features and their Description

Feature	Description	Duration
RR Interval	The interval between an R wave and the next R wave. Normal resting heart rate is between 60 and 100 bpm	0.6 to 1.2s
P Wave	During normal atrial depolarization, The P Wave is generated on the ECG.	80 ms
QRS Complex	The QRS complex reflects the rapid depolarization of the right and left ventricles. They have a large muscle mass compared to the atria and so the QRS complex usually has a much larger amplitude than the P-wave.	80 to 120 ms
ST Segment	The ST segment represents the period when the ventricles are depolarized. It is isoelectric.	80 to 120 ms
T wave	the apex of the T wave is referred to as the absolute refractory period. The last half of the T wave is referred to as the relative refractory period (or vulnerable period).	160 ms

3.2 The Significance of ECG

The jeopardy of Cardiac diseases is aggravating in the modern world leading to death of many people because of sudden heart attack. At the same time, a large number of people die because of the delay or errors in diagnosing their cardiac diseases. Cardiac specialists use Electrocardiogram (ECG) signal to effectively diagnose cardiovascular diseases. The electrocardiogram (ECG) is a representative signal that contains information about the heart's working condition. The shape and size of the P-QRS-T wave and the time intervals between various peaks contain useful information about the nature of abnormalities in the heart. Being the bio-signals having highly subjective nature, the symptoms sometimes may appear at random in the timescale. The presences of cardiac abnormalities are generally reflected in the shape of ECG waveform and heart rate.It is the best way to measure and diagnose abnormal rhythms of the heart. These abnormal rhythms are a result of some cardiac abnormalities.

The ECG is the most commonly performed cardiac test. This is because the ECG is a useful screening tool for a variety of cardiac abnormalities; ECG machines are readily available in most medical facilities; and the test is simple to perform, risk-free and inexpensive. From the ECG tracing, the following information can be determined:

- the heart rate and rhythm
- whether there are "conduction abnormalities" (abnormalities in how the electrical impulse spreads across the heart)
- whether there has been a prior heart attack
- whether there may be coronary artery disease
- whether the heart muscle has become abnormally

Limitations of ECG:

- The ECG reveals the heart rate and rhythm only during the time that the ECG is taken. If intermittent cardiac rhythm abnormalities are present, the ECG is likely to miss them. Ambulatory monitoring is needed to record transient arrhythmias.
- The ECG can often be normal or nearly normal in patients with undiagnosed coronary artery disease or other forms of heart disease (false negative results.)
- Many "abnormalities" that appear on the ECG turn out to have no medical significance after a thorough evaluation is done (false positive results).

3.3 ECG Waveforms describing various Cardiac Abnormalities

Figure below shows a typical ECG tracing of a normal Heartbeat or Cardiac cycle. We shall now discuss various cardiac abnormalities depicted through ECG Waveforms. [22]

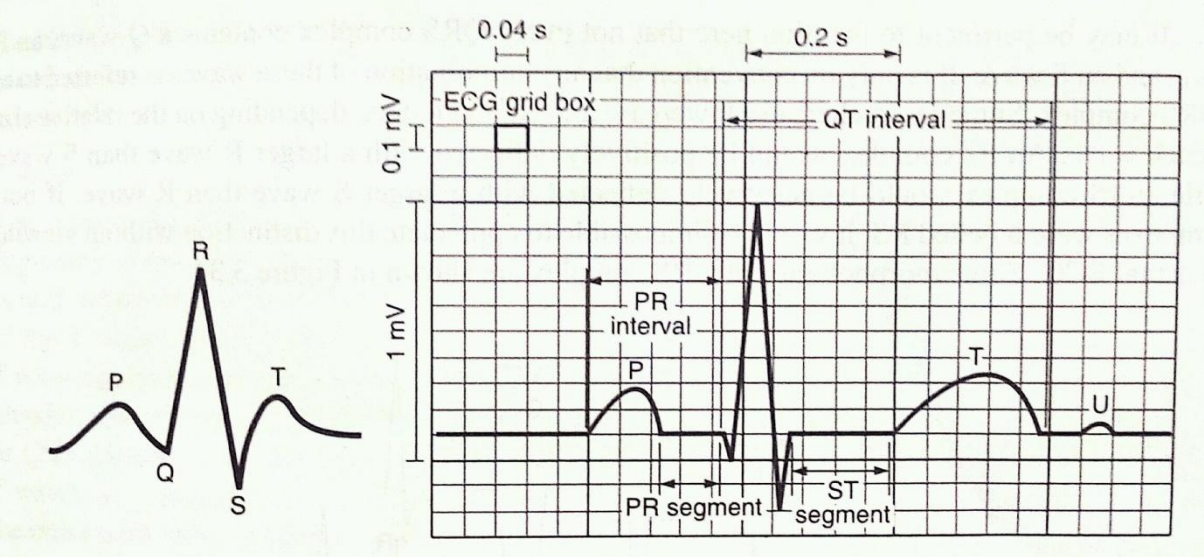


Figure 3.4. ECG Waveform on the Graph Paper

P wave: The p wave of ECG represents the start of a cardiac cycle or heartbeat. Its shape and duration helps in determining many clinical conditions:-

1. **Cardiac Arrhythmias**: Heartbeat may be too fast or too slow, or may be regular or irregular.
2. **Atrial Hypertrophy**: It means Atrial Enlargement.
3. **Atrial Fibrillation**: It is indicated by absence of P waves which tells about a condition that involves quivering of the heart muscle of the atria.
4. **Atrial Flutter**: Indicated by Saw-tooth-formed P wave, which is due to rapid, uncoordinated contraction of the atria.

QRS waves: The QRS complex is a composition of three characteristic points Q, R and S on the ECG which corresponds to Depolarisation of the ventricles. The duration, amplitude and the morphology of the QRS complex are helpful in diagnosing many cardiac abnormalities such as **Cardiac Arrhythmias, Conduction Abnormalities, Ventricular Hypertrophy,** and **Myocardial Infraction.** Some Morphologies of QRS complex are shown below.

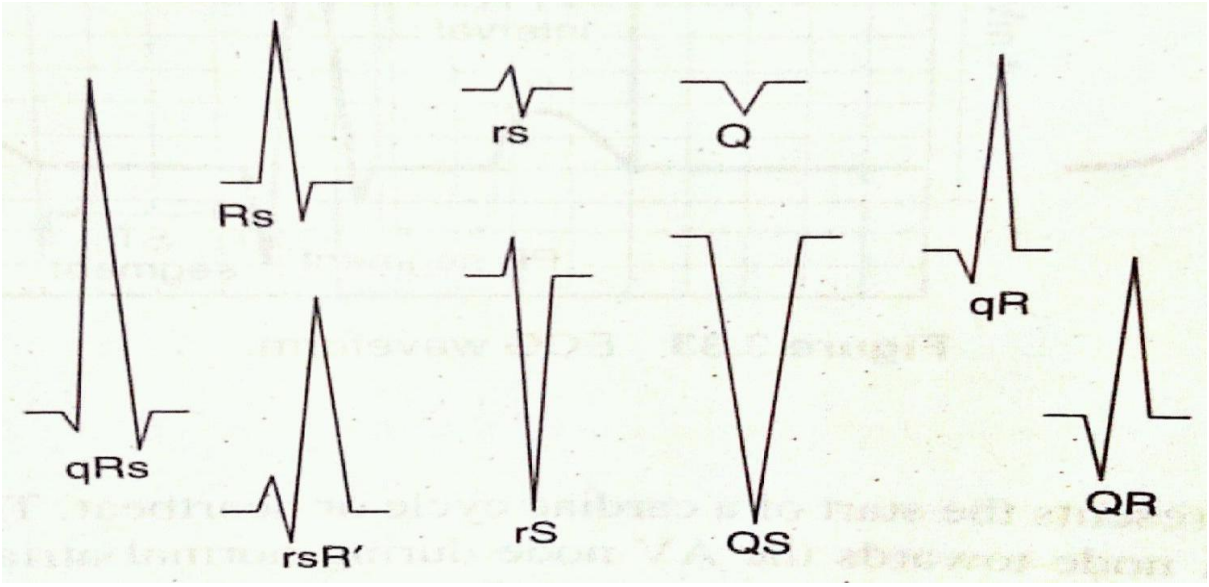


Figure 3.5 Morphologies of QRS complex [18]

PR/PQ Interval: The PR interval is measured from the beginning of the P wave to the beginning of the QRS complex**.** A number of clinical conditions of the heart can be diagnosed from PR interval. For example a PR interval of over 0.2 sec. may indicate **First degree heart block**, and a short PR interval may indicate a pre-excitation syndrome. One such condition is known as **Wolff-Parkinson-White syndrome.**

ST Segment: The ST segment connects the QRS complex and T wave**.** Flat, Down-scoping, or depressed ST segments may indicate **Coronary Ischemia**. Also, an ST segment elevation may indicate **Myocardial Infraction.**

T Wave: The T wave characterizes the repolarisation or recovery of the ventricles. An inverted or negative T wave can be a sign of **Coronary Ischemia, Left Ventricular Hypertrophy.** A tall symmetrical T waves indicate **Hyperkalemia** and a flat T wave indicates **Hypokalemia.**

QT interval: This is measured from the beginning of the QRS complex to the end of the T wave. They are helpful in providing information about the QT syndrome. A syndrome is a group of symptoms that together are characteristic of a specific disorder or disease.

U wave: A prominent U wave may be indicative of **Hypokalemia, Hypercalcemia, Thyrotoxicosis.** An inverted U wave may be on account of **Myocardial Ischemia** or **left ventricular volume overload.**

Some pathological entities which can be seen on the ECG

Shortened QT interval	Hypercalcemia, some drugs, certain genetic abnormalities.
Prolonged QT interval	Hypocalcemia, some drugs, certain genetic abnormalities.
Flattened or inverted T waves	Coronary ischemia, hypokalemia, left ventricular hypertrophy, digoxin effect, some drugs.
Hyperacute T waves	Possibly the first manifestation of acute myocardial infarction, where T waves become more prominent, symmetrical, and pointed.
Prominent U waves	Hypokalemia.

3.5 Symptoms generally indicating use of ECG

- **Cardiac murmurs**.
- **Syncope or collapse.**
- **Seizures**.
- **Perceived cardiac dysrhythmias**.
- **Symptoms of myocardial infarction.**

3.6 Cardiac Arrhythmias

During normal rhythm, the heart beats are regular, producing a single coordinated electrical wave that can be seen as a normal electrocardiogram (ECG). During arrhythmias such as ventricular tachycardia and ventricular fibrillation, this normal behaviour is disrupted and the ECG records rapid rates with increased complexity. The underlying cause of many arrhythmias is the development of a reentrant circuit of electrical activity that repetitively stimulates the heart and produces contractions at a rapid rate. During tachycardia, a single wave can rotate as a spiral wave, producing fast rates and complexity. During fibrillation, a single spiral wave can degenerate into multiple waves. Because contraction is stimulated by the pattern of electrical waves, arrhythmias can compromise the heart's ability to pump blood and sometimes may be lethal.

Sinus rhythm is the normal rhythm of the heart and results from proper activation of the entire heart in proper sequence. Any variation from normal sinus rhythm is termed an arrhythmia.

Ventricular Fibrillation: It is a life threatening arrhythmia which is characterized by rapid, irregular activation of the ventricles and thereby prevents an effective mechanical contraction. During ventricular fibrillation, the ECG has no distinctive QRS complexes but instead consists of an undulating baseline of variable amplitude. Although the sinus node continues to function properly, P waves cannot be discerned in the VF waveform.

Ventricular tachycardia is a rhythm characterized by wide, bizarre QRS complexes and frequent ventricular premature contractions in a row. VT may be paroxysmal or chronic and

often signifies underlying myocardial disease. Mechanisms include reentry, increased automaticity, and triggered activity.

Atrial flutter is an AV node-independent intra-atrial macro-reentry rhythm, in which the atrial anatomy sustains a loop of continuous depolarization, often around the tricuspid valve annulus in the right atrium. Atrial flutter can be paroxysmal or chronic and may be associated with extremely rapid ventricular response rates.

Atrial fibrillation is characterized by rapid, irregular activation of the atria. Causes can include reentry and abnormal automaticity. AF can be paroxysmal or chronic.

Chapter 4
MATERIALS AND METHODOLOGIES

4.1 Principal Component Analysis (PCA)

4.1.1 Introduction

Principal Component Analysis (PCA) is a multivariate technique used to analyse a data table that contain observations which are described by several inter-correlated quantitative dependent variables. It is a way of identifying patterns in data and represents the data in such a way as to highlight their similarities and differences. [23]

A Variable Reduction Procedure

Principal component analysis is a variable reduction procedure. It is useful when you have obtained data on a number of variables (possibly a large number of variables), and believe that there is some redundancy in those variables. Redundancy here means that some of the variables are correlated with one another, possibly because they are measuring the same construct. Because of this redundancy, it should be possible to reduce the observed variables into a smaller number of artificial variables (principal components) that will account for most of the variance in the observed variables. [24]

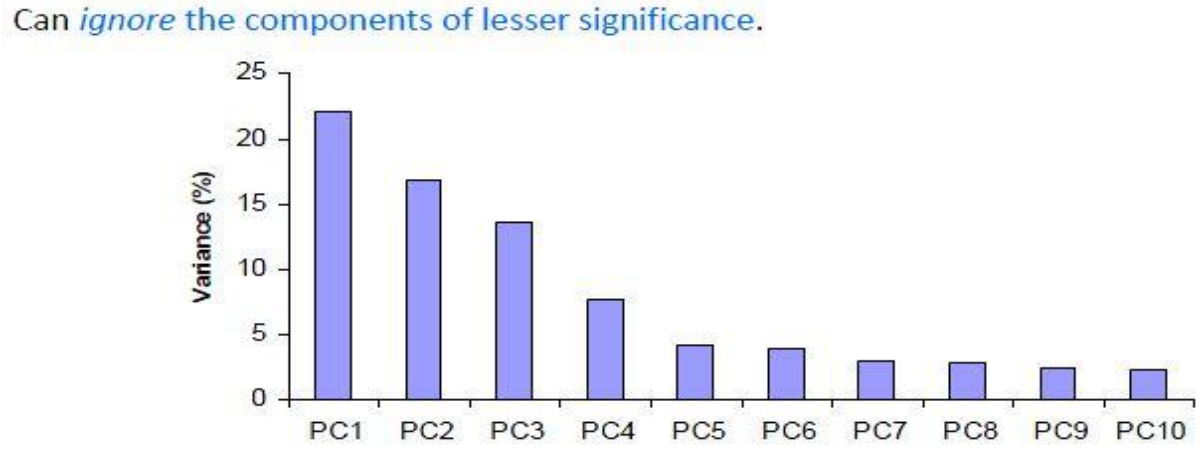


Figure 4.1. Principal Components with Variance Percentage

4.1.2 The Assumptions of PCA

1. **Linearity:** PCA assumes the data set to be linear combination of the variables.

2. The importance of Mean and Covariance: It is not necessary that the directions of maximum variance will contain good features for discrimination.

3. Large Variances have important dynamics: Components with larger variance correspond to interesting dynamics and lower ones correspond to noise.

Where regression determines a line of best fit to a data set, factor analysis determines several orthogonal lines of best fit to the data set. Orthogonal means "at right angles", the lines are perpendicular to each other in n-dimensional space. [25]

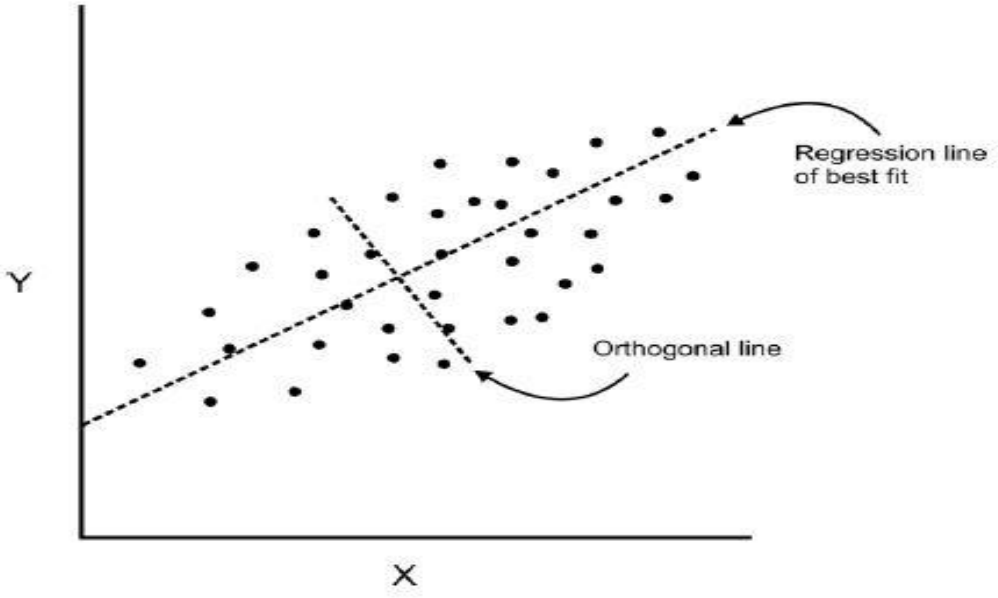


Figure 4.2. Plot of data showing Orthogonal line and Regression line

4.1.3 Terms related to PCA

n- dimensional space: It is the variable sample space. It depends on the number of variables in the data set. So, in a data set with 5 variables, the sample space is 5-dimensional.

Components: A variable system for the data set is chosen by linear transformation, such that the greatest variance of the data set lies on the first axis, called the first Principal Component, the second greatest variance on the second axis and so on …(figure) The components are uncorrelated, since they are orthogonal to each other in the sample space.

Component Loadings: In the multivariate space, the correlation between the component and the original variables is called the component loadings. It tells how much of the variation in a variable is explained by the component. (figure 4.3)

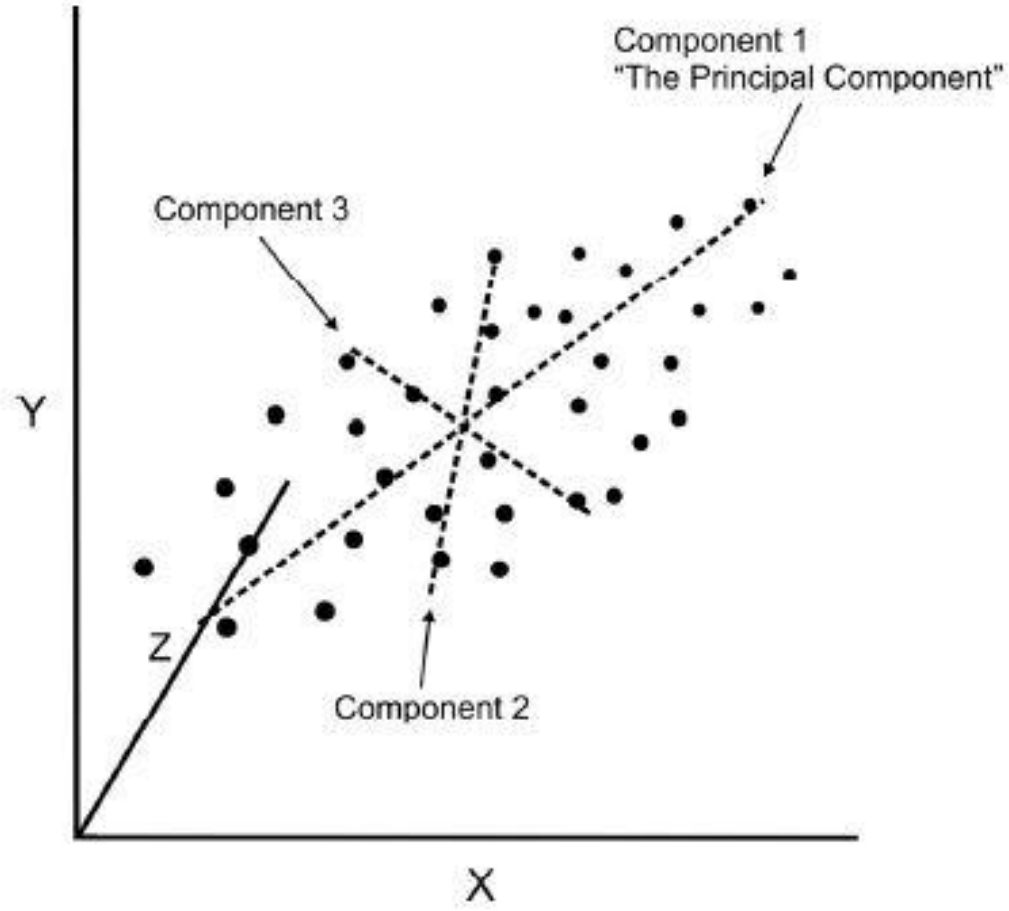


Figure 4.3. Axis of the Components showing the Variance of Data

4.1.4 The Significance of PCA

Sometimes it is hard to find the pattern in the data where the data is of very high dimension and that is where the PCA comes into picture. PCA is a powerful analysing tool. The goal is to extract the important information from the table of observations, represent it as a set of new orthogonal variables called principal components, and to display the pattern of similarity or dissimilarity of the observations and of the variables. PCA , being a simple, non-parametric method of extracting relevant information from confusing data sets, is used abundantly in all forms of analysis.PCA shows the way for

reducing a complex data set to a lower dimension to reveal the sometimes hidden, simplified structure that often underlie it.

One major advantage of PCA is that once we have found these patterns in the data, and we can compress the data, i.e., by reducing the number of dimensions, without much loss of information. This technique can be used in image compression, data compression, variable reduction and many other applications. [23, 24, 25]. Applications like pattern recognition also use PCA. A pattern recognition process cannot be accomplished with a large amount of data efficiently. The large data set usually intervenes with the classification process. So we apply Principal Component Analysis (PCA) to a dataset. Principal Component Analysis (PCA) is a technique that is commonly used in multivariate statistical analysis. Its goal is the reduction in the number of dimensions from a numerical measurement of several variables. With this dimensional reduction, this technique looks for simplifying a statistical problem with the minimal lost of information. This method is also used in signal processing for separating a linear combination of signals generated from sources that are statistically independent. This is performed by representing the data with a new coordinate system. This transformation is bidirectional and no information is lost [25].

4.1.5 Applications of PCA

There are many different areas of applications where PCA is successfully employed and has played the role of an information extracting tool. Applications like Feature Extraction [26, 27], Dimensionality Reduction [28], Pattern Recognition [29,30,31], Image Compression [32, 33] etc are good examples showing PCA as a powerful and reliable data and signal analysis technique.

4.1.6 Algebra of PCA

The PCA technique takes a data matrix of *n* objects (the rows) by *p* variables (the columns), which may be correlated and summarizes it by uncorrelated axes (principal components or principal axes) that are linear combinations of the original *p* variables. The steps for calculating principal component are as follows:

Covariance: The first step is general and involves standardization of data set. This is done to observe the total variance of the data set. Here each variable is transformed so it

has a mean of zero and variance of one. This can simply be done by subtracting each column value from its mean. Thus the variance of each variable becomes one and each of them contributes a unit to the total variance. The components which are formed during the analysis partition this variance only. An example of 2- dimensional data is taken to demonstrate the methodology of PCA [23].

Table1. 4.1 PCA Example Data

Original Data		**Mean Subtracted Data**	
X	**Y**	**$X - \bar{X}$**	**$Y - \bar{Y}$**
2.5	**2.4**	**0.69**	**0.49**
0.5	**0.7**	**-1.31**	**-1.21**
2.2	**2.9**	**0.39**	**0.99**
1.9	**2.2**	**0.09**	**0.29**
3.1	**3.0**	**1.29**	**1.09**
2.3	**2.7**	**0.49**	**0.79**
2	**1.6**	**0.19**	**-0.31**
1	**1.1**	**-0.81**	**-0.81**
1.5	**1.6**	**-0.31**	**-0.31**
1.1	**0.9**	**-0.71**	**-1.01**

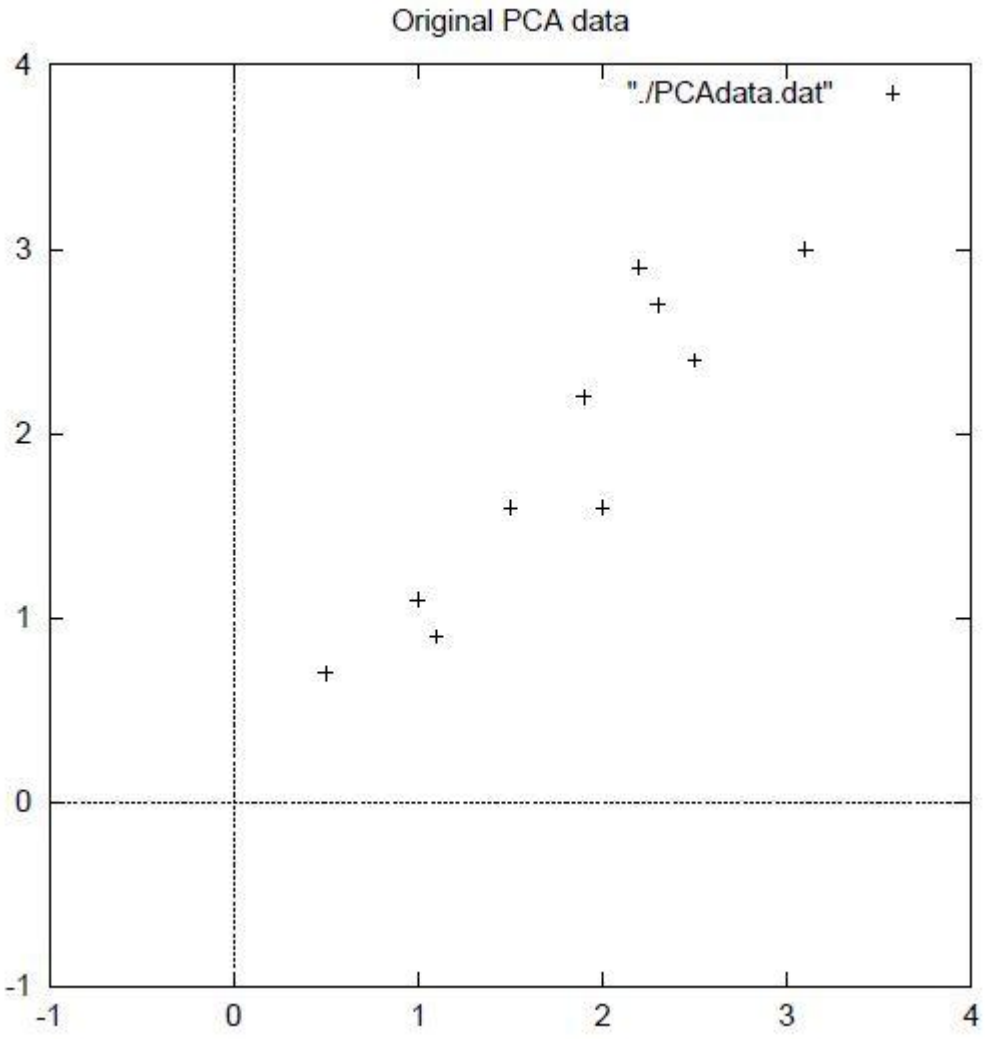


Figure 4.4 Plot of the Data

The second step involves creating a covariance matrix from this obtained standardized data set matrix. The covariance between two variables is a correlation between them divided by the products of their standard deviation. The covariance matrix is thus a square matrix obtained. Since our data is 2- dimensional, the covariance matrix will be 2×2. Here the covariance matrix will be

$$\begin{bmatrix} 0.616555556 & 0.615444444 \\ 0.615444444 & 0.716555556 \end{bmatrix}$$

In the third step the Eigenvalues and eigenvectors of this covariance matrix are calculated. These are important as they tell us useful information about our data.

$$\text{Eigenvalues} = \begin{bmatrix} 0.0490833989 \\ 1.28402771 \end{bmatrix}$$

$$\text{Eigenvectors} = \begin{bmatrix} -0.735178656 & -0.677873399 \\ 0.677873399 & -0.735178656 \end{bmatrix}$$

As shown in the figure 1 it is observed that the data has quite a strong pattern. On top of data both the eigenvectors are also plotted. They appear as diagonal dotted lines on the plot, perpendicular to each other. But, more importantly, they provide information about the patterns in the data. We can see how one of the eigenvectors goes through the middle of the points, like drawing a line of best fit. That eigenvector is showing us how these two data sets are related along that line. The second eigenvector gives us the other, less important, pattern in the data, that all the points follow the main line, but are off to the side of the main line by some amount.

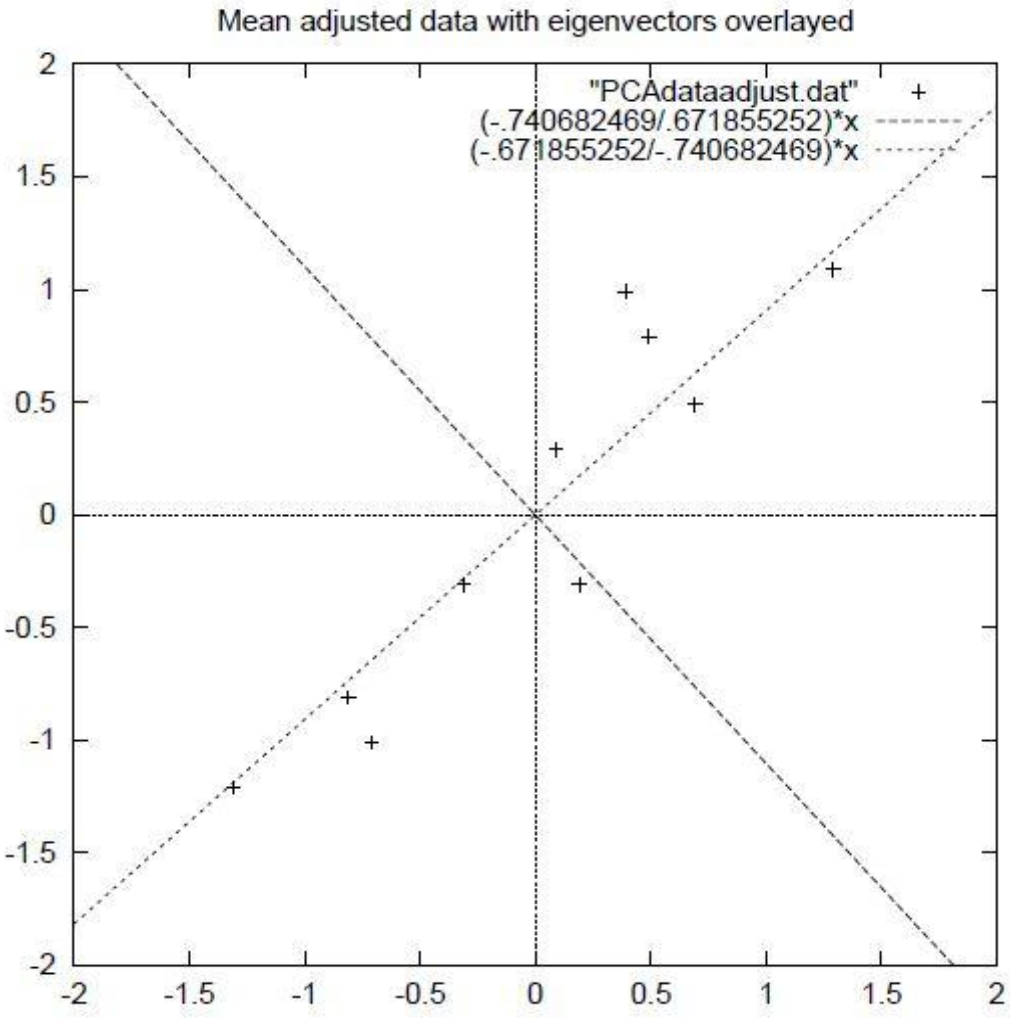


Figure 4.5 A plot of the normalised data (mean subtracted) with the eigenvectors of the covariance matrix overlayed on top.

The eigen values represent the fraction of total variance contained in the corresponding principal component. The diagonal of eigen value matrix gives variance. It contains elements on the diagonal only.

The eigenvectors of the covariance matrix represents the principal components corresponding to their eigen values. It is required that the eigen values should be arranged in decreasing order and then the vectors should be calculated correspondingly. This will give principal components with decreasing variance i.e. the most significant principal component will occur first. We create a Feature Vector which is nothing but matrix of vectors. Our example set of data has 2 eigenvectors, so we have two choices. We can either form a feature vector with both of the eigenvectors:

$$\begin{bmatrix} -0.677873399 & -0.735178656 \\ -0.735178656 & 0.677873399 \end{bmatrix}$$

or, we can choose to leave out the smaller, less significant component and only have a single column:

$$\begin{bmatrix} -0.677873399 \\ -0.735178656 \end{bmatrix}$$

The component loadings can be calculated by multiplying the eigenvector matrix with the square root of eigenvalue matrix. The square of each element of the loading matrix gives the percentage of each variable contained in each component.

Next is the selection of significant principal components by looking at the variance contained in them and ignoring the others.

Now the new data set called the component scores can be calculated in terms of principal components. This can be done simply by multiplying the original data set with the required principal component vectors. This gives a new reduced and completely uncorrelated data set.

The major constituent variables of the significant components have been calculated by calculating their composition percentage in it. This is done by squaring each element of the component loading matrix and then multiplying it by hundred.

Correlation: Principal component analysis depends upon the fact that there exists some correlation between the variables in the data set. The correlation is the association between two variables, the amount by which they co-vary. A worked example is shown below [34].

Table 4.2 Example Data with Standard Scores

Original Data			**Standard Scores of Data**		
$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$
1	2	3	-1.46	-1.10	-0.73
2	3	4	-1.10	-0.73	-0.37
3	1	1	-0.73	-1.46	-1.46
4	6	7	-0.37	0.37	0.73
5	5	5	0.00	0.00	0.00
6	4	2	0.37	-0.37	-1.10
7	8	9	0.73	1.10	1.46
8	9	8	1.10	1.46	1.10
9	7	6	1.46	0.73	0.37

The correlation matrix becomes

$$\begin{pmatrix} 1.000 & 0.8333 & 0.5833 \\ 0.8333 & 1.000 & 0.9167 \\ 0.5833 & 0.9167 & 1.000 \end{pmatrix}$$

Whose eigen values come out to be

$$\Lambda_1 = 2.5636$$

$$\Lambda_2 = 0.4214$$

$$\Lambda_3 = 0.0151$$

Thus the first principal λ_1 accounts for the 85.45% of the total variance. Similarly, λ_2 accounts for 14.05% and λ_3 for 0.5% of the total variance.

And the corresponding eigenvectors are

$$E = \begin{bmatrix} 0.5421 & -0.7620 & -0.3542 \\ 0.6209 & 0.0792 & 0.7799 \\ 0.5662 & 0.6427 & -0.5160 \end{bmatrix}$$

Then we have the principal components

Table 4.3 Scores for the nine data points on the three components.

Component 1	Component 2	Component 3
-1.8855	0.5588	0.0399
-1.2540	0.5422	0.0069
-2.1298	-0.4979	-0.1267
0.4423	0.7765	0.0372
0.0000	0.0000	0.0000
-0.6490	-1.0112	0.0512
1.9031	0.4690	-0.1581
2.1209	-0.0150	0.1859
1.4520	-0.8205	-0.1362

In principal components analysis the transformation of the system of reference from one set of axes (the variables) to another (the principal components) is done. Nothing "new" is created; nothing "old" is lost. PCA is a technique that is commonly used in multivariate statistical analysis. Its goal is the reduction in the number of dimensions from a numerical measurement of several variables.

How many Axes are needed? How many Components should be selected? Well, the determination of number of axes or components to retain is done with the help of some methods. Two common techniques are: The Kaiser criterion and the Scree Test.

The Kaiser Creterion: First, we can retain only factors with eigenvalues greater than 1. In essence this is like saying that, unless a factor extracts at least as much as the equivalent of one original variable, we drop it. This criterion was proposed by Kaiser (1960), and is probably the one most widely used. In our example above, using this criterion, we would retain 2 factors (principal components).

The scree test: A graphical method is the *scree* test first proposed by Cattell (1966). We can plot the eigenvalues in a simple line plot. The place (points) where the smooth decrease of eigenvalues appears to level off to the right of the plot.

4.1.7 PCA: Advantages and Disadvantages

Advantages of PCA

The Principal Component Analysis is a useful technique which offers many advantages when applied on data analysis:-

1. Principal component analysis is a good technique for dimensionality reduction applied on multivariate analysis.

2. It is an optimal and linear scheme for compressing a set of high dimensional vectors into a set of lower dimensional vectors and then reconstructing.

3. Also, it computes the model parameters directly from the data – for example by diagonalising the sample covariance.

4. Since the model parameters require only matrix multiplications, the compression and decompression are easy operations to perform.

5. It is used in many applications that include data compression, image processing, visualisation, exploratory data analysis, pattern recognition and time series prediction. [22, 24, 35].

Disadvantages of PCA

Despite of many good useful features of PCA, it often incorporates many shortcomings which are as follows:-

1. The PCA technique is non-parametric, so no prior knowledge can be incorporated.

2. PCA data reduction often incurs a loss of information.

3. The naïve methods for finding the principal component directions find trouble with high dimensional data or large numbers of data points, as difficulties would be there in terms of computational complexity and data scarcity.

4. Another shortcoming is that one cannot deal with incomplete data set, where some of the points are missing.

5. The standard PCA algorithm is based on the assumption that data have not been spoiled by outliers (error points) [22, 24, 35].

4.2 k-Nearest Neighbour Algorithm (k-NN)

This is one of the oldest, simplest and most popular classification algorithms used in data mining and pattern recognition field. K-NN algorithm tries to find k-nearest points to the testing point and labels(maps) the testing point to a class which appeared maximum number of times in the circle enclosing k-training points. This algorithm is also called as Lazy learning method and it requires very less or no training. The main steps of the algorithms are

1) Define k-value

2) Train the model

3) Distance measurement

4) Labelling.

The figures 4.6(a) and 4.6(b) shows a general approach to label (map, classify) a given data set (point) from a set of predefined labels.

Fig 4.6(a). Scatter plot of two classes.

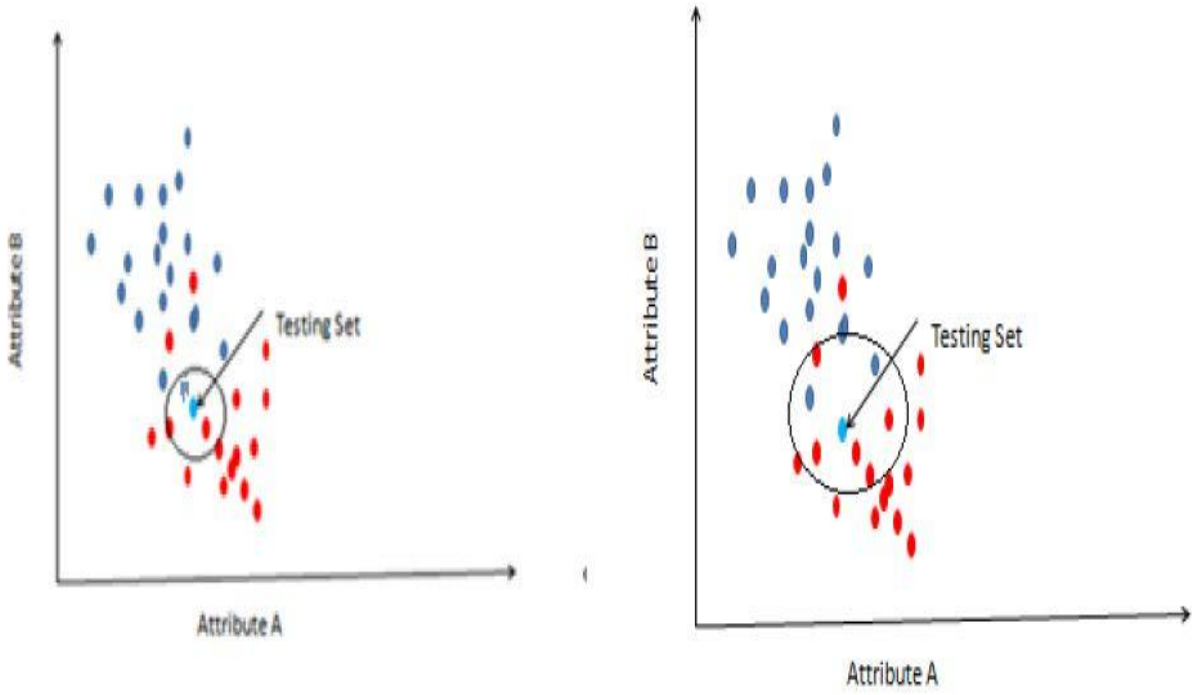


Figure 4.6(b). Decision making for different k-values

Figure 10(a) shows a scatter-plot of two-classes, in which red ‘dots’ belong to Class YES and Blue ‘dots’ belong to class NO. The sky-blue coloured ‘dot’ is a point which is to be labelled. Actually the testing dot here belongs to the class YES. Let’s see how k-NN algorithm labels the testing set in the following section.

Case1. k=3: Here the value of k is 3, i.e. we are considering 3 nearest points to classify the testing set. We find that out of 3-nearest points, three dots belong to the class YES and one-point belongs to the class NO. Majority is the class YES. Hence the model labels the testing set as class YES.

Case2. k=7: Here the value of k is 7, i.e. we are considering 7 nearest points to classify the testing set. We find that out of 7-nearest points, five dots belong to the class YES and otwo-point belongs to the class NO. Majority is the class YES. Hence again the model labels the testing set as class YES.

As we have seen, more the number of k-values more the training set it uses for training the model. In the first case the probability of the testing set belonging to class YES was 0.66. But in the second case the probability of labelling the test set to class YES was increased to 0.71, which is more accurate. Hence we can understand a very important fact that, more the value of k more the accurate is the model.

4.3 Support Vector Machine (SVM)

A standard, binary, SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the input, making the SVM a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

In a two-class learning task, the aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data. The metric for the concept of the "best" classification function can be realized geometrically. For a linearly separable dataset, a linear classification function corresponds to a separating hyper-plane *f (x)* that passes through the middle of the two classes, separating the two. Once this function is determined, new data instance *xn* can be classified by simply testing the sign of the function
f (xn); *xn* belongs to the positive class if *f (xn)* > 0.

Because there are many such linear hyper-planes, what SVM additionally guarantee is that the best such function is found by maximizing the margin between the two classes. Intuitively, the margin is defined as the amount of space, or separation between the two classes as defined by the hyper-plane. Geometrically, the margin corresponds to the shortest distance between the closest data points to a point on the hyper-plane. Having this geometric definition allows us to explore how to maximize the margin, so that even though there are an infinite number of hyper-planes, only a few qualify as the solution to SVM. To ensure that the maximum margin hyper-planes are actually found, an SVM classifier attempts to maximize the following function with respect to **_w** and *b*: a_i y_i

$$\mathbf{L_p = 1/2\ ||\vec{W}|| - \sum_{i=1}^{t} ai\ yi\ (\vec{W}.\ \vec{X_l} + b) + \sum_{i=1}^{t} ai} \qquad 4.1$$

where *t* is the number of training examples, and *αi , i* = 1, . . . , *t*, are non-negative numbers such that the derivatives of *L P* with respect to *αi* are zero. *αi* are the Lagrange multipliers and *L P* is called the Lagrangian. In this equation, the vectors **_w** and

constant *b* define the hyper-plane. The figure 11 shows how the separation-line is drawn and maximization of the margin.

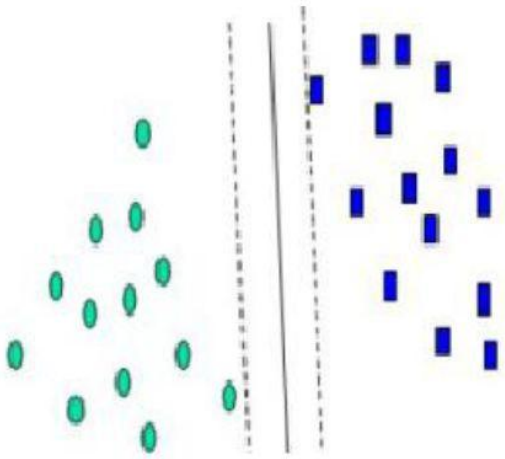


Figure 4.7(a) Linera Dividing Line

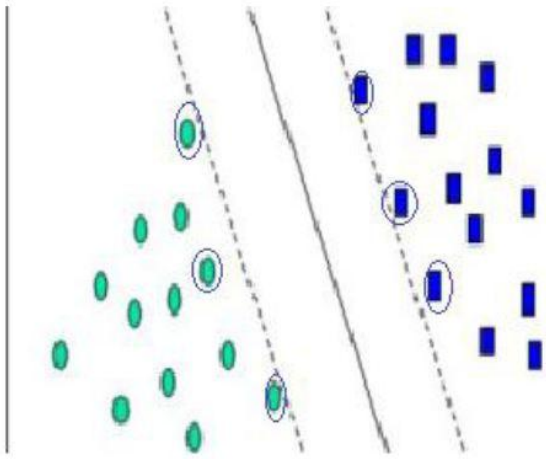


Figure 4.7(b) Linear Dividing Line

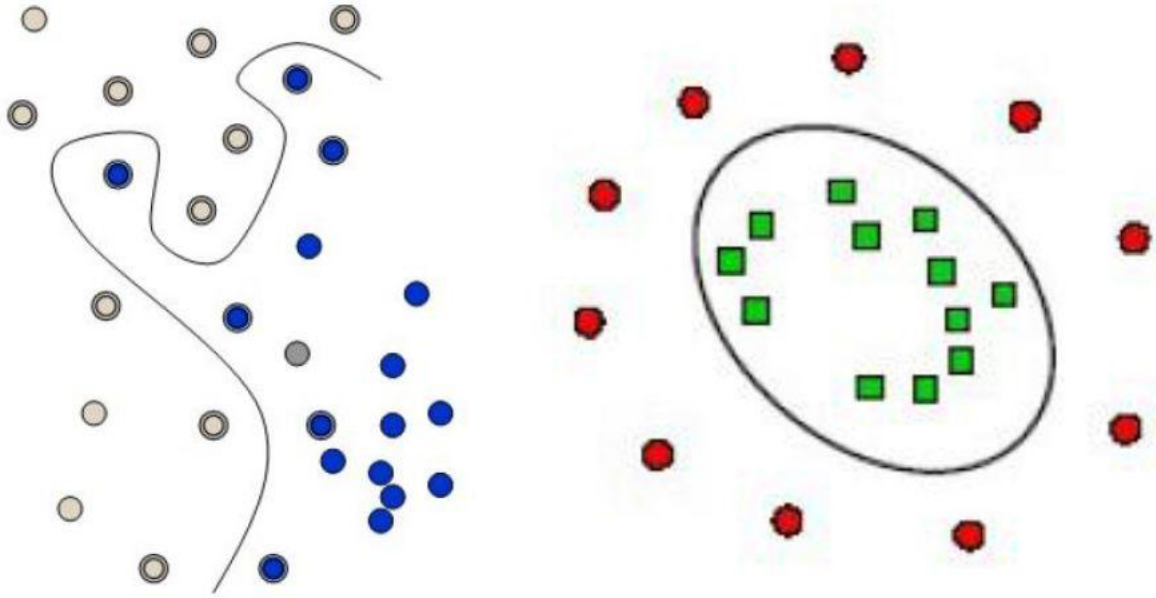


Figure 4.8. Non-linear dividing line.

The figure 4.7(a) shows that, the margin of separation line is very small and hence it can be proved that generalization error will be high in this case. Figure But the margin in figure 4.7(b) is greater than the margin shown in fig 4.7(a). The encircled dots are the points which actually helps define the margin and those points are called Support Vectors, hence the name Support Vector Machine for this classifier algorithm. Figure 4.8 shows how a hyper-plane is drawn in case non-linear problems.

4.4 Bayesian Network Classifier (BayesNet)

Let $U = \{x1, \ldots, xn\}$, $n \geq 1$ be a set of variables. A Bayesian network B over a set of variables U is a network structure BS, which is a directed acyclic graph (DAG) over U and a set of probability tables $BP = \{p(u|pa(u))|u \in U\}$ where pa(u) is the set of parents of u in BS. A Bayesian network represents a probability distributions.

$$\mathbf{P(U) = \Pi_{u \in U}\, p(u|pa(u)).} \qquad 4.2$$

The classification task consist of classifying a variable y = x0 called the class variable given a set of variables x = x1 . . . xn, called attribute variables. A classifier h : x → y is a function that maps an instance of x to a value of y. The classifier is learned from a dataset D consisting of samples over (x, y). The learning task consists of finding an appropriate Bayesian network given a data set D over U.

4.4.1 Inference algorithm

To use a Bayesian network as a classifier, one simply calculates argmaxyP(y|x) using the distribution P(U) represented by the Bayesian network. Now note that

$$\mathbf{P(y|x) = P(U)/P(x)}$$

$$\mathbf{\propto P(U)}$$

$$\mathbf{\Pi}$$
$$\mathbf{u \in U}$$

$$\mathbf{p(u|pa(u))} \qquad 4.3$$

And since all variables in x are known, we do not need complicated inference algorithms, but just calculate (1) for all class values.

4.4.2 Learning Algorithms

The dual nature of a Bayesian network makes learning a Bayesian network as a two stage process a natural division: first learn a network structure, then learn the probability tables.

There are various approaches to structure learning and inWeka, the following areas are distinguished:

• **Local score metrics**: Learning a network structure BS can be considered an optimization problem where a quality measure of a network structure given the training data Q(BS|D) needs to be maximized. The quality mea- sure can be based on a Bayesian approach, minimum description length, information and other criteria. Those metrics have the practical property that the score of the whole network can be decomposed as the sum (or product) of the score of the individual nodes. This allows for local scoring and thus local search methods.

• **Conditional independence tests**: These methods mainly stem from the goal of uncovering causal structure. The assumption is that there is a network structure that exactly represents the independencies in the distribution that generated the data. Then it follows that if a (conditional) indepen- dency can be identified in the data between two variables that there is no arrow between those two variables.

• **Global score metrics**: A natural way to measure how well a Bayesian net- work performs on a given data set is to predict its future performance by estimating expected utilities, such as classification accuracy. Cross- validation provides an out of sample evaluation method to facilitate this by repeatedly splitting the data in training and validation sets. A Bayesian network structure can be evaluated by estimating the network's parameters from the training set and the resulting Bayesian network's performance determined against the validation set. The average performance of the Bayesian network over the validation sets provides a metric for the quality of the network. Cross-validation differs from local scoring metrics in that the quality of a network structure often cannot be decomposed in the scores of the individual nodes. So, the whole network needs to be considered in order to determine the score.

• **Fixed structure**: Finally, there are a few methods so that a structure can be fixed, for example, by reading it from an XML BIF file3.

Chapter 5
The Proposed Work

5.1 Overview

Our work is related to ECG pattern recognition and classification. The ECG consists of three basic waves: the P, QRS, and T. These waves correspond to the far field induced by specific electrical phenomena on the cardiac surface, namely, the atrial depolarization (P wave), the ventricular depolarization (QRS complex), and the ventricular repolarization (T wave). ECG signal does not look the same in all the leads of the standard 12-lead system used in clinical practice. They usually change over different leads. Most cardiovascular diseases are caused by some kind of physical malfunction of one or several parts of the heart. This can certainly have a reflection on the shape of the ECG signal. So we conduct our research on a single-beat basis, attempting to discriminate different diseases from ECG data. The following block diagram in the figure 5.1 shows the workflow of the proposed work.

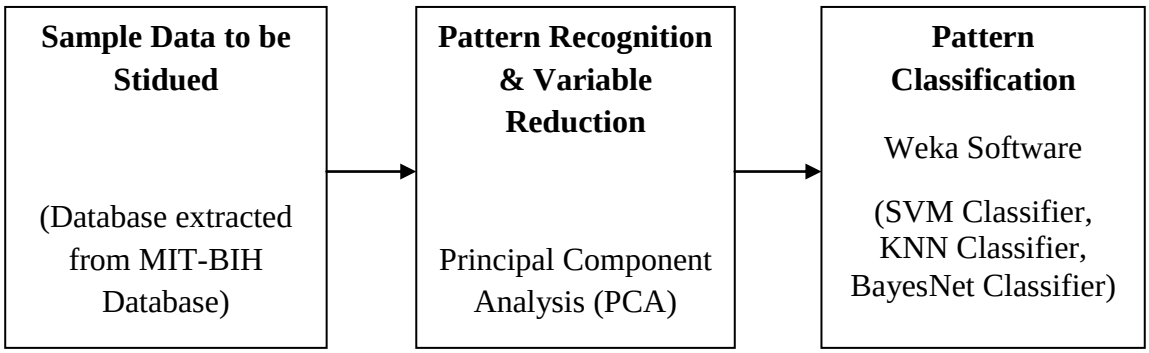


Figure 5.1 Workflow of the Proposed Work

5.2 The Database

We obtain our data set from MIT-BIH Arrhythmia Database [36]. The MIT-BIH Arrhythmia Database contains 48 half-hour excerpts of two-channel ambulatory ECG recordings. The recordings were digitized at 360 samples per second per channel with 11-bit resolution over a 10 mV range. Two or more cardiologists independently annotated each record; disagreements were resolved to obtain the computer-readable reference annotations for each beat (approximately 110,000 annotations in all) included with the database [37].

A new database from the original database for our experiment, where each signal represents a single heart beat. To test the integrated method, a set of signals is extracted from MIT-BIH. Since there are a few categories of abnormal QRS complexes in one record, we select different abnormal signals from several records so that the classification ability of the method can be studied conveniently. Five types of signals appeared frequently in the database are selected. The extracted data of ECG complexes is centered around R peak. Considered that some PVC duration is great and sometimes R peak detection may be not the centre of the complex, we have selected segment of 75ms before the fiducial point and 75ms after it. Signals are arranged in order to form a matrix where each column represents a signals.

The features of our database are:

1. We are using only one lead signal (limb lead II or MLII) to generate our data.
2. Each data in our experimental data set represents a single beat. To obtain the data, the signals are extracted with 151 samples, where the R wave was the 76th sample. The whole data set has 450 signals.
3. The data set has five types of signals: Atrial Premature Beat, Left Bundle Branch Block Beat, Right Bundle Branch Block Beat, Ventricular Premature Beat and Paced Beat. These reference signals were taken from the following records from MIT BIH Arrhythmia database.

Table 5.1 Sample Data to be Studied

Record	**A**	**L**	**P**	**R**	**V**
Signal Type	**100**	**118**	**102**	**109**	**106**

Table 1 shows the sample data to be studied where **A** is Atrial Premature beat, **L** is Left Bundle Branch Block Beat, **P** is Paced Beat, **R** is Right Bundle Branch Block Beat and **V** is Ventricular Premature Beat.

5.3 Results and Discussion

Applying PCA to *n* ECG leads that are statistically independent gives *n* new signals or principal components. The first signal corresponds to the principal component with highest variance while the *n-th* signal corresponds to the principal component with the lowest variance.

We apply PCA to our dataset containing 450 signals. These signals are of 5 different arrhythmias namely, Atrial Premature beat, Left Bundle Branch Block beat, Paced beat, Right Bundle Branch Block beat and Ventricular Premature beat. Samples in the form of text are saved in excel file and then called in MATLAB. Then in MATLAB we apply PCA on the data. We get 450X450 matrix of Eigen vectors or Principal Components. And we get New Data by multiplying this Eigen Vectors Matrix with Original data Matrix. This new data is nothing but the old data organised in a new way. The first few Principal Components contain the highest energy of the dataset.

Table 5.2 Principal Components Contribution.

Principal Component	Relative Contribution	Density Contribution
1	60.9318	60.9318
2	16.3021	77.2339
3	5.5650	82.7989
4	3.4770	86.2759
5	0.7222	86.9981
6	0.4891	87.4872
7	0.2408	87.7280
8	0.1986	87.9266
9	0.1275	88.0541
10	0.0823	88.1364
30	0.0034	88.4969
40	0.0016	88.5194

5.4 PCA Results

We project two types of data onto the subspace generated by 2 most significant components which gives us some insight about how data is distributed in the subspace, shown in Fig. 5.7. Also the first 5 components with highest energy are shown in fig. 5.8. The contribution of principal components is shown in table 5. First 10 components contain 88.14% of the total dataset energy.

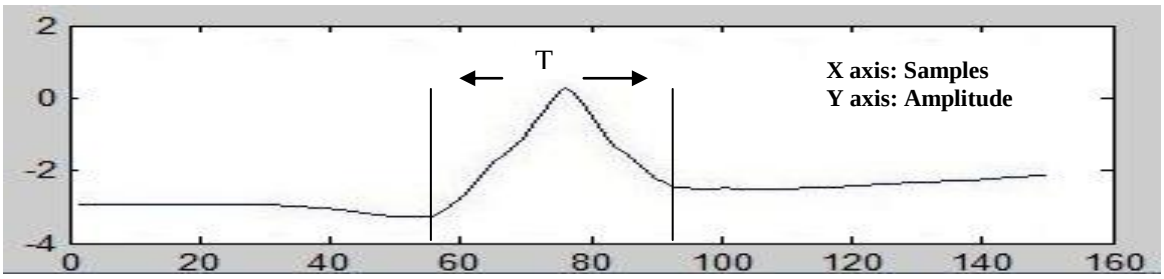


Figure 5.2 Principal Component representing Atrial Premature Beat

The figure above is the first principal component which is showing a Atrial Premature Beat. This beat is having 151 samples. According to the sampling frequency that we have i.e 360 samples per second, the duration of this beat is approximately 419ms. Then accordingly the duration of QRS complex (T as shown in figure) is around 88.96 ms which is less than 120 ms, and the morphology proves the case of Atrial Premature Beats.

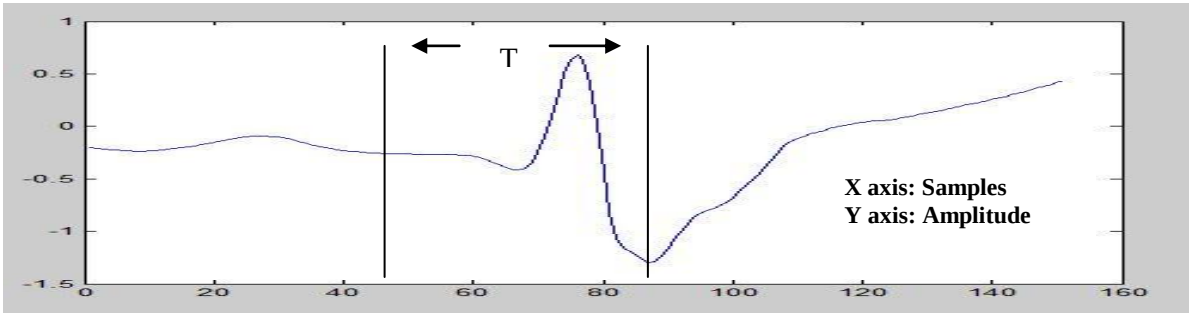


Figure 5.3 Principal Component representing Left Bundle Branch Block Beat

The figure above is the second principal component which is showing a Left Bundle Branch Block Beat. This beat is having 151 samples. According to the sampling frequency that we have i.e 360 samples per second, the duration of this beat is approximately 419ms. Then accordingly the duration of QRS complex (T as shown in figure) is around 125.1 ms which is greater than 120 ms, and the morphology proves the case of Left Bundle Branch Block Beats.

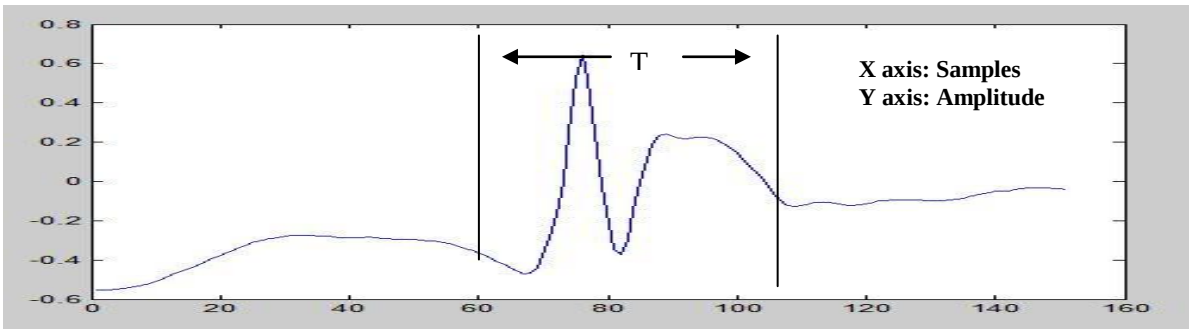


Figure 5.4 Principal Component representing Paced Beat

The figure above is the second principal component which is showing a Paced Beat. This beat is having 151 samples. According to the sampling frequency that we have i.e 360 samples per second, the duration of this beat is approximately 419ms. Then accordingly the duration of T as shown in figure is around 139 ms which is greater than 120 ms, and the morphology proves the case of Paced Beats.

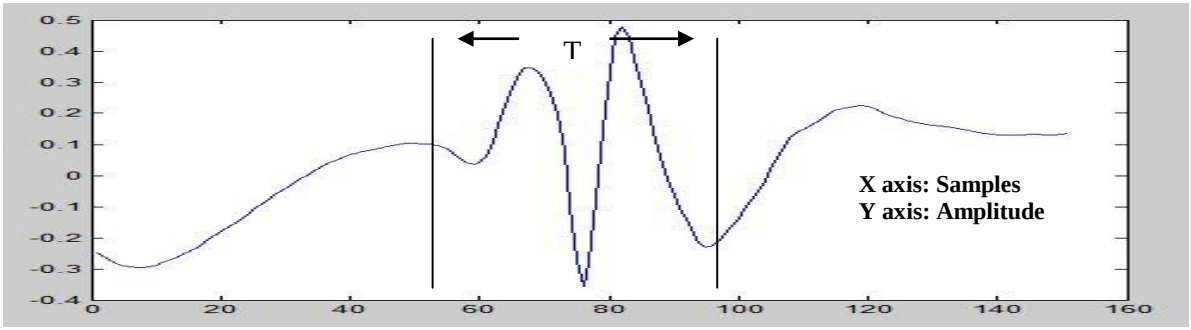


Figure 5.5 Principal Component representing Right Bundle Branch Block Beat

The figure above is the fourth principal component which is showing a beat for Atrial Premature Beat. This beat is having 151 samples. According to the sampling frequency that we have i.e 360 samples per second, the duration of this beat is approximately 419ms. Then accordingly the duration of QRS complex (T as shown in figure) is around 117ms which is almost equal to 120 ms, and the morphology proves the case of Right Bundle Branch Block Beats.

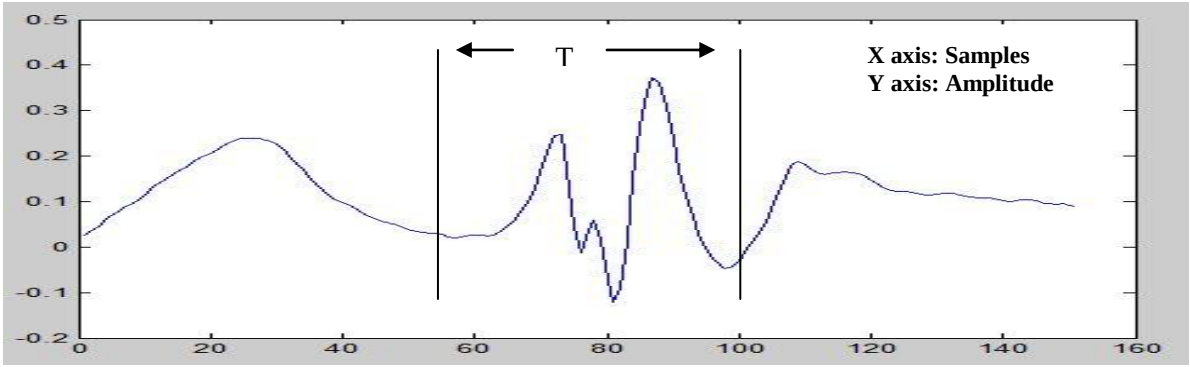
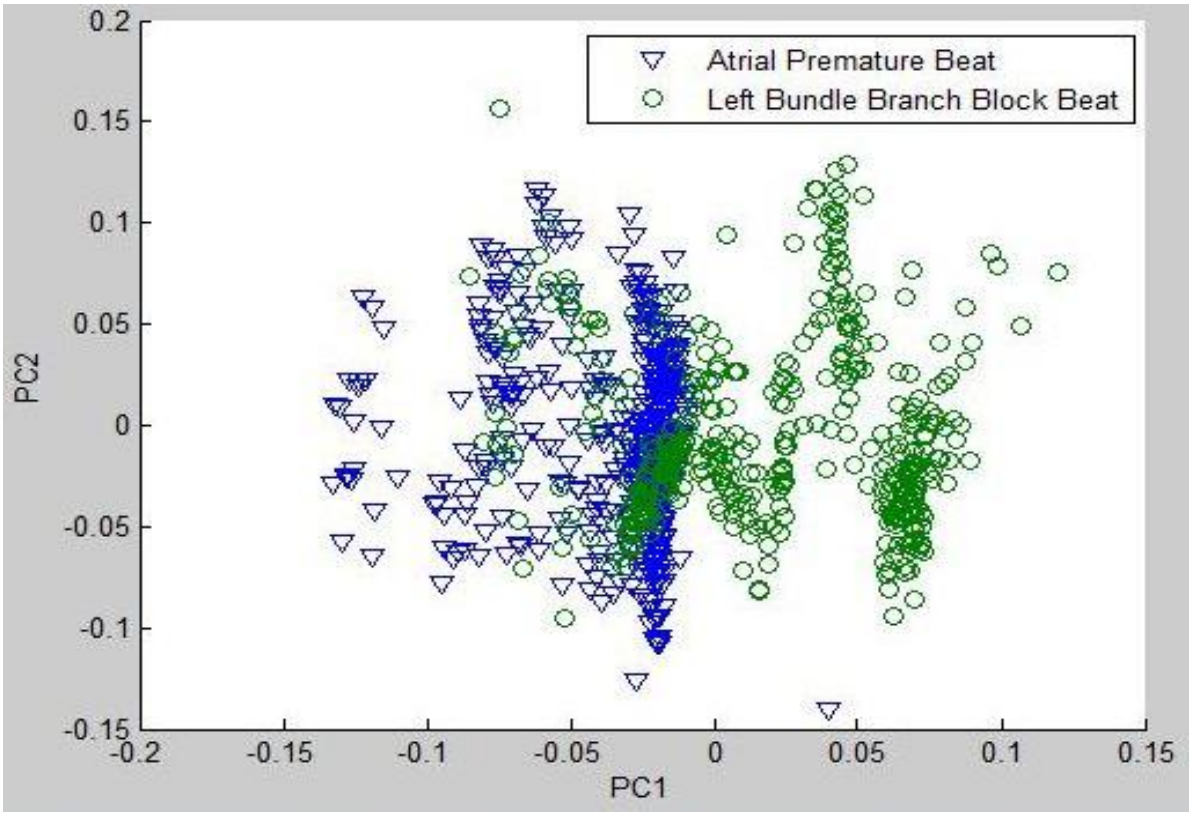


Figure 5.6 Principal Component representing Ventricular Premature Beat

The figure above is the fourth principal component which is showing a beat for Ventricular Premature Beat. This beat is having 151 samples. According to the sampling frequency that we have i.e 360 samples per second, the duration of this beat is approximately 419ms. Then accordingly the duration of QRS complex (T as shown in figure) is around 125.1 ms which is almost equal to 120 ms, and the morphology proves the case of Ventricular Premature Beats.



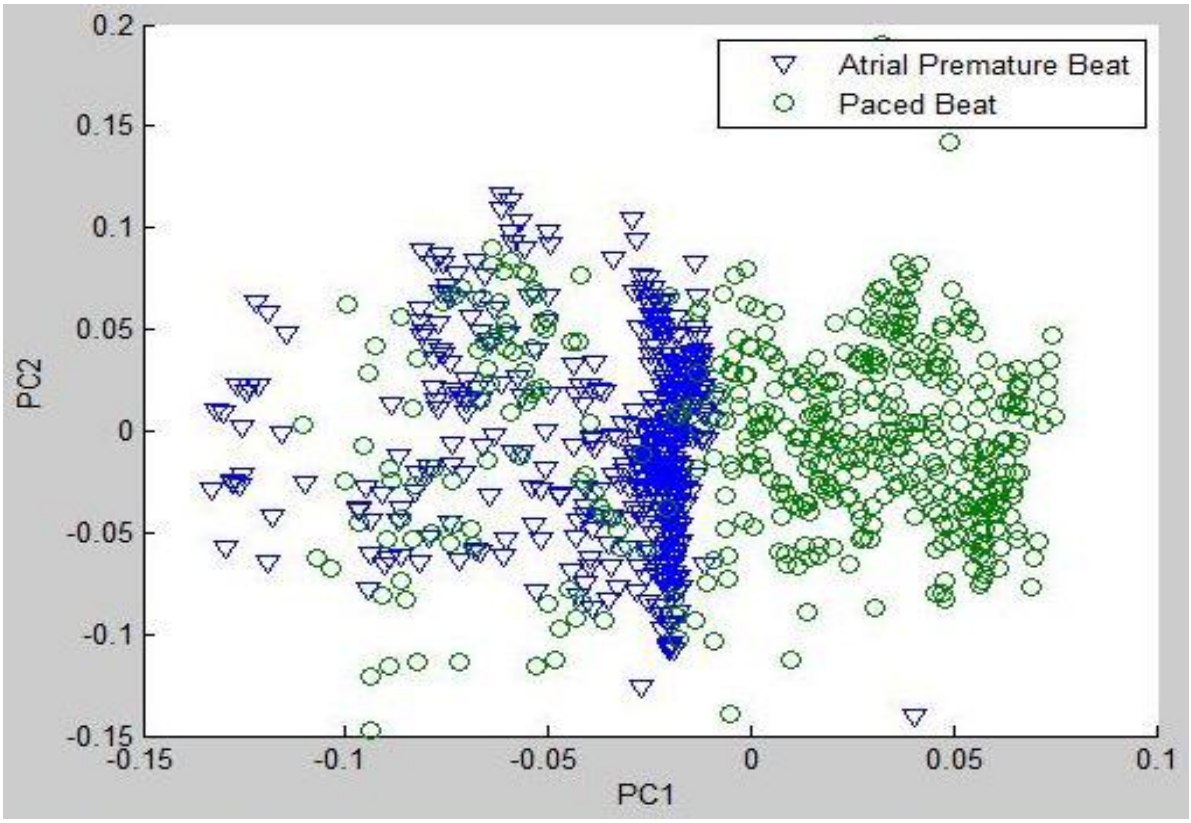


Figure 5.7 Plots of Different beats on two most significant components.

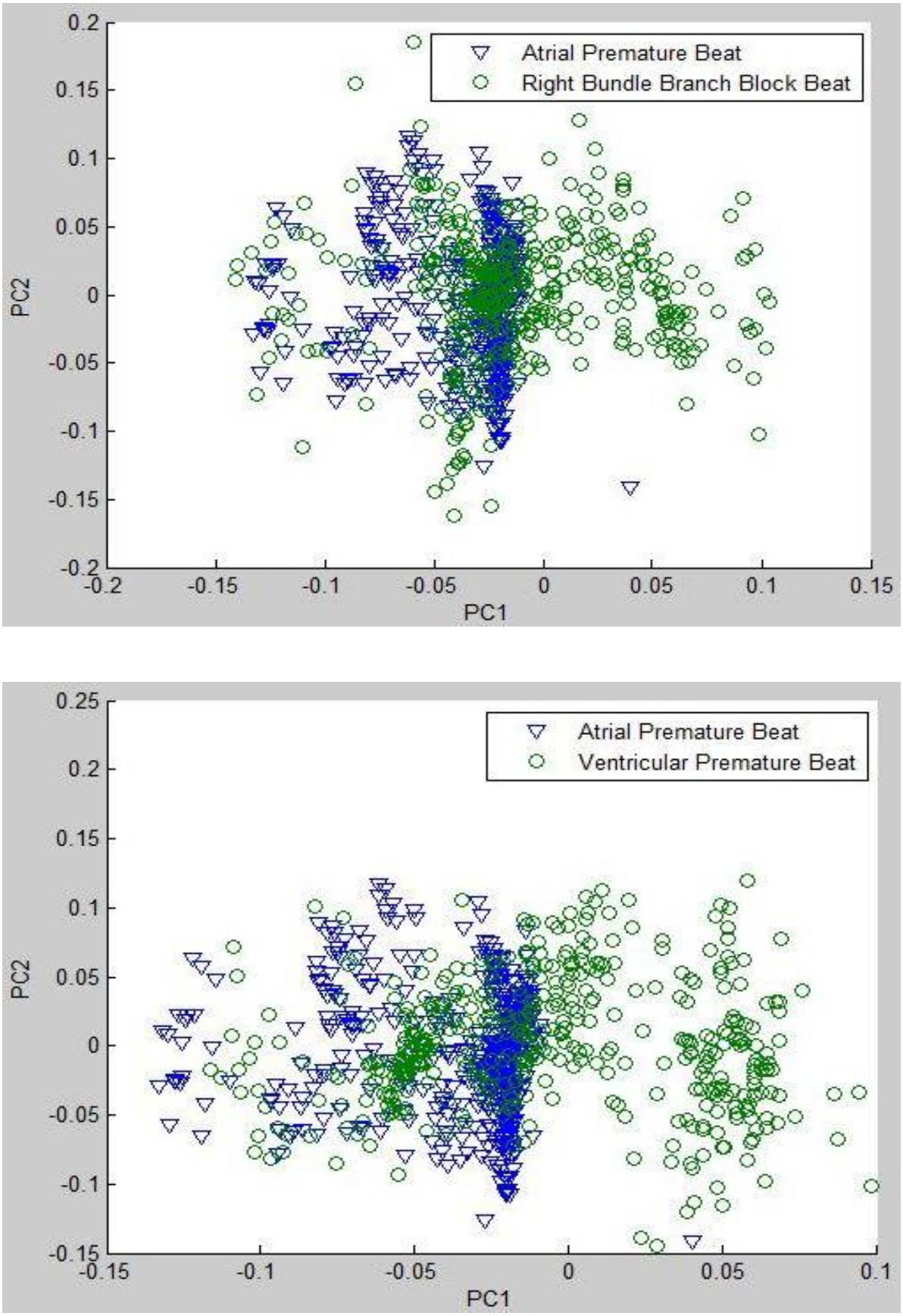


Figure 5.8 Plots of different beats on two most significant components

In figure 5.7, the first plot shows the scatter plot of Left bundle branch block beats(LBB) and Atrial (APB) onto the 2 most significant components, and the second figure shows the scatter plot of Atrial Premature Beat (APB) and Paced beats (PB) onto the 2 most significant components. Similarly, in figure 5.8, the first plot shows the scatter plot of Atrial Premature Beat (APB) and Right bundle branch block beats (RBB) onto the 2 most significant components. And the second figure shows the scatter plot of Atrial Premature beats (APB) and Ventricular Premature beats (VPB) onto the 2 most significant components. We can see that each of these two types of data is loosely concentrated and tend to have a slight hazy classification plane. But the data set is still highly correlated. So we apply the classifiers to accomplish the classification task.

5.5 Classification Results

Classification Analysis is done using several classifiers. Principal Components are different assigned classes according to the arrhythmias. The evaluation method adopted is “Cross-validation” with 10 folds. And “Percentage split” is 66%. It means that the classifier takes 66% of the dataset as training data and rest 33% as test data. Cross validation is done and it gives the accuracy percentage. We use BayesNet Classifier, RBFNetwork Classifier and IBK Classifier. The classification results of the three classifiers used are given in table 6.

Results are shown with 30 and 40 components. We see that using 40 principal components gives approx. 99% classification accuracy in all the three cases.

5.6 Confusion Matrix

The three tables below show the classification results of different classifiers in the form of Confusion Matrix. Tables 5.7, 5.8, 5.9, show the Confusion Matrix for BayesNet Classifier, SVM Classifier and KNN Classifier respectively.

The Table 5.10 summarizes the results obtained by the other algorithms proposed in the literature. Our work is showing high performance in comparison to some other algorithms for ECG arrhythmias classification. the algorithm and the database used by the other related algorithms is also given and the comparison is done on the basis of Accuracy, which is highest in our proposed work.

Table 5.3 Classification Results

Classifier Factors	**30 Components**			**40 Components**		
	KNN	**SVM**	**BayesNet**	**KNN**	**SVM**	**BayesNet**
Total Instances	**755**	**755**	**755**	**755**	**755**	**755**
Correctly Classified Instances	739 **97.88%**	739 **97.88%**	709 **93.90%**	750 **99.34%**	750 **99.34%**	744 **98.54%**
Incorrectly Classified Instances	16 **2.12%**	31 **4.106%**	46 **6.09%**	5 **0.6623%**	5 **0.6623%**	11 **1.457%**
Kappa Statistic	**0.9735**	**0.9487**	**0.9238**	**0.9917**	**0.9917**	**0.9818**
Mean Absolute Error	**0.0108**	**0.0204**	**0.0352**	**0.005**	**0.0029**	**0.0076**
Root Mean Absolute Error	**0.0918**	**0.1185**	**0.1372**	**0.0514**	**0.0495**	**0.0611**
Relative Absolute Error	**3.3601%**	**6.3658%**	**10.9914%**	**1.5522%**	**0.8959%**	**2.3727%**
Root Relative Squared Error	**22.95%**	**29.63%**	**34.30%**	**12.84%**	**12.37%**	**15.27%**

The Accuracy is measured in terms of Correctly Classified instances out of the total instances, as shown in the table the total instances are 755 and out of which the correctly and incorrectly classified are shown. Then some others factors are also calculated like Mean Absolute Error, Root Mean Absolute Error, Relative Absolute Error and Root Relative Squared Error. Also, in tables 5.4, 5.5, 5.6, the detailed accuracy by class is also given. In this table we can see the performance of a classifier in terms of True Positive Rate (TP Rate), False Positive Rate (FP Rate), Precision, Receiver Operating Characteristics (ROC) etc.

Table 5.4 Detailed Accuracy by Class for k-NN

k-NN	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	1	APB
	1	0	1	1	1	1	LBB
	1	0	1	1	1	1	PB
	0.987	0.005	0.987	0.987	0.983	0.991	RBB
	0.98	0.003	0.98	0.98	0.983	0.99	VPB
Weighted Average	0.993	0.002	0.993	0.993	0.993	0.996	

Table 5.5 Detailed Accuracy by Class for BayesNet

BayesNet	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	1	APB
	1	0	1	1	1	1	LBB
	0.993	0.003	0.987	0.993	0.99	1	PB
	0.98	0.012	0.995	0.98	0.967	0.999	RBB
	0.954	0.003	0.986	0.954	0.97	1	VPB
Weighted Average	0.985	0.004	0.986	0.985	0.985	1	

Table 5.6 Detailed Accuracy by Class for SVM

SVM	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	1	APB
	1	0	1	1	1	1	LBB
	0.993	0	1	0.993	0.997	0.998	PB
	0.98	0.003	0.987	0.98	0.983	0.996	RBB
	0.993	0.005	0.98	0.993	0.987	0.997	VPB
Weighted Average	0.993	0.002	0.983	0.993	0.993	0.998	

Table 5.7 BayesNet Confusion Matrix

Classified as →	a	b	c	d	e
APB = a	151	0	0	0	0
RBB = b	0	151	0	0	0
PB = c	0	0	150	1	0
LBB = d	0	0	1	148	2
VPB = e	0	0	1	6	144

Table 5.8 SVM Confusion Matrix

Classified as →	a	b	c	d	e
APB = a	151	0	0	0	0
RBB = b	0	151	0	0	0
PB = c	0	0	150	1	0
LBB = d	0	0	0	148	3
VPB = e	0	0	0	1	150

Table 5.9 KNN Confusion Matrix

Classified as →	a	b	c	d	e
APB = a	151	0	0	0	0
RBB = b	0	151	0	0	0
PB = c	0	0	151	0	0
LBB = d	0	0	0	149	2
VPB = e	0	0	0	3	148

Confusion matrix is another way of representing the classifier result. It gives numbers of correctly and falsely classifier instances out of the total instances. Higher the values of the diagonal and lower the values of the other elements, better is the result. The table 5.7 shows the confusion matrix resulted from classifier-ensemble. The diagonal elements are shown in shaded cells. It is seen from the table that, in some instances non-diagonal elements show very low or zero values – which is always expected from a good classifier.

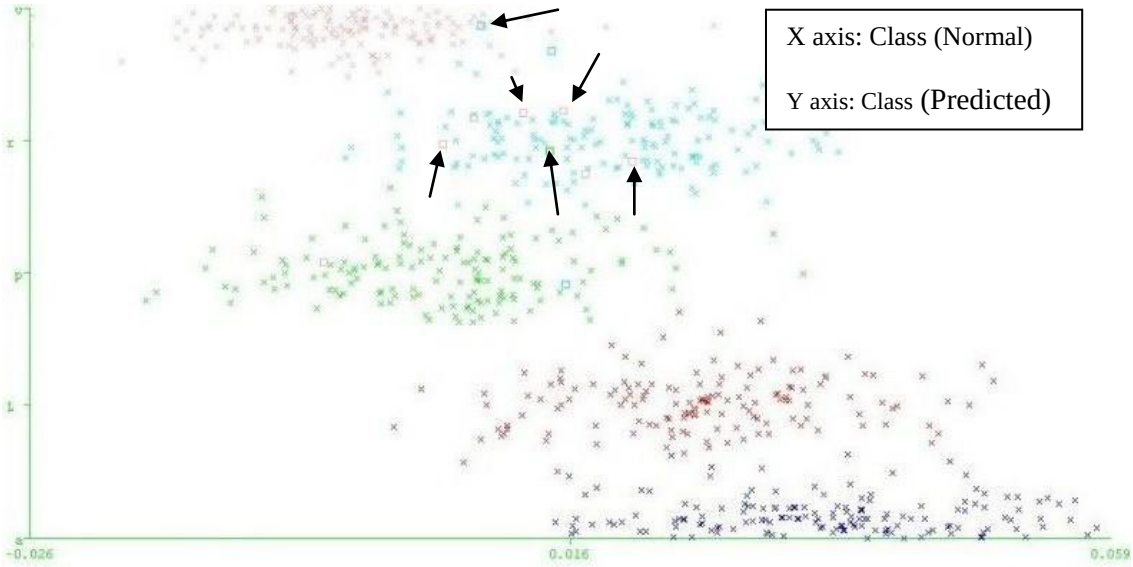


Figure 5.9 BayesNet Classification Plot

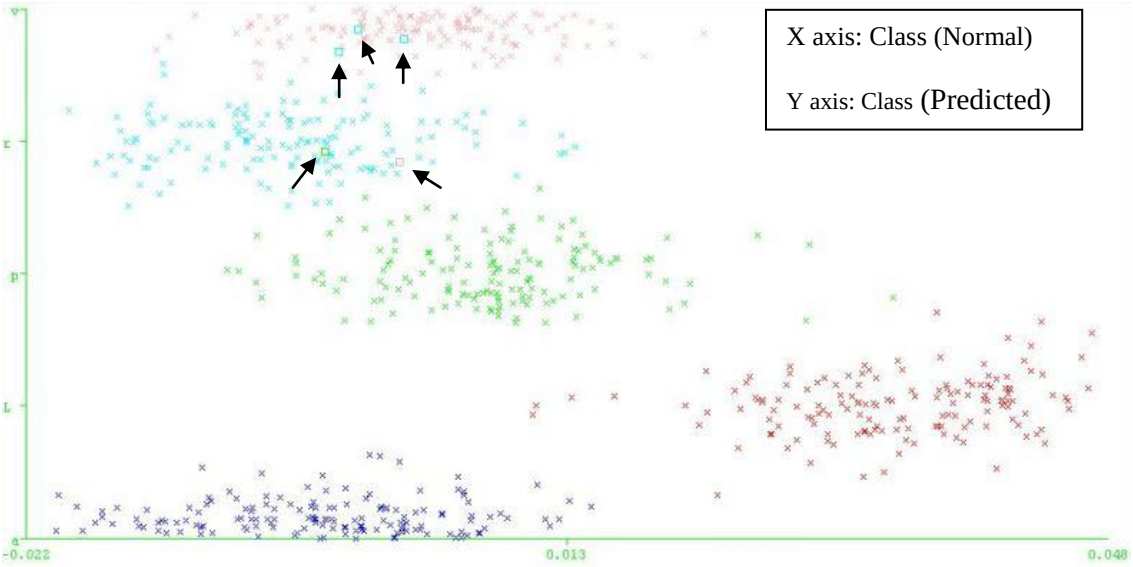


Figure 5.10 SVM Classification Plot

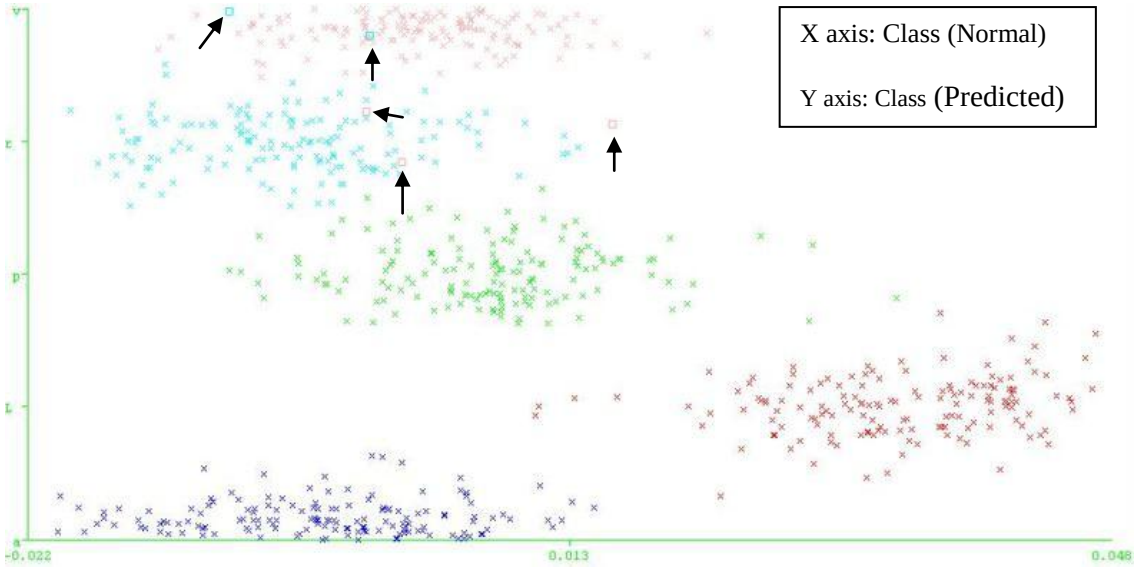


Figure 5.11 k-NN Classification Plot

The plots 5.4, 5.5, 5.6 show the classification result by plotting the data points. The data points with different colours represent the classified beats.

Blue dots represent the data points for Atrial Premature Beat.
Red dots present the data points for Left Bundle Branch Block Beat.
Green dots represent the data points for Paced Beat.
Blue dots represent the data points for Right Bundle Branch Block Beat and
Pink dots represent the data points for Ventricular Premature beat.

In this we can easily see that the highly concentrated data represent the accurately classified data and some falsely classified points. The colour shows their actual class but the misclassified points are shown by arrow. It shows that the data point actually belongs to some other class but as it is falsely classified it falls into some other class.

Table 5.10 Comparison between other related algorithms and our proposed algorithm.

Author,Year	Method	Data Base	Arrhythmia	Accuracy (%)
[11] Roshan Joy Martis *et. al.*, 2012	LS-SVM with RBF kernel	MIT-BIH Arrhythmia	N, RBBB, LBBB, APC, VPC	98.11
[12] Ataollah Ebrahim Zadeh *et. al.*, 2010	Wavelet transform and SVM	MIT-BIH Arrhythmia	N, PVC, Others	97.14
[8] R. Ganesh Kumar *et. al.*, 2012	Random forest Classifier	MIT-BIH Arrhythmia	N, PVC, APB, LBBB, RBBB, PB	92.16
[9] Maedeh Kiani Sarkaleh *et. al.*, 2012	DWT and Neural Network	MIT-BIH Arrhythmia	N, PB, APB	96.50
[10] Abhinav Vishwa *et. al.*,	Machine Learning Technique	MIT-BIH Arrhythmia	Normal & Abnormal	96.77
[38] Nasiri *et. al.*, 2009	SVM & Genetic Algorithm	MIT-BIH Arrhythmia	N, APB, LBBB, VPC	93.00
[39] Jadhav *et. al.*, 2010	MNN Model	UCI Arrhythmia Dataset	Normal & Abnormal	82.22
[40] A. R. Sahab *et. al.*, 2010	Wavelet Transform and Neural Network	MIT-BIH Arrhythmia	N, RBBB, LBBB	97.33
[7] B.Anuradha *et. al.*, 2005-08	Fuzzy Classifiers	MIT-BIH Arrhythmia	LBBB, N, PVC, AF, VF, CHB etc.	93.13
Our Proposed Work	KNN & PCA	MIT-BIH Arrhythmia	APB, LBBB, PB, RBBB, VPB	99.34
	SVM & PCA			99.34
	BayesNet & PCA			98.54

Chapter 6
CONCLUSION

ECG signal can be used as a reliable indicator of heart diseases. It is very important to identify arrhythmia and provide a method of machine pre-diagnosis. Here we have proposed a method of pattern recognition and classification of Electrocardiogram. It consists of two phases: the feature extraction phase and the classification phase. The feature extraction is done with the help of PCA and then three classifiers, SVM, KNN and BayesNet are used in classification phase. In order to overcome the difficulty of intensive computational time taken using several classifiers, attempt has been made to reduce the number of input data points using PCA. Only 40 components are used to train the classifiers and the accuracy percentage of SVM, KNN and BayesNet Classifiers is 99.34%, 99.34% and 98.54% respectively. The proposed structure enhances the performance to recognize and classify ECG beats in terms of faster rate and better accuracy. The good performance of this approach is supported by experimental results on MIT/BIH arrhythmia ECG database and shows potential real-time application in intensive care units or telemedicine networks. This study proves that it is an excellent model for the computer-aided diagnosis of heart diseases based on ECG signals. We have obtained improved results and can be used in practical arrhythmia monitoring systems. Our proposed methodology has immense applications in electronic remote patient monitoring and in intensive care units.

References

[1] Stamkopoulos, T. Diamantaras, K. Maglaveras, "ECG analysis using nonlinear PCA neural networks for ischemia detection", Signal Processing, vol. 46, 1998.

[2] Silipo, R. Marchesi, C., "Artificial neural networks for automatic ECG analysis", Signal Processing, Vol. 46, 1998.

[3] 0. Meste, H. Rix, P. Caminal, and N. Thakor, "Ventricular late potentials characterization in time-frequency domain by means of a wavelet transform", IEEE Trans. Biomed. Eng., vol. 41, 1994.

[4] N. V. Thakor and Z. Yi-Sheng, "Applications of adaptive filtering to ECG analysis: Noise cancellation and arrhythmia detection", IEEE Trans. Biomed. Eng., vol. 38, 1991.

[5] W. Philips, "Adaptive noise removal from biomedical signals using warped polynomials", IEEE Trans. Biomed. Eng., vol. 43, 1996.

[6] Hao Zhang, Li-Qing Zhang, "ECG analysis based on PCA and Support Vector Machines", IEEE, 2005.

[7] B.Anuradha, V.C.Veera Reddy, "Cardiac Arrhythmia Classification Using Fuzzy Classifiers", Journal of Theoretical and Applied Information Technology, 2005-2008.

[8] R. Ganesh Kumar, Dr. Y S Kumaraswamy, "Investigating Cardiac Arrhythmia in ECG using Random Forest Classification", International Journal of Computer Applications, 2012.

[9] Maedeh Kiani Sarkaleh, Asadollah Shahbahrami, "Classification Of ECG Arrhythmias Using Discrete Wavelet Transform And Neural Networks", International Journal of Computer Science, Engineering and Applications (IJCSEA), 2012.

[10] Abhinav Vishwa, Mohit K. Lal, Sharad Dixit, Dr. Pritish Vardwaj, "Clasification Of Arrhythmic ECG Data Using Machine Learning Techniques", International Journal of Artificial Intelligence and Interactive Multimedia.

[11] Roshan Joy Martis, U. Rajendra Acharya, K.M. Mandana, A.K. Ray, Chandan Chakraborty, "Application of principal component analysis to ECG signals for automated diagnosis of cardiac health", Expert Systems with Applications, Elsevier, 2012.

[12] Ataollah Ebrahim Zadeh, Ali Khazaee, Vahid Ranaee, "Classification of the electrocardiogram signals using supervised classifiers and efficient features", computer methods and programs in biomedicine, Elsevier, 2010.

[13] Sung-Nien Yu, Ying-Hsiang Chen, “Electrocardiogram beat classification based on wavelet transformation and probabilistic neural network”, Pattern recognition letters, Elsevier, 2008.

[14] Peng Li, Kap Luk Chan, Y. W. Chan, “A Mixed SVM-Based Hierarchical Learning Approach for Abnormal ECG Beat Recognition”.

[15] S. S. Mehta, and N. S. Lingayat, “Support Vector Machine for Cardiac Beat Detection in Single Lead Electrocardiogram”, IAENG International Journal of Applied Mathematics, 2007.

[16] Mi Hye Song, Jeon Lee, Sung Pil Cho, Kyoung Joung Lee, Sun Kook Yoo, “Support Vector Machine Based Arrhythmia Classification Using Reduced Features”, International Journal of Control, Automation, and Systems, 2005.

[17] Dayong Gao, Michael Madden, Des Chambers, and Gerard Lyons, “Bayesian ANN Classifier for ECG Arrhythmia Diagnostic System: A Comparison Study”.

[18] Dipti Patra, Manab Kumar Das, Smita Pradhan, “Integration of FCM, PCA and Neural Networks for Classification of ECG Arrhythmias”, IAENG International Journal of Computer Science, 2010.

[19] Ye Wenyu, Li Gang, Lin Ling, Yu Qilian, “ECG Analysis Based On PCA And Som”, Int. Conf. Neural Networks 8 Signal Processing, IEEE, 2003.

[20] Francisco Castells, Pablo Laguna, Leif Sornmo, Andreas Bollmann, and Jos´eMillet Roig, “Principal Component Analysis in ECG Signal Processing”, EURASIP Journal on Advances in Signal Processing, 2007.

[21] Mehmet Korürek, Ali Nizam, “Clustering MIT–BIH arrhythmias with Ant Colony Optimization using time domain and PCA compressed wavelet coefficients”, Digital Signal Processing, Elsevier, 2009.

[22] Mandeep Singh, “Introduction to Biomedical Instrumentation”,PHI Learning Pvt. Ltd., 2010.

[23] Lindsay I Smith “A tutorial on Principal Components Analysis”, 2002.

[24] Cooley, W.W. and Lohnes, P.R. “Multivariate Data Analysis”, John Wiley & Sons, Inc., New York, 1971.

[25] Herve Abdi ¢ Lynne J. Williams “Principal Component Analysis”, Wiley Interdisciplinary Reviews: Computational Statistics, 2011.

[26] I . A. Ismail,Professor, M.A. Ramadan, T. El danf,lecturer, A. H. Samak,lecturer, "Automatic Signature Recognition And Verification Using Principal Components Analysis", IEEE, 2008.

[27] Fengxi Song, Zhongwei Guo, Dayong Mei, "Feature selection using principal component analysis", IEEE, 2010

[28] Khaled Labib and V. Rao Vemuri, "An Application of Principal Component Analysis to the Detection and Visualization of Computer Network Attacks", IEEE.

[29] Mohammad Asmatullah Khan, Aurangzeb Khan, Tariq Mahmood, Muzahir Abbas, Nazir Muhammad, "Fingerprint Image Enhancement Using Principal Component Analysis (PCA) Filters", IEEE, 2010.

[31] Alisher Kholmatov, Berrin Yanikoglu Sabanci, "Identity authentication using improved online signature verification method", ELSEVIER, 2005.

[32] Dimitri Pissarenko, "Eigenface-based facial recognition", IEEE, 2002.

[33] J.S. Taur, W. Tao, "Medical Image Compression Using Principal Component Analysis", IEEE, 1996.

[34] Stu Daultrey, "Principal Component Analysis", Catmog 8, University College Dublin, 1976.

[35] Lindsay I Smith, "A tutorial on principal component analysis", Derivation, Discussion and Singular Value Decomposition", Version 1, 2003.

[36] "MIT/BIH database distributor", Beth Israel Hospital, Biomedical Engineering, Division, http://ecg.mit.edu/, USA, 1979.

[37] MIT-BIH Arrhythmia Database. www.physionet.org/physiobank/database/mitdb/.

[38] Nasiri, J.A., Naghibzadeh, M. ,Naghibzadeh, B. , "ECG Arrhythmia Classification with Support Vector Machines and Genetic Algorithm", Computer Modeling and Simulation, 2009.

[39] Jadhav, Nalbalwar, Ghatol, "ECG arrhythmia classification using modular neural network model", Biomedical Engineering and Sciences (IECBES), IEEE, 2010.

[40] A. R. Sahab, Y. Mehrzad Gilmalek, "ECG arrhythmias classification using wavelet transform and neural networks", MMES'10 Proceedings of the 2010 international conference on Mathematical models for engineering science, 2010.