BACHELOR OF SCIENCE IN
COMPUTER SCIENCE AND ENGINEERING

Inspiring Excellence

Machine Learning Approach for ECG Analysis and predicting different heart diseases

Authors

Sushmita Roy Tithi
Afifa Aktar
Fahimul Aleem

Supervisor
Dr. Amitabha Chakrabarty
Associate Professor
Department of CSE

A thesis submitted to the Department of CSE
in partial fulfillment of the requirements for the degree of
B.Sc. Engineering in CSE

Department of Computer Science and Engineering
BRAC University, Dhaka - 1212, Bangladesh

December 2018

q

Declaration

We, hereby declare that this thesis is based on results we have found ourselves. Materials of work from researchers conducted by others are mentioned in references. This thesis report or any part of it has not been submitted elsewhere for the award of any Degree.

Authors:

Sushmita Roy Tithi
Student ID: 14201051

Afifa Aktar
Student ID: 15101015

Fahimul Aleem
Student ID: 15101126

Supervisor:

Dr. Amitabha Chakrabarty
Associate Professor, Department of CSE
BRAC University

December 2018

The thesis titled Machine Learning Approach for ECG Analysis and predicting different heart diseases
Submitted by:
Sushmita Roy Tithi Student ID: 14201051
Afifa Aktar Student ID: 15101015
Fahimul Aleem Student ID: 15101126
of Academic Year 2018 has been found as satisfactory and accepted as partial fulfillment of the requirement for the Degree of Computer Science And Engineering

1. Dr. Amitabha Chakrabarty
Associate Professor
Department of CSE
BRAC University

2. Dr. Md. Abdul Mottalib
Professor and Chairperson
Department of CSE
BRAC University

Acknowledgements

It is a matter of joy that we had a successful investigation for our thesis research. This would not have been possible without the support and co-operation of some people. Therefore, we want to show our gratitude and sincerest appreciation to the following people.

In the first place, we want to thank our supervisor, Dr. Amitabha Chakrabarty, for showering us with his knowledge and guiding us with patience throughout our entire research. Then we want to thank our friend Mesbahul Islam Siam (final year student at Dhaka National Medical College) for his constant support throughout our research. In addition, we want to thank our parents for the financial assistance and motivating us during this entire paper. Finally, we want to thank our team members, fellow class mates and friends for their valuable opinions and ideas. All in all, this token of gratitude is for everyone who helped us directly and indirectly in this paper.

Abstract

In the modern world, there have been some revolutionary advancement in the field of medical science and research and this is no different for electrocardiogram. Electrocardiogram (also abbreviated as ECG) illustrates the electrical activity of one's heart over a period of time. Over the years, number of people suffering from heart disease have increased to some extent. Therefore, in our research, we aim to design a model using supervised machine learning that can find anomalies in one's ECG report by analyzing it. We have applied six supervised machine learning algorithms to distinguish between normal and abnormal ECG. In addition, we used them to predict the chances of a patient suffering from a certain disease. We divided our data set into two parts. 75 percent data in one group for training the model and rest 25 percent data in another group for testing. To avoid any kind of anomalies or repetitions, Cross Validation and Random Train-Test Split was used to obtain an answer as accurate as possible. We have compared the results with each other for a better understanding.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

ACC	Accuracy
ANN	Artificial Neural Network
AV	Atrioventricular
CAD	Coronary Artery Disease
CV	Cross Validation
DT	Decision Tree
ECG	Electrocardiogram
FN	False Negative
FP	False Positive
LR	Logistic Regression
MI	Myocardial Infarction
NB	Naïve Bayes
N	Negative
NN	Nearest Neighbour
NPV	Negative Predictive Value
P	Positive
PPV	Positive Predictive Value

PVC Ventricular Premature Contraction

RBBB Right Bundle Branch Block

ROC Receiver operating characteristics

SA Sinoatrial

SB Sinus Bradycardy

ST Sinus Tachycardy

SVM Support Vector Machine

TNR True Negative Rate

TN True Negative

TPR True Positive Rate

TP True Positive

UCI University of California, Irvine

Chapter 1

Introduction

In this era, the 21st century, almost every other person is dependent on technology. This remarkable development of technology has made life better in many possible ways. Number of people suffering from heart disease is increasing day by day mostly because of the unhealthy lifestyle led by people. So data acquisition methods have been developed for the past decades to accurately sense, collect, record, and analyze the patient's' physical condition [2].

Machine learning allows systems the ability to automatically learn and improve from experience without programming explicitly. This is an application of artificial intelligence. Machine learning methods are often classified as supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning [25]. Supervised learning algorithms figure out model relationships and dependencies between the target prediction output and the input features [15]. So we can predict the output values for new data based on those relationships which it learned from the previous data sets. For unsupervised learning the model is trained with unlabeled data. Semi-supervised learning uses both labeled and unlabeled data, so it falls between supervised and unsupervised learning [16]. Reinforcement machine learning algorithms interacts with its environment by producing actions and then discovers errors or rewards. Machines and software agents are allowed to automatically determine the ideal behavior within a specific context in order to maximize its performance. We have trained our model by supervised machine learning.

ECG provides us with series of sinus rhythm which defines the condition of heart. ECG is useful for detecting certain types of conditions and because of its low cost, it is the most common test for screening heart diseases [29]. An ECG telemonitoring system supports patients from avoiding frequent visit at hospitals and allows monitoring ECG patients. ECG signals of patients can be useful to identify different type of heart problems. Wireless health care services provide remote monitoring diagnosis to patient every now and then. At the

present time, these services evolved tremendously that it is not only restricted to give services for the cognitive and physical disabilities patients but also it can also maintain high-quality services for the newborn and young children [2]. Furthermore, ECG pattern recognition is often useful as an early warning system for heart diseases. So, we plan about finding anomalies in Heart Rate of ECG reports and figure out which algorithm gives better and reliable result for some particular heart disease.

1.1 Motivation

Often the doctors and medical staffs face problem in interpreting an ECG report. Very little change in any section of the ECG graph can result in different kind of diseases. During medical emergencies, like in ER or ICU, where time is of the essence, it would be more advantageous to find out what is ailing the patient for immediate treatment [29]. Moreover, it is very difficult for a doctor to read an ECG report with bare eyes. At times, there is high chance to miss out any abnormality in the ECG report as the change in the ECG wave shape is hardly noticeable. So doctors often deny to conclude any disease from the ECG report of a patient with 100 percent accuracy until they conduct some more tests for the patient.

Recently, in a hospital at Dhaka, a patient with extreme chest pain was admitted. After doing electrocardiogram, doctors concluded that this was a mere gastric pain from his ECG report and discharged him. No less than 2 hours, the dead body of the patient arrived back in the same hospital. When his ECG report was re-analyzed, it was seen that there was a slight distortion in one of his ECG graph which was an indication of an heart attack. This was so small that the doctors didnot notice at the first place.

With the recent development in Machine Learning field, the scope of performing in different sectors and concluding with better accuracy and optimized performance has increased [29]. Medical science has also improved over time. Considering all these factors, we decided to do our thesis using machine learning technique to distinguish between normal and abnormal heart rate. In addition, we have also done our research on some diseases that occur due to change in any section of the ECG graph.

1.2 Objective

Our initial idea was to distinguish between normal and abnormal ECG using different machine learning algorithms. Afterwards, we also did the research on predicting few diseases depending on the availability of the data for a particular disease. The model that we trained to predict the outcome for different diseases and classify between normal and abnormal ECG

can also be used to predict outcome for other diseases that we did not work for. In addition, our target was to figure out which algorithm gives the best result in predicting the diseases we worked with. We analyzed the results that we predicted by using our model and calculated the accuracy level of different algorithms for our selected diseases.

1.3 Thesis Outline

The rest of the dissertation is organized as follows:

Chapter 2 contains the literature review. It is divided into separate sections which has more sub-sections. In the section 2.1, there is details about the background study in this field. Section 2.1 is divided into 3 subsections where subsection 2.1.1 contains details about the structure of the heart with the diagram of a human heart, subsection 2.1.2 contains information about the cardiac cycle, that is how blood flows inside a human heart and the changes that take place in the cardiac muscle during this time, with a diagram of the phases of the cardiac system and subsection 2.1.3 contains details about the process of a heartbeat with another diagram to explain it better. In the Section 2.2, there is details about the cardiac background that includes description about the ECG, how the 12 leads are connected to the body to do ECG and also the duration table for each section in the ECG PQRST graph. In section 2.3, there is information about related works in this field.

Chapter 3 has information about our data set and methods that we used in our research. Section 3.1 contains details about the data set that we worked with, its instances and attributes. Subsection 3.1.1 holds more information about data refining and categories. Section 3.2 contains details about the algorithms that we have implemented to analyze our data. There are 6 subsections in section 3.2 each containing information about the six different algorithms we have implemented including their mathematical formula and diagrams if available. section 3.3 has details about the confusion matrix. Lastly section 3.4 contains our workflow diagram and its explanation.

Chapter 4 contains details about the diseases we analyzed and our experimental setup. we have analyzed 5 diseases, and they are explained under section 4.1 in different subsections for each diseases. In section 4.2, we have discussed about how we have used our data to train and test our model. This also have two sub-sections, each defining the different techniques.

In chapter 5, there is information about the results we found and discussion on them. We have also compared our results to get a better overview. Section 5.1 contains result about classification between normal and abnormal heartbeat using different algorithms that we obtained from our research . Section 5.2 contains the results for individual diseases and algorithms in tabular form. Furthermore, we also analyzed for prediction of disease with two

different input data set. Both the outputs are explained in Section 5.3 and 5.4 respectively, each containing results about the prediction of different diseases by the algorithms we implemented. The results are shown in a pie-chart for each algorithm and also explained under different subsections. In section 5.5, the results we found are analyzed and lastly, there is a comparison between previous works and our result. There are graphical representations throughout this section for a better understanding of the results obtained.

Chapter 6 contains details about the conclusion and future works that can be done under this topic.

At the end, all the references that we used are given.

Chapter 2

Literature Review

Machine learning is the study of making computers act using the knowledge gathered from historical data without being expressly modified. This chapter contain details about the background study, structure of heart and its cardiac cycle, how a heartbeat occurs, cardiac background and how to read an ECG graph. Furthermore it also has information about the related past works.

2.1 Background Study

The general objective of an ECG is to recognize the condition and function of the heart. The ECG has been found as the most used and uncomplicated way for precise and rapid diagnosis of numerous cardiovascular incidents [29]. In the United States, the United States Preventive Services Task Force does not prescribe an ECG for routine screening in patients without any indications and those who has lower chances to have coronary heart disease [48]. This is on the grounds that an ECG may indicate an erroneous presence of an issue, prompting misdiagnosis, the proposal of obtrusive methodology, or over-treatment. To avoid such predicament machine learning can be used. Over the years, a large amount of raw data that has been gathered that can be of use. Data acquisition methods have been developed for the past decades to accurately sense, collect, record, and analyze the patient's' physical condition [2]. To an ever increasing extent, information is being gathered and put away in databases or then again spreadsheets [43]. Machine learning is the perfect approach to find beforehand obscure regularities and patterns from different data sets, with the expectation that machines can help in the frequently repetitive and mistake inclined procedure of getting information from experimental information, and help individuals to clarify and arrange their perception [43]. In our thesis, we also did experiment to predict the chances of a certain heart disease for a patient to experience. We chose five heart diseases for this which are

Ischemic changes (Coronary Artery Disease), Myocardial Infarction, Sinus tachycardy, Sinus bradycardy, Right bundle branch block.

2.1.1 The structure of the heart

The heart is a hollow muscular organ that lies in the middle of the chest cavity and is surrounded by a double membrane known as the pericardium. This protects the heart and facilitates its pumping action. The heart is responsible for controlling the circulatory system in mammals and other animals. The heart is divided into four chambers internally. Two Atria (singular : atrium) and two ventricles. The upper chambers are the atria (left atrium and right atrium) and the lower chambers are the ventricles (left ventricle and right ventricle) [1][49]. Muscular walls (septum) divide two sides of the heart. Each atrium is connected to its ventricle by a opening that is guarded by a valve (the bicuspid valve on the left and the tricuspid valve on the right). The bicuspid valve which is also known as mitral valve is made up of two flaps and the tricuspid valve has three flaps. These valves are also known as atrioventricular valves and they prevent backflow of blood into the atria. All the four chambers have a smooth membranous lining internally known as the endocardium. The walls of the heart are made up of cardiac muscles which is unlike any other muscle in our bodies. It never gets fatigue like skeletal muscle. The cardiac muscle fibres contract and then relax again about 70 times a minute on average.

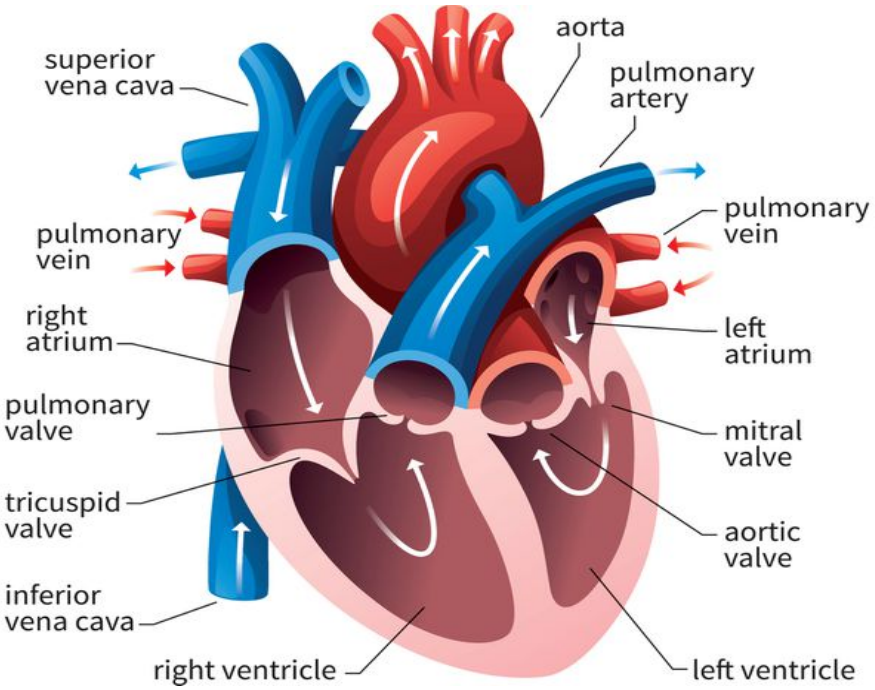

Fig. 2.1 The Human Heart [42]

The heart is attached to four blood vessels which are Aorta, Vena cava, Pulmonary Artery and Pulmonary vein. The vena cava is connected to the right atrium and pulmonary vein is connected to the left atrium. Similarly, the aorta is connected to the left ventricle and pulmonary artery is connected to the right ventricle. The ventricles are separated from the aorta and pulmonary artery by means of semilunar valves. These valves are also responsible for preventing backflow of blood into the heart [22][50][45]. The atrial walls are thinner than the wall of the ventricles since ventricles need to withstand more blood pressure.The vena cava supplies de-oxygenated blood from the body, which then flows into the right atrium then the right ventricle. This gets pumped through the pulmonary artery to the lungs where it gets oxygenated, before returning to the heart via the pulmonary vein [22][50][49][45]. This flows through the left atrium into the left ventricle, and then gets pumped to the body via the aorta. It finally returns to the heart through the vena cava, and the process repeats [22] . Fig 2.1 shows a Human Heart.

2.1.2 The cardiac cycle

The human heart is a pump and it pumps blood around the body at different speeds and at different pressures according to the need of the body. Blood is moved through the heart by a series of contractions and relaxations of the muscle in the walls of the four chambers. The contraction of the heart is called ‘systole’ and the relaxation is called ‘diastole’. The contraction and relaxation together constitute the heartbeat. Cardiac cycle is formed by the cyclical repetition of contraction and relaxations of these muscles and consists of three main stages which are atrial systole, ventricular systole and diastole. The cardiac cycle is explained briefly below.

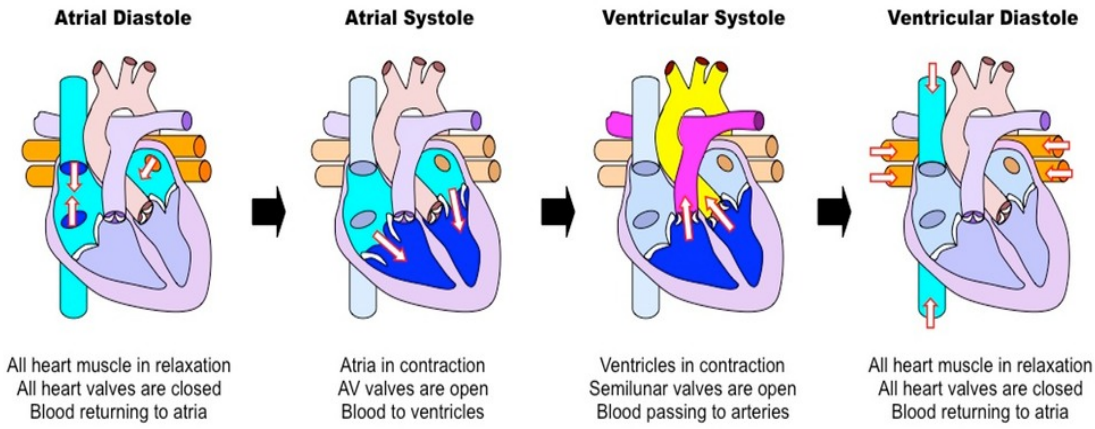

Fig. 2.2 The phases of a cardiac cycle [11]

As blood enters the atria, it can not pass into the ventricles as the atrioventricular valves are closed and during atrial systole, the walls of the atria contracts and this raises the blood pressure in the atria which forces the atrioventricular valves to open, then the blood then passes to the ventricles through these valves [50]. During ventricular systole, the ventricles contract when they are full and this raises the blood pressure in the ventricles which closes the atrioventricular valves and thus preventing any backflow of blood to the atria, therefore, The pressure continues to increase as the ventricle contracts and this forces open the semilunar valves and blood travels to the two arteries [50]. The pulmonary artery carries blood to the lungs and aorta has several branches that carries blood to all other parts of the body [50][1]. As blood leaves the ventricles, the semilunar valves are closed due to the increase of pressure in the pulmonary artery and aorta. This happens during the diastole. The cycle then begins again as the atria start to fill with blood. Fig 2.2 is an image of the phases of the cardiac cycle.

2.1.3 The process of a Heartbeat

A heartbeat begins in a region of tissue known as the sinoatrial node (SA node) that is a specialized bundle of thin cardiac muscular fibers located above the right atrium. SA node is also known as the heart's natural pacemaker. It generates electrical pulses regularly over the two atria which are then stimulated to contract as a part of atrial systole. When the electrical impulses reach the atrioventricular node (AV node), the impulses spread rapidly through special fibers from the interventricular septum to the walls of both ventricles, where the muscles are stimulated to contract [50].

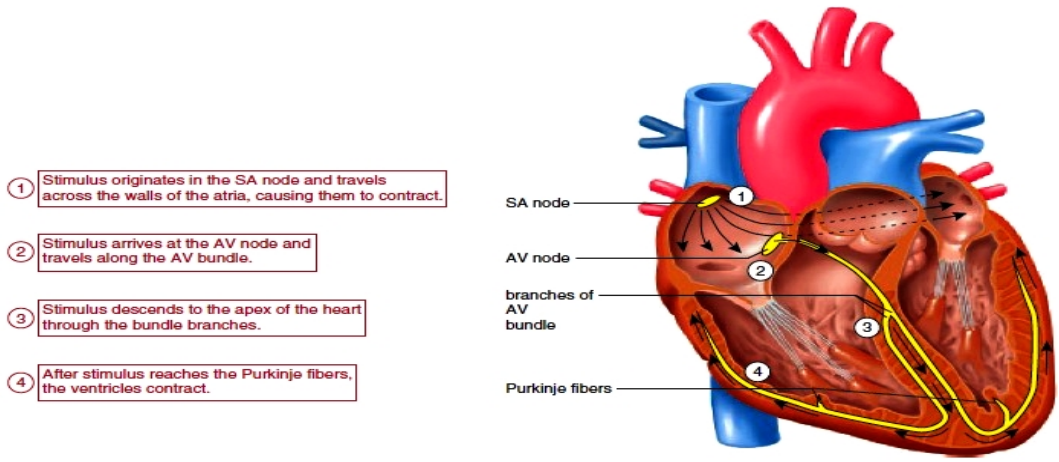

Fig. 2.3 Location of SA node and AV node [23]

The AV node is a cluster of cells in the center of the heart between the atria and ventricles,

and acts like a gate that slows the electrical signal before it enters the ventricles. This delay gives the atria time to contract before the ventricles do [13]. The waves then travels down the inter-ventricular wall through special conducting tissue known as the Purkyne (also known as His-Purkinje Network) which is a pathway of fibers that sends the impulse to the muscular walls of the ventricles and causes them to contract. All of the Purkinje fibres together are known as the Bundle of His [1]. Fig 2.3 shows the location of SA node and AV node.

2.2 Cardiac Background

Number of people suffering from heart disease is increasing day by day mostly because of the unhealthy lifestyle led by people. ECG provides us with series of sinus rhythm which defines the condition of heart.

An Electrocardiogram (ECG) is a procedure that shows the electrical activity of the human heart over a period of time [1]. It consists of several sensors placed around the body connected to a monitor. The electrical signals are recorded by the device, by attaching electrodes to the outer surface of the skin. Then the electrodes detect the small electrical changes on the skin that arise from the heart muscle's electrophysiological pattern of depolarizing and repolarizing during each heartbeat [1].

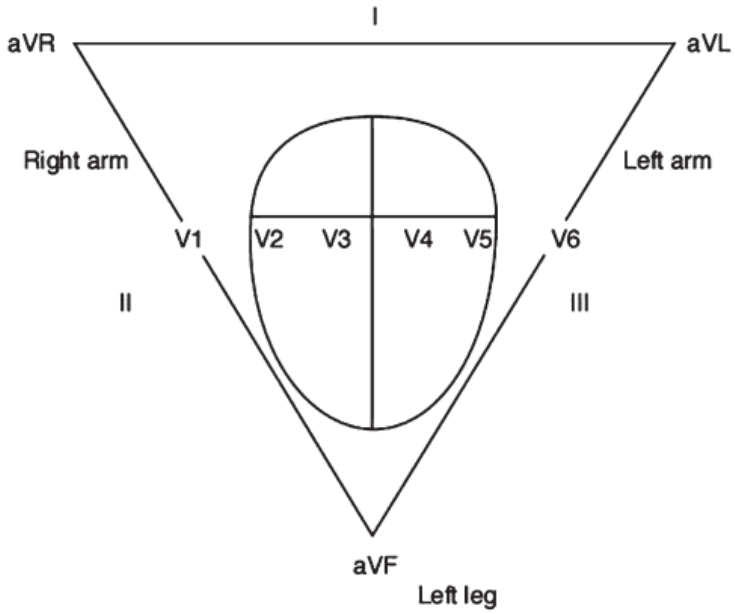

Fig. 2.4 Schematic representation of 12 leads (electrodes) placed around the heart [49]

The electrical signals travel through the electrodes to the graph device, which records them as characteristic waves. Different waves reflect the activity of different areas of the heart which generate the respective flowing electrical currents [29].

The electrocardiogram comprises of 12 leads around the heart. For scientific purposes, the heart is at the focal point of a triangle (fig. 2.4) [1][49]. The cathode situations are assigned as follows:The three appendage drives: lead I joins the privilege and left arms, lead II associates the correct arm and left leg and lead III joins the left arm and left leg [49]. The three enlarged leads: aVL is situated confronting the heart from the correct arm VL from the left arm and aVF from the left foot and these cathodes are set in a frontal plane [1][49]. The precordial leads (V1– V6): these are put on the front of the thorax and record even driving forces [1][49].

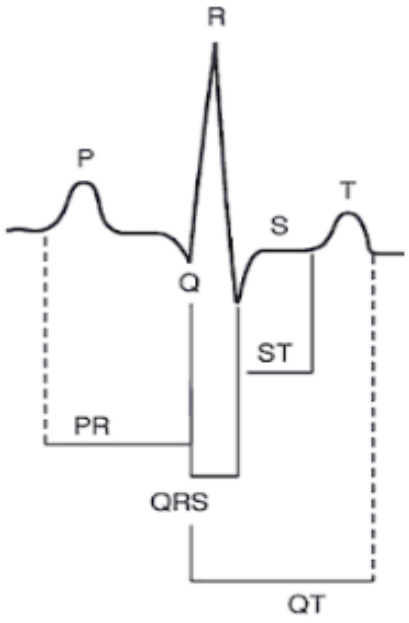

Fig. 2.5 Normal ECG recording [49]

An ECG cycle consists of 5 waves which are PQRST (fig. 2.5). The P wave corresponds to the atrial depolarization and the pumping of blood from the atrium to the ventricle. The QRS complex usually consists of Q, R and S waves which corresponds to the depolarization of blood and the pumping of blood out from the ventricle to the body and lung. Both the P wave and the QRS complex are depolarization waves. Finally the T wave corresponds to the repolarization of the ventricle and the recovery of the ventricle for next cycle [1][49].

Table 2.1 ECG Features and their normal duration [18]

Feature	Description	Duration
RR	interval between R wave and the next R wave	0.6-1.2 s
P	first short upward movement of the ECG tracing	80ms
PR	measured from the beginning of the P wave to the beginning of the QRS complex	120-200 ms
QRS	normally begins with a downward deflection Q, a larger upwards deflection R and ends with a downward S wave	80-120 ms
PR	connects the P wave and the QRS complex	50-120 ms
J-point	The point at which the QRS complex finishes and the ST segment begins is called J-point.	Not applicable
ST	connects the QRS complex and the T wave	80-120 ms
T	normally a modest upward waveform	160 ms
ST	measured from the J point to the end of the T wave	320 ms
QT	measured from the beginning of the QRS complex to the end of the T wave	420 ms

One ECG signal consists of several ECG beats and each ECG beat contains P wave, QRS complex, and T wave. Each peak (P, Q, R, S and T), intervals (PR, RR, QRS, ST, and QT) and segments (PR and ST) of ECG signals have their normal amplitude or duration values [18]. These peaks, intervals, and segments are called ECG features. All these features for one ECG cardiac cycle are described in Table 2.1.

2.3 Related Works

We live in an era where computation is being moved from vast centralized servers to PCs and cloud [34]. For a long time researchers have been working on identifying and predicting different diseases using machine learning. Exploration supervised learning, unsupervised learning and reinforcement learning, which is better for machine learning are analyzed in [16], [15] and [25]. There has been much work in the field of classification and most work has been based on neural networks, Markov chain models and support vector machines (SVMs) [43]. In [14] a comparison between three different machine learning algorithms were discussed. It has been done by many to improve past algorithm or create new one for machine learning. In [15] a new algorithm VF15 was developed to classify arrhythmia. In [18] a definite review of preprocessing strategies, ECG databases, highlight extraction methods, classifiers and execution measures are displayed Whereas in [38], it exhibits a survey on the best machine learning approach on reading ECG. Machine learning in medical science is an

unmistakable research related on osmosis of present day innovations: programming, PC and data advancements [2]. There is even research on defining the ECG wave from other wave or mixed signal. For example Independent component analysis (ICA) is connected on the blended signs and the isolated signs are recreated utilizing wavelet remaking and correlating the results demonstrate that Lifting Wavelet Transformation and FASTICA algorithm creates the best SNR estimation of 11.39 for maternal and 10.10 for fetal Electro Cardio Gram signals [17]. In [51]. The purpose was to make an algorithm with elevated amounts of precision and less dimensions of false cautions by classifying the hearts electrical signals as demonstrative of ischemia or not. ECG early warning system comes to light for meeting public health, medical informatics referring to health services and information delivered or enhanced through the internet and related technologies [2]. For example, ROC (Receiver operating characteristics) diagrams are helpful for sorting out classifiers and observing their execution, moreover they are normally utilized in medical decision making, and lately have been utilized progressively in machine learning and information mining research[10]. Ultimately, there are many research work involving machine learning and medical science, however we choose to work with ECG or the functionality of the human heart. During the time most common defect in human body primarily related with heart.

Chapter 3

Data and Methods

In this chapter, the details about our data set and the machine learning algorithms that we implemented are given.

3.1 Dataset

We worked with the data set we found in the UCI Machine Learning Repository (UCI) database [9]. The data set or the input was a .data file. A screen shot of a single row of the data set is shown below in fig 3.1.

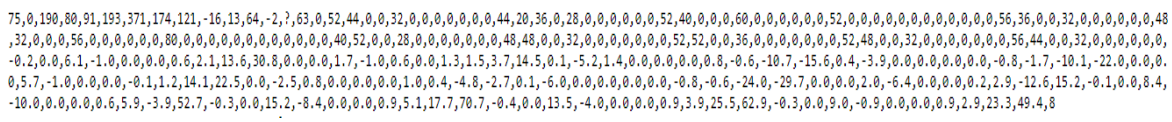

Fig. 3.1 A snapshot of the Dataset [9]

Here, in this .data file the data are given in rows where each rows are the instances and the columns are the attributes. However the attribute name of each column is not given in the .data file. In a different file the names of the attributes are given for each column. It is a pre processed data which contains 452 instances and 279 attributes, 206 of them are linear valued and the rest are nominal where 16 types of cardiac arrhythmia are classified in this data set whereas the first one being the normal ECG and class distribution of this data set is really unjust as three of the class's (11, 12, 13) instances does not exist and Class01 (normal) is most repeated [15]. These data are taken by using 10 electrodes. There are repetitive attributes for 12 different leads. The class distributions are Normal, Ischemic changes (Coronary Artery Disease), Old Anterior Myocardial Infarction, Old Inferior Myocardial Infarction, Sinus tachycardia, Sinus bradycardia, Ventricular Premature Contraction (PVC), Supra ventricular

Premature Contraction, Left bundle branch block, Right bundle branch block, First-degree atrioventricular block (AV block), Second-degree atrioventricular block (AV block), Third-degree atrioventricular block (AV block), Left ventricle hypertrophy, Atrial Fibrillation or Flutter, and Others[9]. Attributes of this data set are Age, Sex, Height, Weight, QRS duration: Average of QRS duration in msec., P-R interval: Average duration between onset of P and Q waves in msec., Q-T interval: Average duration between onset of Q and offset of T waves in msec., T interval: Average duration of T wave in msec., P interval: Average duration of P wave in msec., Vector angles in degrees on front plane of:, QRS , T , P, QRST , J, Heart rate: Number of heart beats per minute which are the first fifteen columns [15][9]. Additionally attributes from channel DI, channel DII, channel DIII, channel AVR, channel AVL, channel AVF, channel V1, channel V2, channel V3, channel V4, channel V5, channel V6 are Average width in msec. of Q wave, R wave, S wave, R' wave, small peak just after R, S' wave, and Number of intrinsic deflections, Existence of ragged R wave, Existence of diphasic derivation of R wave, Existence of ragged P wave, Existence of diphasic derivation of P wave, Existence of ragged T wave, Existence of diphasic derivation of T wave, Amplitude of JJ wave, Q wave, R wave, S wave, R' wave, S' wave, P wave, T wave and QRSA , Sum of areas of all segments divided by 10 (Area= width * height / 2), QRSTA = QRSA + 0.5 * width of T wave * 0.1 * height of T wave. [15][9]. Around 0.33 percent of data of the dataset is missing [15]. Data representation of different attributes are given below in figure 3.2, 3.3, 3.4 and 3.5.

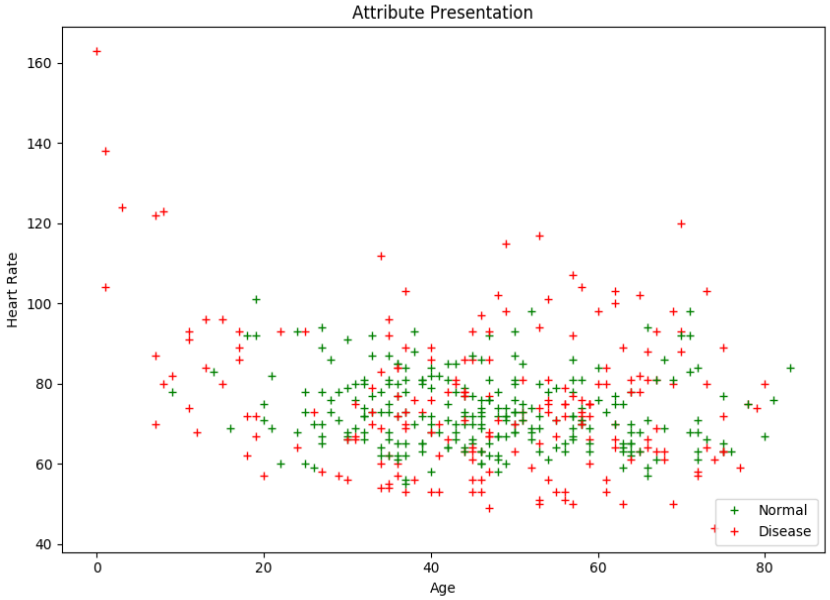

Fig. 3.2 Scatter diagram of age vs heart rate

In figure 3.2 normal and abnormal heart rate of various ages are plotted in a heart rate vs age graph.

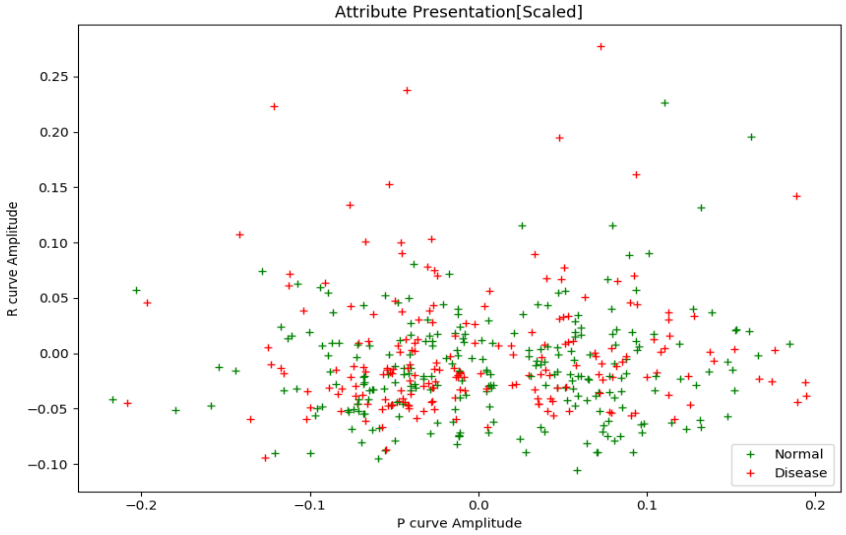

Fig. 3.3 Scatter diagram of scaled amplitude of P curve vs R curve

In figure 3.3, scaled representation of the amplitude of P curve vs R curve for normal and abnormal ECG are plotted.

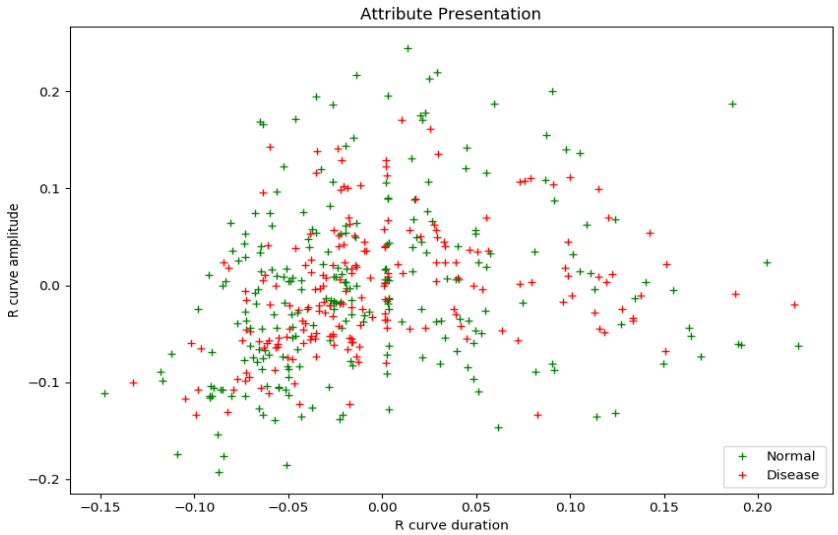

Fig. 3.4 Scatter diagram of R curves duration vs amplitude

Figure 3.4 represents the amplitude and duration of R curve for normal and abnormal ECG.

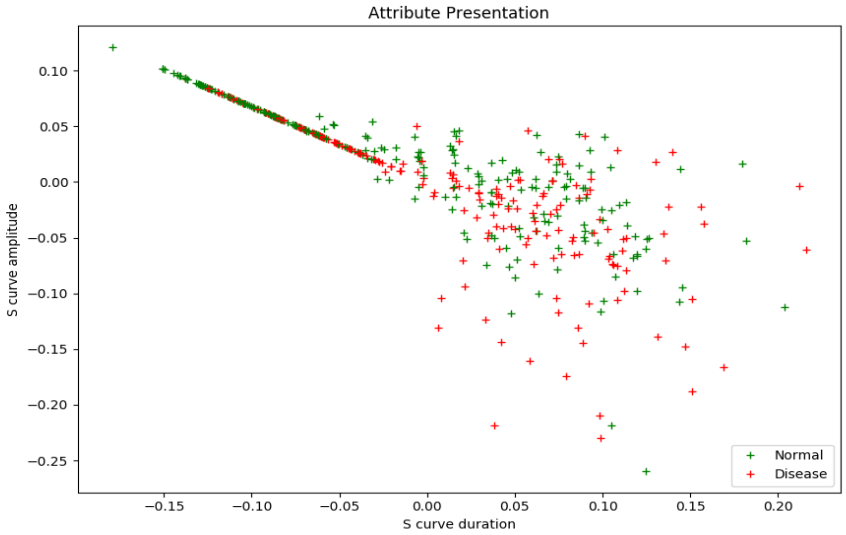

Fig. 3.5 Scatter diagram of S curves duration vs amplitude

In figure 3.5,the attribute representation for the amplitude vs duration of S curve for normal and abnormal ECG is given.

3.1.1 Data Refining and Categories

The dataset we found from UCI database, the columns are attributes and the rows are instances. It has some data missing for different attributes, as well instances. There are 10 instances which has missing data. That is why, we deleted them from the dataset before using it. The respective rows are 5, 66, 91, 200, 213, 238, 242, 360, 372 and 412. Moreover, the maximum data is missing in column 12, 14 and 199 which are respectively vector angles in degree P, vector angles in degree J and amplitude of channel AVR, T wave. Additionally 17 columns have 0.0 values. They are 20:(of channel DI) S' wave, 68:(of channel AVL) S' wave , 70:(of channel AVL) Existence of ragged R wave, 84:(of channel AVF) Existence of ragged P wave, 132:(of channel V4) Existence of ragged P wave, 133:(of channel V4) Existence of diphasic derivation of P wave, 140:(of channel V5) S' wave, 142:(of channel V5) Existence of ragged R wave, 144:(of channel V5) Existence of ragged P wave, 146:(of channel V5) Existence of ragged T wave, 152:(of channel V6) S' wave, 157:(of channel V6) Existence of diphasic derivation of P wave, 158:(of channel V6) Existence of ragged T wave, 165:(of channel DI) S' wave, 205:(of channel AVL) S' wave, 265:(of channel V5) S' wave,

and 275:(of channel V6) S' wave. These columns were deleted just for predicting abnormal and normal ECG.

After this, we considered all data into two categories, normal and abnormal. Considering all ECGs of heart diseases as abnormal and normal ECGs are normal where normal is 1 and abnormal 0. Although for specific disease prediction the complete dataset is used without those 10 instances or rows mentioned before.

3.2 Algorithms

We have implemented six algorithms separately to analyze our data. The algorithms are as followed:

1. Logistic regression
2. Decision tree
3. Nearest neighbour
4. Naïve Bayes
5. Support Vector Machine (SVM)
6. Neural Network

A brief description of all these algorithms are given below.

3.2.1 Logistic regression

Logistic regression is a linear model for classification. It is used to predict the probability of occurrence of an event by fitting the data to a sigmoid function. In other words, it estimates values based on given set of independent variables. The estimated values are discrete (binary values) [41][8][4][34].

$$f(x) = \frac{L}{(1 + e^{-}k(x - x_0))} \tag{3.1}$$

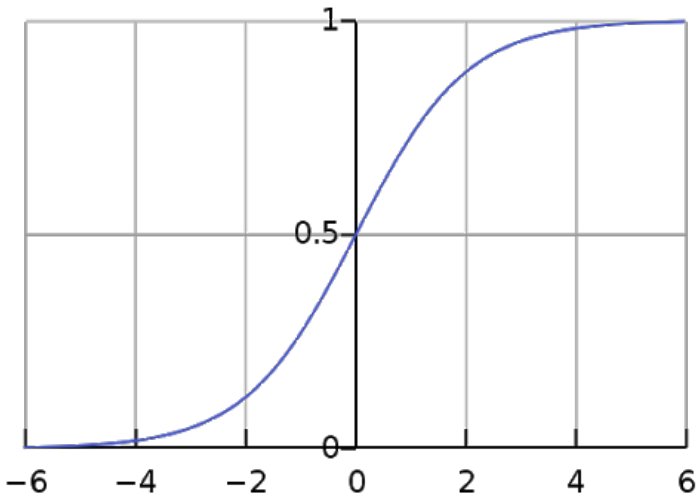

Fig. 3.6 The logistic curve [36]

3.2.2 Decision tree

It is a supervised learning method used to solve classification problems mostly. It works for both categorical and continuous input and output variables. The model learns simple decision rules inferred from its data features and then predicts the value of a target variable. In other words, the population or sample is divided into two or more homogeneous sets (or sub-populations) based on most significant differentiator in input variables.

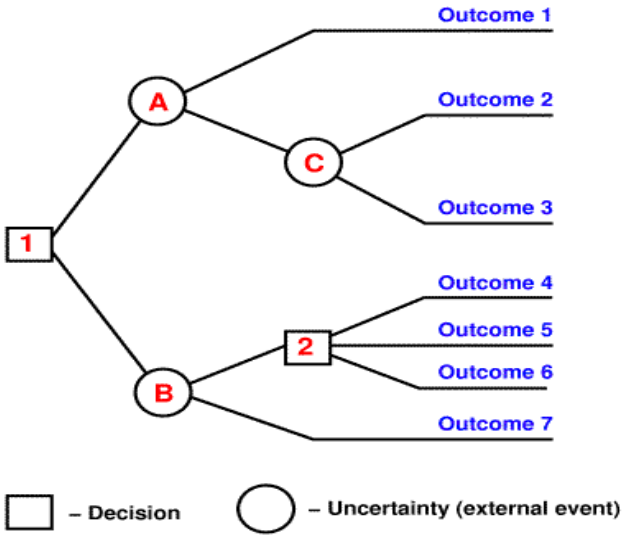

Fig. 3.7 Decision tree [28]

In equation 3.2 H(S) is entropy which measure the quantity of variability of the data in the dataset S [31][44][14][27].

$$H(S) = \sum_{x \in X} -p(x) \log_2 p(x) \tag{3.2}$$

In equation 3.3 IG(A) is information gain which measure the difference of the entropy after dividing set S on attribute A [31][44][14][27].

$$IG(S,A) = H(S) - \sum_{t \in T} p(t)H(t) = H(S) - H(S|A) \tag{3.3}$$

Decision tree identifies the most significant variable and it's value that gives best homogeneous sets of population [5][39][16] [34].

3.2.3 Nearest neighbour

It is a type of instance-based learning that simply stores instances of the training data. In short, it involves a training set of cases and the aim is to predict the label from these predefined number of training samples closest in distance to the new point.

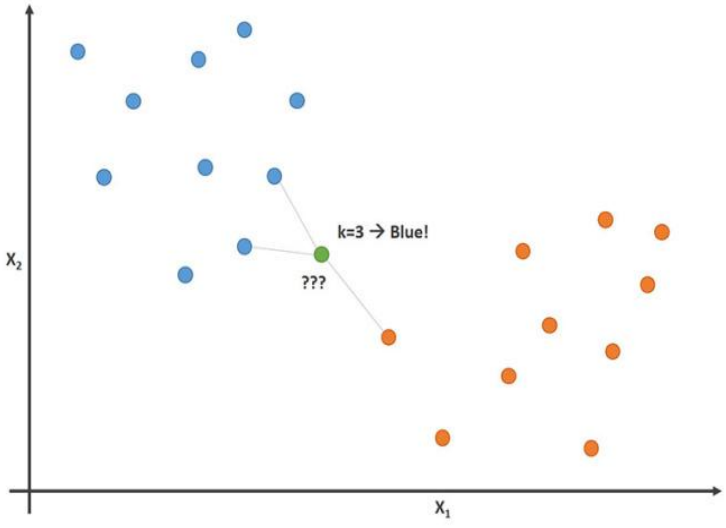

Fig. 3.8 Nearest Neighbors [20]

The number of samples can be a user-defined constant or vary, based on the local density

of points. The distance can, in general, be any metric measure. A new sample is classified from a simple majority vote of the nearest neighbors of each points [30].

3.2.4 Naïve Bayes

It is a training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions.It is suitable for linear classification [33]. There may be multiple features but each one is assumed to be a binary-valued variable.

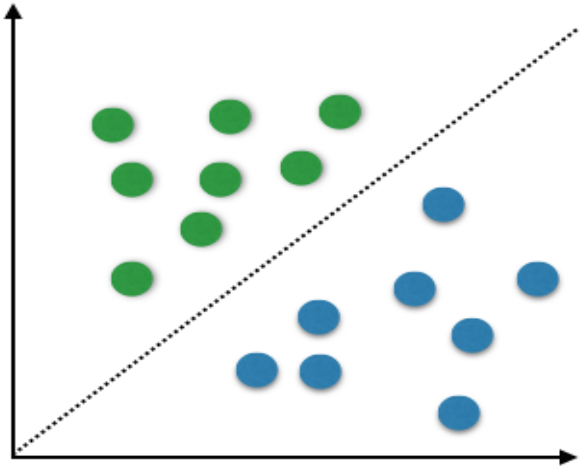

Fig. 3.9 Naïve Bayes, linear classifier [33]

Therefore, this class requires samples to be represented as binary-valued feature vectors [30] and the decision rule for the algorithm is based on

$$P(x_i|y) = P(i|y)x_i + (1 - P(i|y))(1 - x_i) \tag{3.4}$$

3.2.5 Support Vector Machine (SVM)

This is a set of supervised learning method that is used for classification, regression and outliers detection. The algorithm outputs an optimal hyperplane which categorizes new examples given labeled training data [7]. It performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels.

$$[\frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i(w.x_i - b))] + \lambda ||w||^2 \tag{3.5}$$

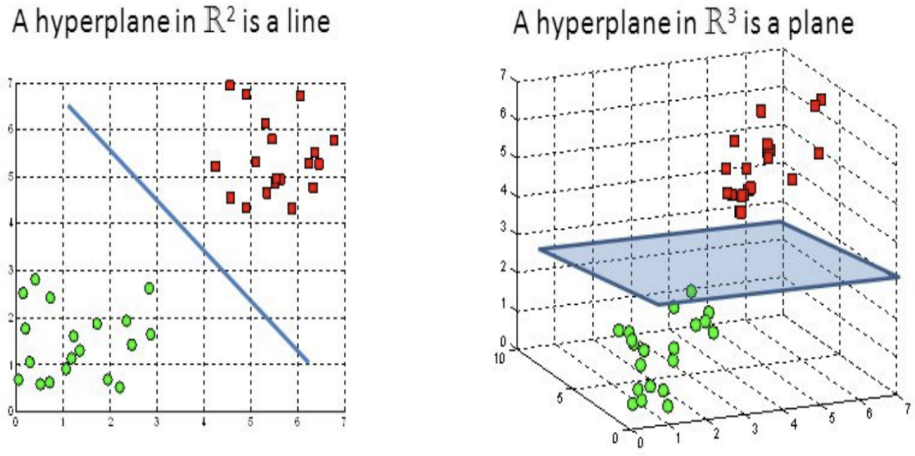

Fig. 3.10 SVM: Hyperplanes in 2D and 3D feature space [20]

3.2.6 Artificial Neural Network

They are organized in layers that are made up of a number of interconnected nodes which contain an activation function.

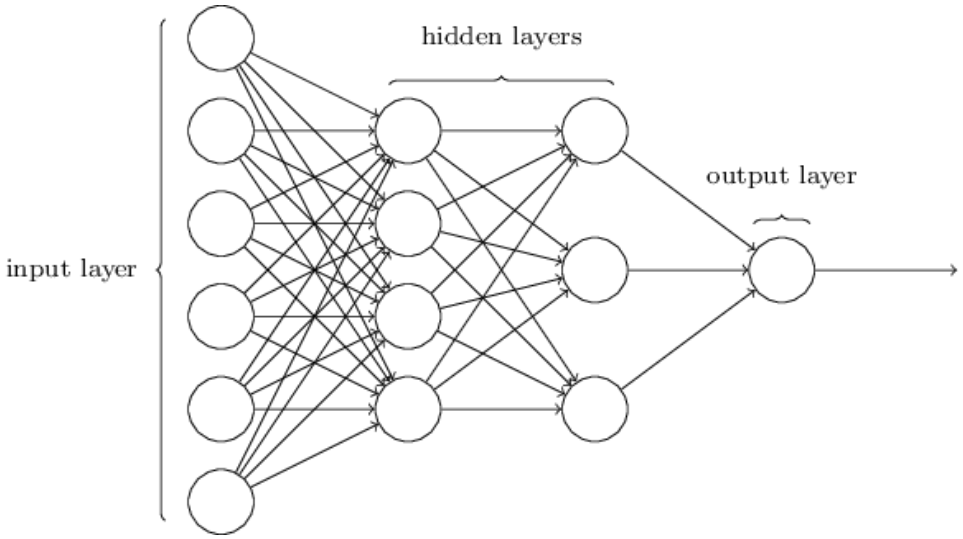

Fig. 3.11 ANN [12]

Patterns are presented to the network via the input layer, which communicates to one or more hidden layers where the actual processing is done via a system of weighted connections. The hidden layers then link to an output layer which gives the answer [7].

$$w_{ij}(t+1) = w_{ij}(t) + \eta \frac{\partial C}{\partial w_{ij}} + \xi(t) \tag{3.6}$$

3.3 Confusion Matrix

A confusion matrix is a table that helps to figure out the performance of a classification model when the true values are known for a set of data. A general confusion matrix is given below and the outcomes are explained following it.

Table 3.1 Confusion Matrix for a binary classifier [10]

	Predicted: No	Predicted: Yes
Actual: No	TN	FP
Actual: Yes	FN	TP

The matrix contains two prediction classes; a yes and a no. In simpler words, when we predict an answer to be false and the answer is actually false, the result is True Negative (TN) that is we predicted a correct negative answer. When we predict an answer to be true but the answer is actually false, the result is False Positive (FP) that is we predicted a negative answer as correct.

When we predict an answer to be true and the answer is actually true, the result is True Positive (TP) that is we predicted a correct positive answer. And finally when we predict an answer to be false but the answer is actually true, the result is False Negative (TN) that is we predicted a correct answer to be wrong.

During our research, we also found out the values of the terms of confusion matrix to get a better overview of our result.

3.4 Workflow

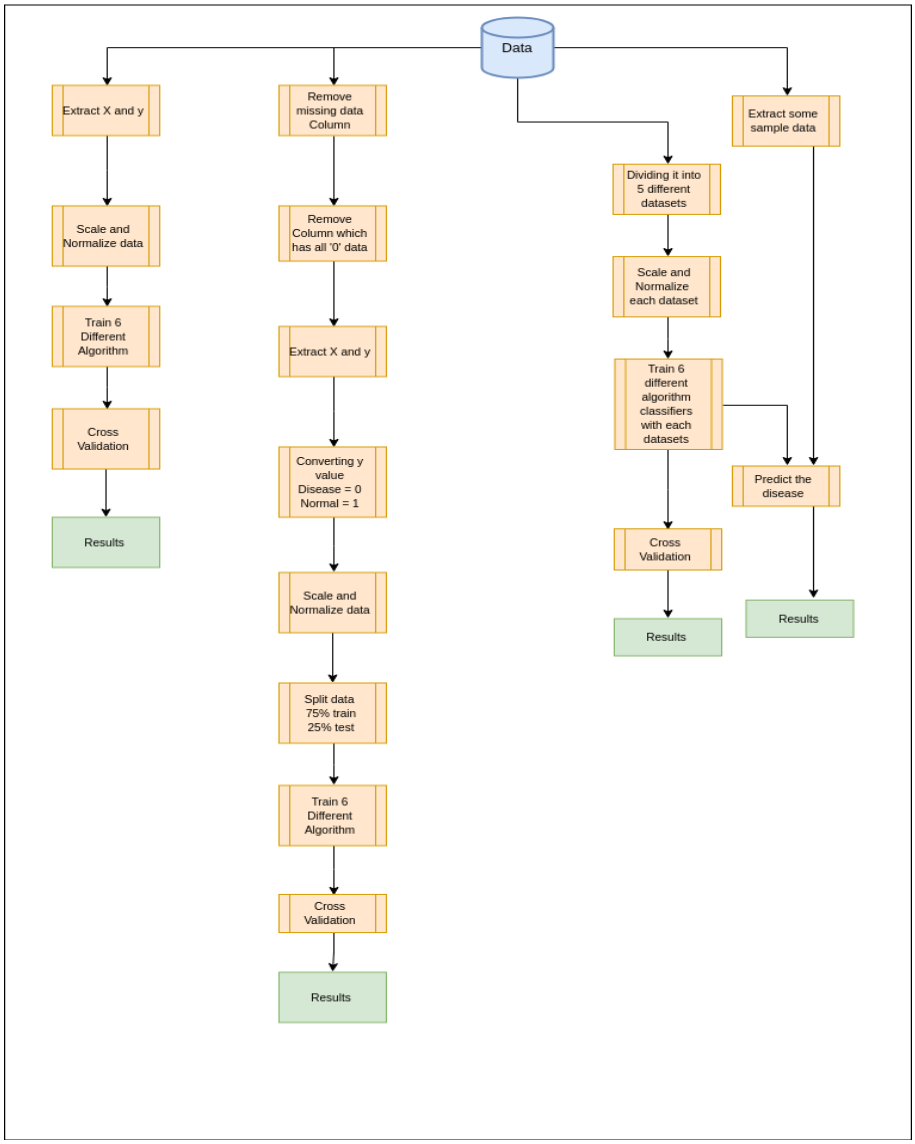

Fig. 3.12 The diagram of the workflow

In fig 3.8, there is a flowchart that briefly shows our work that has been discussed throughout the paper. It states how we divided the dataset to train and test our classifier with different algorithms that we have discussed in subsection 3.2 of this very chapter. We also divided some sample data from the main data set only to predict the disease of that particular using our classifiers which are discussed in chapter 5. At the beginning, we removed all the rows of the data set which had missing values, there were 10 such rows. So our instances came down to 442 from 452. After that we did three different works here.

First, we extracted x and y where x is the attributes and y is their respective classes of diseases. We normalized and scaled those data and train them through 6 different algorithm using cross validation. Subsequently, we did our second experiment. We considered all diseases in y are abnormal ECG valued 0 and normal ECG valued 1. We removed the columns where all values are '0'and columns with too much missing data. We again scaled and normalized the data, trained them through 6 different algorithms using both random train test split and cross validation. Furthermore, we created 5 different data sets according to 5 different diseases. The attributes we took for each data set are related to those diseases. After that, we again trained those data set by the 6 different algorithms using cross validation, making 30 classifiers. Additionally, we used the sample data as input to test and the chances of a heart disease to occur.

Chapter 4

Disease Analysis and Experimental Work

This chapter contains information about all the diseases we predicted and the setting up of our experiment.

4.1 Heart Diseases

A slight change in the PQRST section can lead to different kind of heart disease. We We did our research on few diseases caused by this change based on the availability. To understand a normal ECG figure is shown in fig 4.1.

Fig. 4.1 Normal ECG [24]

4.1.1 Coronary artery disease

It is the damaging of major blood vessels (coronary arteries) that supply our heart with blood, oxygen and nutrients. The damage can be caused by smoking, high blood pressure, high cholesterol, diabetes and sedentary lifestyle. Cholesterol-containing deposits (plaque) in our

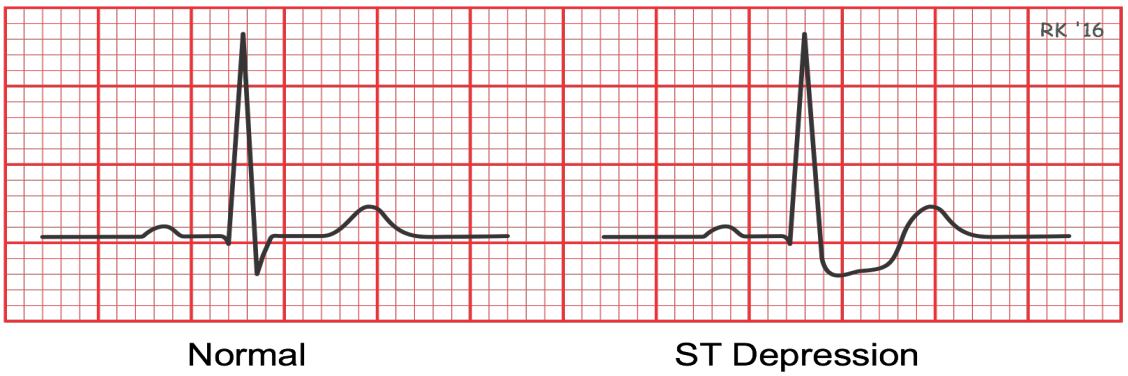

Fig. 4.2 Coronary artery heart disease ECG depression [21]

arteries and inflammation usually result in coronary artery disease. The building up of plaque narrows our coronary arteries and decreases the blood flow to our heart. Eventually, the decreased blood flow may result in chest pain (angina), shortness of breath, or other coronary artery disease signs and symptoms. A complete blockage can lead to heart attack. Some symptoms of coronary artery disease includes chest pain and shortage of pain. In the fig. 4.2 we observe the difference of the normal ECG and CAD ECG. Whereas the CAD ECG shows a ST depression. That is why, we choose to train our model by the different values of the attributes, Weight in kilograms, heart rate, QRS duration: Average of QRS duration in msec., P-R interval: Average duration between onset of P and Q waves in msec., Q-T interval: Average duration between onset of Q and offset of T waves in msec., T interval: Average duration of T wave in msec., Q wave: Average width in msec., S wave: Average width in msec.,Existence of ragged T wave, Existence of diphasic derivation of T wave, for all 12 leads to detect CAD.

4.1.2 Myocardial Infarction

A myocardial infarction (MI), commonly known as a heart attack, occurs when a portion of the heart is deprived of oxygen due to blockage of a coronary artery. Coronary arteries supply the heart muscle (myocardium) with oxygenated blood. Without oxygen, muscle cells served by the blocked artery begin to die (infarct). Myocardial infarction can be characterized from a sum of different perspectives related to clinical, electrocardiographic (ECG), biochemical and pathologic characteristics. It is accepted that the term myocardial infarction reflects death of cardiac myocytes caused by prolonged ischemia [2]. Changes in ECG when MI detected are ST rise and ST depression and T-wave changes shown in fig 4.3. New ST rise at the J-point in two bordering leads with the cut-off focuses: 0.2 mV in men or 0.15 mV in ladies in leads V2– V3 and additionally 0.1 mV in different leads and new flat or down-slanting

Fig. 4.3 ECG deflection from II, III and AVF leads for Myocardial Infarction [26]

ST depression >0.05 mV in two adjacent leads; and additionally T reversal 0.1 mVin two coterminous leads with noticeable R-wave or R/S proportion 1 [46]. We took different values of the attributes, Weight in kilograms, heart rate, QRS duration: Average of QRS duration in msec., P-R interval: Average duration between onset of P and Q waves in msec., Q-T interval: Average duration between onset of Q and offset of T waves in msec., T interval: Average duration of T wave in msec., Q wave: Average width in msec., S wave: Average width in msec.,Existence of ragged T wave, Existence of diphasic derivation of T wave, for all 12 leads for training our model for MI.

4.1.3 Sinus Tachycardia

It is a sinus rhythm with a heart rate greater than 100 beats/min. However, the PQRST wave stays normal. Some of the physiological reasons for this to occur are anxiety, emotion, exercise, pregnancy and pain. Other pathological conditions include fever, heart failure, shock, chronic constrictive pericarditis etc. It can also occur in response to certain drugs

Fig. 4.4 ECG of Sinus Tachycardia [47]

like nifedipine, sympathomimetic agents, atropine, adrenaline, isoprenaline etc. Persistent increase in resting heart rate unrelated to or proportion with the level of physical activity or emotional stress inappropriate sinus tachycardia. This is likely to be caused from a combination of factors that includes, a problem with the sinus node, unusual nerve signaling that causes heart rate to increase and dysfunction of the nerves that work to lower heart rate [1][3].As we can see increasing amount of heart rate in fig 4.4. For training our model we used the attributes, age in years, heart rate, Q wave: Average width in msec., R wave: Average width in msec.,S wave: Average width in msec.,Q wave: Amplitude , * 0.1 milivolt, R wave: Amplitude , * 0.1 milivolt, S wave: Amplitude , * 0.1 milivolt, for all 12 leads for training our model for detecting Sinus Tachycardia.

4.1.4 Sinus Bradycardia

It is a sinus rhythm with a heart rate less than 60 beats/min. From the fig. 4.5 there is a long gap between two PQRST waves dispite being normal like sinus Tachycardia. Many people with sinus bradycardia are not aware of the fact that they have it. In some cases, sinus bradycardia is a sign of a problem with the heart or another medical condition. It

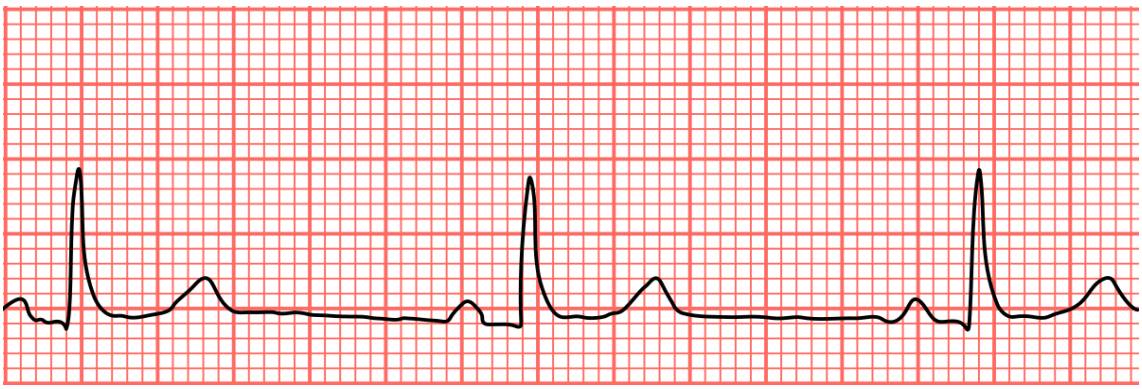

Fig. 4.5 ECG of Sinus Bradycardia [19]

is physiological in athletes due to increased vagal tone. Sleep, vomiting and coughing are another physiological factor. Other pathological conditions include Hypothyroidism, Hypothermia, Raised intracranial pressure, obstructive jaundice etc. It occurs mainly in patients taking drugs such as beta-blockers, digoxin and verapamil. This also results from complete heart blockage, diabetes and high blood pressure. Some symptoms of sinus bradycardia includes dizziness, fainting, shortness of breath and chest pain [1]. For training our model we used the attributes, age in years, heart rate, Q wave: Average width in msec., R wave: Average width in msec.,S wave: Average width in msec.,Q wave: Amplitude , * 0.1

milivolt, R wave: Amplitude , * 0.1 milivolt, S wave: Amplitude , * 0.1 milivolt, for all 12 leads for training our model for detecting Sinus bradycardia.

4.1.5 Right Bundle branch block (RBBB)

It is a blockage in the right bundle of the electrical conduction system. The impulses travelling through the right bundle branch does not activate the right ventricle during a RBBB[49]. So, there is a delay or blockage along the pathway that electrical impulses travel to make our heart beat. Therefore, pumping blood efficiently through our body gets difficult in such

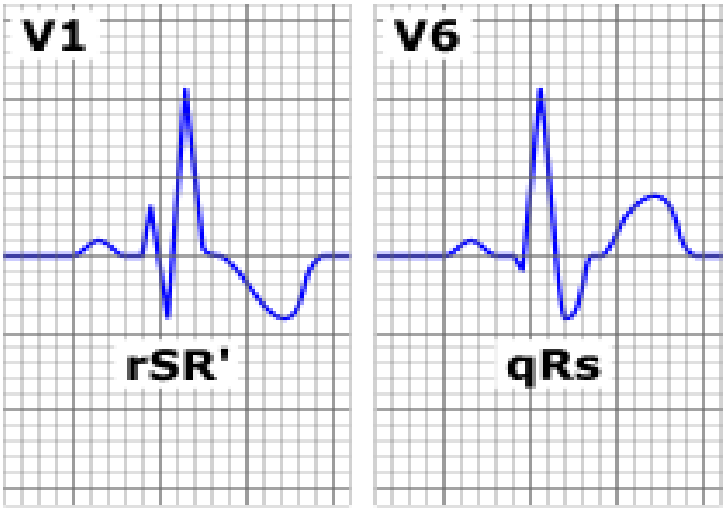

Fig. 4.6 RBBB ECG [32]

situations at times. V1 and V6 leads mainly shows the QRS deflection, fig.4.6. During this, the QRS duration is greater than 120 milliseconds. One of the most common symptom is fainting. Some causes for RBBB includes a heart abnormality that is present since birth, a heart attack, a bacterial infection on the heart muscle, high blood pressure in the pulmonary arteries and a blood clot in the lungs. For training our model we used the attributes, Q wave: Average width in msec., R wave: Average width in msec.,S wave: Average width in msec.,Q wave: Amplitude , * 0.1 milivolt, R wave: Amplitude , * 0.1 milivolt, S wave: Amplitude , * 0.1 milivolt, for all 12 leads for training our model for detecting RBBB.

4.2 Experimental set up

As we have already mentioned before that we decided to train our model by supervised machine learning, we had to divide our filtered data into two sets. One for training and the

other for testing. We used 75 percent of data to train our model. After the model was trained completely, the rest 25 percent of data was used to test our model for different algorithms and predict result for different diseases. It is also used to evaluate the model. To figure out how reliable our model is, we figured out scores for both random train-test split and cross validation. The results are discussed in the next chapter.

4.2.1 Random Train-Test Split

When a data is separated into a training set and a testing set, majority of the information is utilized for training, and a smaller part of the information is utilized for testing. A training data set is a data set of models utilized for learning on which the model is trained on [37]. A test data set is a data set that is autonomous of the training data set, however that pursues indistinguishable likelihood appropriation from the preparation data set and a superior fitting of the training data set rather than the test data set for the most part indicates over-fitting [40]. A test set is along these lines an arrangement of models utilized just to evaluate the execution of a completely determined classifier [37].

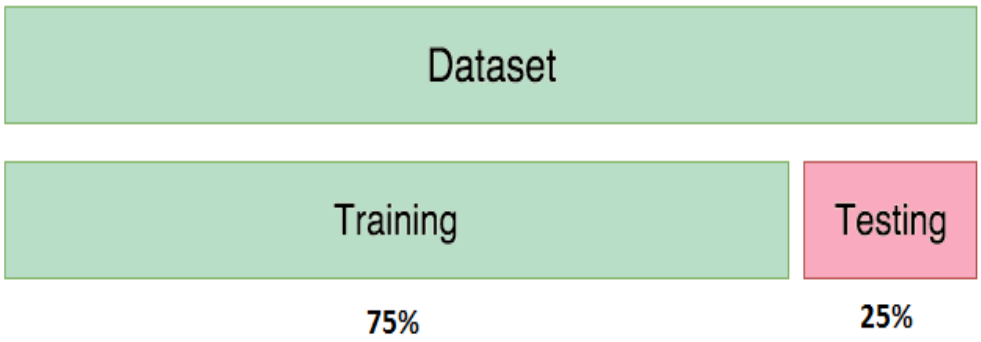

Fig. 4.7 Visualization of the splits

As we already know, for this case, the data set is separated into training data and testing data where training data is 75 percent and testing data is 25 percent as shown in fig 4.7.

4.2.2 Cross Validation

Cross validation is simple and widely used in machine learning [6]. Moreover, result is significantly less biased then other methods [6]. There are around three steps for doing cross validation [35],

1. Save some segment of the dataset .

2. Utilizing the rest of the dataset to train the model.

3. Test the model utilizing the save part of the dataset.

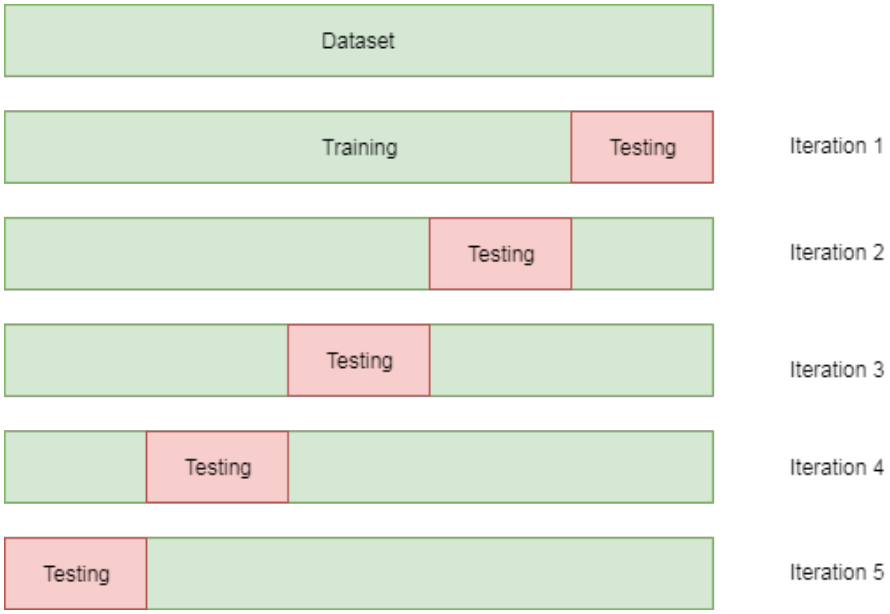

Fig. 4.8 Visualization of the test and train dataset

To do cross validation, we divided our data into 'n' subsets and every time we used 'n-1' of these subsets to train our model. The rest '1' subset was used to test our model. This was done repeatedly for all possible combinations which has been visualize in fig. 4.8. In addition, this was also done to predict outcomes for the diseases we analyzed using our dataset.

Chapter 5

Results and Discussions

This chapter contains all the mathematical and graphical results that we found out in our research from the six different algorithms implemented by us. Each graph have a brief description of its outcome alongside. Furthermore, the results are also discussed and compared where needed.

5.1 Results for normal and abnormal ECG

The bar chart in fig 5.1 shows the cross validation score for predicting all 15 diseases without removing any missing data and the bar chart in fig 5.2 and 5.3 shows the cross validation and random train-test split result for normal and abnormal ECG classified by Logistic regression, Naïve Bayes, Decision tree, Support Vector Machine and Nearest Neighbour algorithms. The missing data are removed where needed. In both cases, the results are shown for both scaled and not scaled.

5.1.1 Cross Validation outcome including missing data

In fig 5.1, this bar chart shows the cross validation score for all 15 diseases predicted by each algorithm for both with and without scaling. All the outcomes are below 70 percent. This is because it includes all the missing data for few attributes. Moreover, the data overlaps with each other while they are being calculated. Here, we can see that scaling gives a better result in most cases except for the Decision Tree algorithm which fell from 0.63 to 0.60 and the Logistic Regression algorithm which fell from 0.63 to 0.56 when scaled. However it increased slightly by 0.01 for Nearest Neighbour (from 0.44 to 0.45) and by 0.02 for Naïve Bayes (from 0.64 to 0.66) and to a great extent for SVM that is from 0.63 to 0.68 almost by 0.05 which is the most increase here. From this result, we can say that SVM gives the

best score when scaling is used whereas SVM gives the second lowest result (which has a difference of only 0.01) when it is not scaled. Decision tree and Logistic regreesion also gives the same result as SVM when not scaled. Naïve Bayes gives the best result when it is not scaled. Nearest Neighbour gives the worst result for both scaled and not scaled. The most change in score is seen for Artificial neural network, which increased by 0.13 (from 0.45 to 0.58) when scaled.

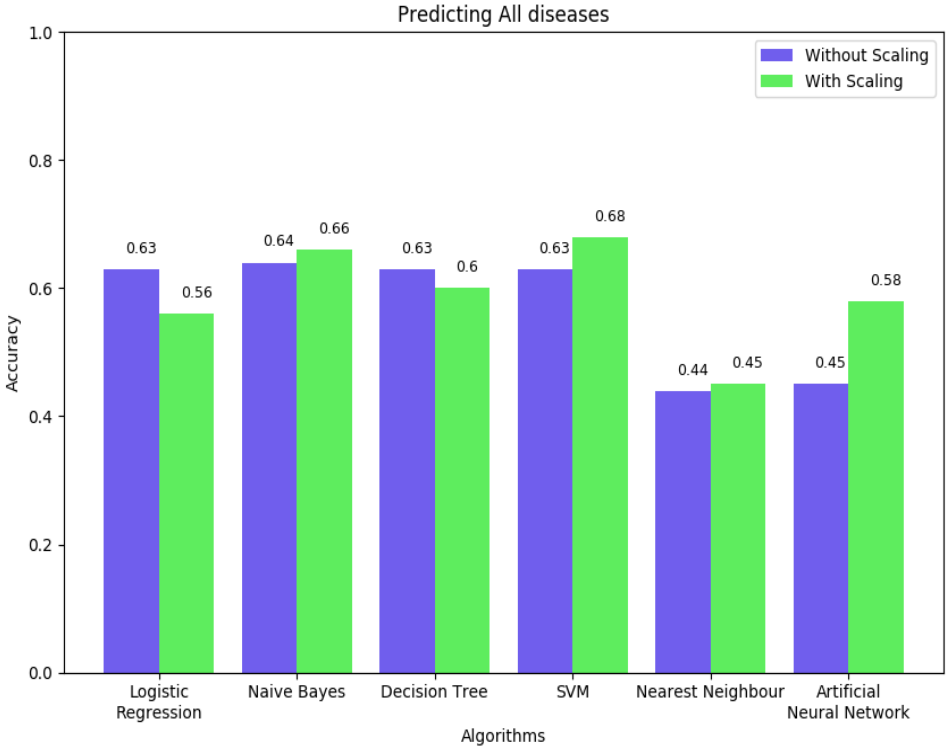

Fig. 5.1 Cross Validation Score of predicting all 15 diseases

5.1.2 Cross Validation outcome for normal and abnormal ECG

In fig 5.2, this bar chart shows the cross validation score for normal and abnormal ECG for both with and without scaling. It can be deduced that scaling gives a better result in most cases here except for the decision tree algorithm which fell from 0.77 to 0.71 when scaled. The score is constant for logistic regression and Naïve Bayes that is 0.72 and 0.74 respectively. However it increased slightly by 0.01 for nearest neighbour(from 0.67 to 0.68) and to a great extent for SVM that is from 0.68 to 0.78 almost by 0.1. The score for ANN changes from 0.73 to 0.74 when scaled. From this result, we can say that SVM gives the

best score when scaling is used whereas SVM gives the second lowest result (which has a difference of only 0.01) when it is not scaled. Decision tree gives the best result when it is not scaled.

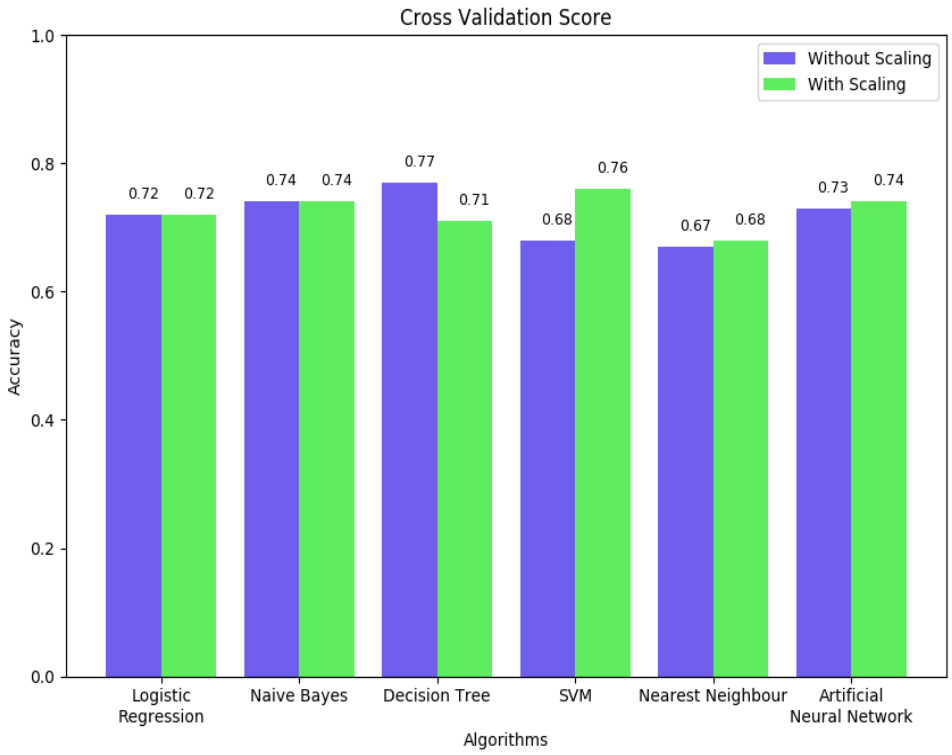

Fig. 5.2 Cross Validation Score for predicting normal and abnormal ECG

5.1.3 Random Train-Test Split outcome for normal and abnormal ECG

In fig 5.3, this bar chart shows the random train-test split score for normal and abnormal heartbeat for both with and without scaling. It can be deduced that scaling gives a better result in most cases here except for the decision tree algorithm which fell from 0.78 to 0.68 when scaled. The score has increased for all other algorithms when scaled. It increased by 0.03 for nearest neighbour and to a great extent for logistic regression that is from 0.63 to 0.72 almost by 0.09. Naive Bayes increased by 0.04 and SVM increased by 0.06. The score of ANN changes from 0.72 to 0.73 when scaled. From this result, we can say that SVM gives the best score when scaling is used and Decision tree gives the best result when it is not scaled.

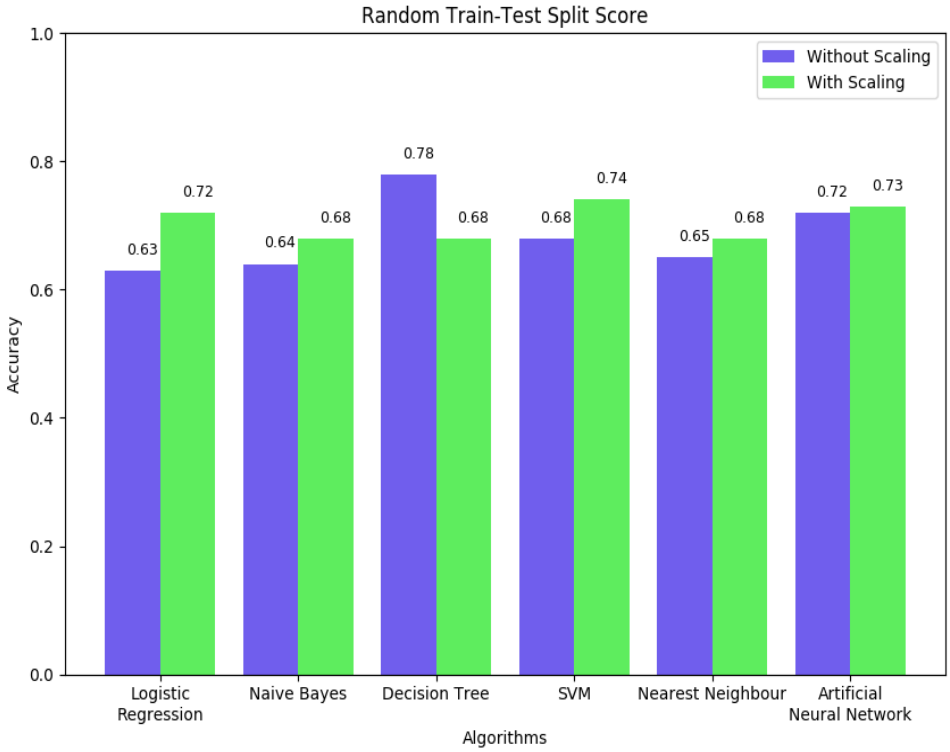

Fig. 5.3 Random Train-Test Split Score for predicting normal and abnormal ECG

5.1.4 Comparison between Cross Validation Score and Random Train-Test Split Score for normal and abnormal ECG

In both the cases (fig 5.2 and fig 5.3), SVM gave the best result when scaling was used and Decision tree gave the best result when it was not scaled. In both cases (fig 5.2 and fig 5.3), nearest neighbour gives the worst result with scaling (note that there are three algorithm with minimum value in case of Random Train-Test Split Score).

5.2 Results for individual disease and algorithm

This section contains result for the classification of individual diseases and algorithms. Table 5.1 and fig 5.4 shows the CV score. Table 5.2 and fig 5.5 shows the score for Random Train-Test split. Table 5.3 and fig 5.6 shows the CV score of ANN with different neuron counts. Each disease is distinguished by a separate colour which is mentioned in the fig 5.4, 5.5 and 5.6

5.2.1 CV score

Here is the summary of the cross validation accuracy score in table 5.1 that we found by classifying each individual disease with different algorithms.

Table 5.1 CV score for individual disease and algorithm

	LR	NN	DT	NB	SVM
CAD	0.86	0.76	0.87	0.94	0.86
MI	0.89	0.92	0.96	0.92	0.94
ST	0.95	0.64	0.95	0.95	0.95
SB	0.9	0.64	0.95	0.9	0.91
RBBB	0.85	0.8	0.88	0.9	0.92

A line graph generated from the table 5.1 is shown in fig 5.4

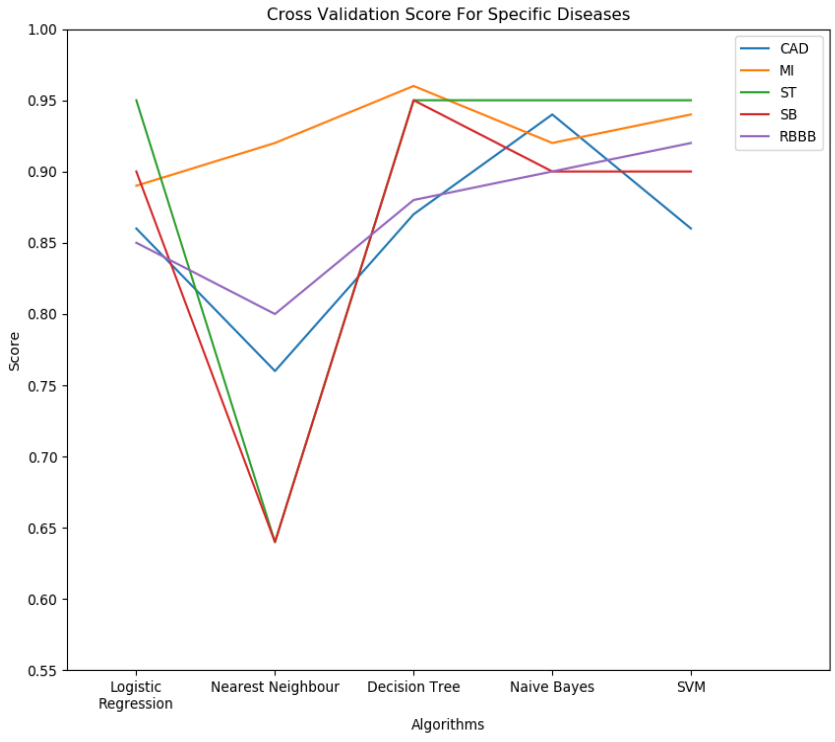

Fig. 5.4 Cross Validation Score for specific diseases

Naive Bayes gives the a score above 0.90 for all the diseases. We got the lowest accuracy score from Nearest Neighbor algorithm which is 0.64 for both ST and SB. For Coronary

Artery disease, NB gives the highest score of 0.94 and NN gives the lowest score which is 0.76. For Myocardia Infarction, DT gives the highest score of 0.96 and LR gives the lowest score which is 0.89. For Sinus Tachycardy, all the algo gives a score of 0.95 except NN which gives a score of 0.64. For Sinus Bradycardy, DT gives a highest score 0.95 and NN gives the lowest score which is 0.64. For RBBB, SVM gives the highest score 0.92 and NN gives the lowest score which is 0.8. Overall, Right Bundle Branch Block disease's scores is pretty consistent.

5.2.2 Random Train-Test Split score

Here is the summary of the Random Train-Test accuracy score in table 5.2 that we found by classifying each individual disease with different algorithms.

Table 5.2 Random Train-Test Split score for individual disease and algorithm

	LR	NN	DT	NB	SVM
CAD	0.9	0.89	0.87	0.83	0.89
MI	1	0.96	0.99	0.91	0.99
ST	0.94	0.7	0.97	0.94	0.94
SB	0.91	0.69	0.94	0.91	0.96
RBBB	0.96	0.86	0.91	0.81	0.93

Logistic Regression gives the a score above 0.90 for all the diseases. We got the lowest accuracy score from Nearest Neighbor algorithm which is 0.67 for SB. For Coronary Artery disease, LR gives the highest score of 0.9 and NB gives the lowest score which is 0.83. For Myocardia Infarction, LR gives the highest score, 1 and NB gives the lowest score that is 0.91. For Sinus Tachycardy, DT gives the highest score, 0.97 and NN which gives a score of 0.7 which is the lowest. For Sinus Bradycardy, SVM gives a highest score 0.96 and NN gives the lowest score which is 0.69. For RBBB, LR gives the highest score 0.96 and NB gives the lowest score which is 0.81. Myocardial Infarction has got the best accuracy score for all the algorithms. Overall, all the algorithms gave relatively very good result for all diseases. A line graph generated from the table 5.2 is shown in fig 5.5.

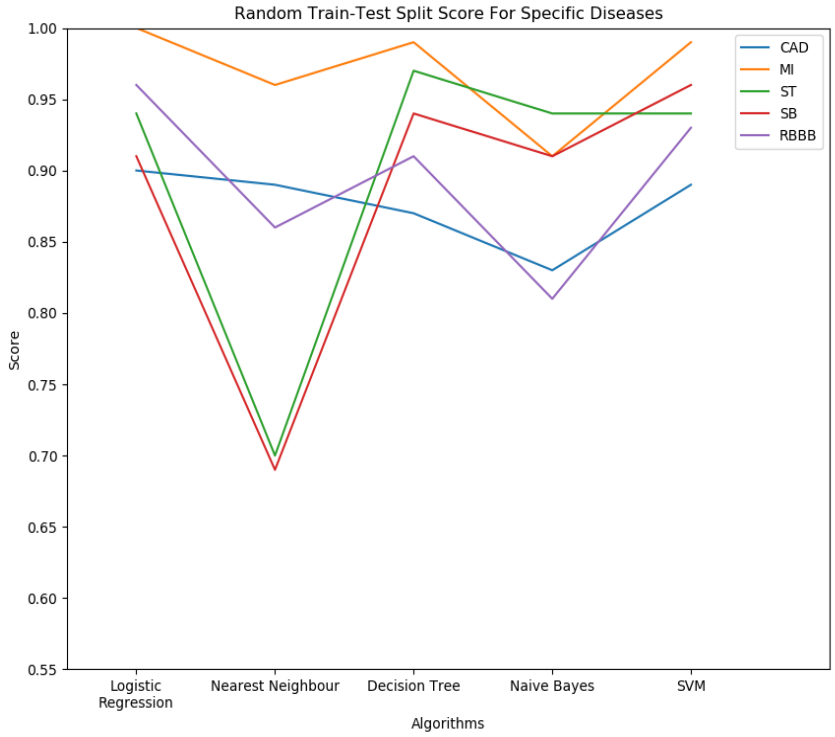

Fig. 5.5 Random Train-Test Split score for specific diseases

5.2.3 Cross Validation Score for ANN

Table 5.3 shows the Artificial Neural Network accuracy score of different neuron count. We used 3 hidden layers and every layers has the same number of neurons as given in table 5.3. The range of the neuron is between 50 to 400. The iteration count is 5000. We can deduce that every disease has the accuracy score between 0.85 to 0.93.

Table 5.3 CV score of ANN of individual disease vs Neuron count

	Neuron Count							
	50	100	150	200	250	300	350	400
CAD	0.854	0.861	0.861	0.861	0.861	0.861	0.865	0.865
MI	0.941	0.937	0.912	0.908	0.897	0.901	0.904	0.926
ST	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949
SB	0.903	0.907	0.907	0.907	0.907	0.907	0.907	0.907
RBBB	0.931	0.925	0.88	0.907	0.928	0.915	0.918	0.914

A line graph generated from the table 5.3 is shown in fig 5.6

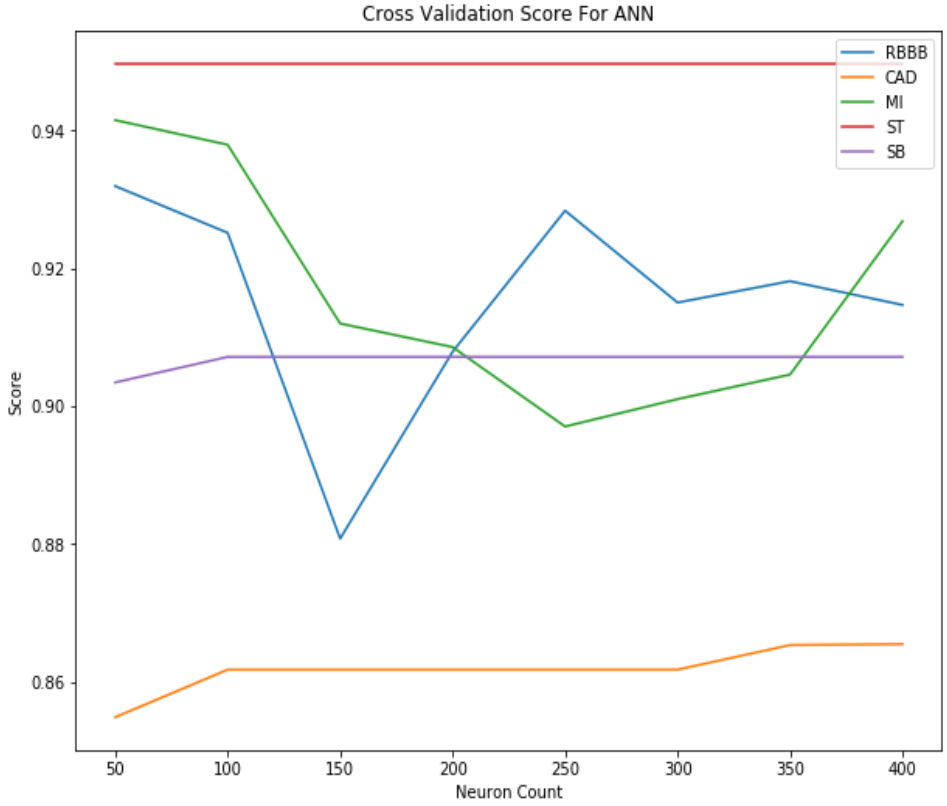

Fig. 5.6 Cross Validation Score of ANN for specific diseases

5.3 Prediction of disease with an input data set

An unknown data set is used as an input for all the six algorithms to figure out the possible disease for this input. The result for each algorithm is shown in a different pie-chart below respectively. Each pie-chart shows the best two possible results.

5.3.1 Result on Logistic Regression

In fig 5.7, We can see that it predicts the disease to be either RBBB with an maximum score of 21.8 percent or Coronary Artery Block with an equal score of 21.8 percent.

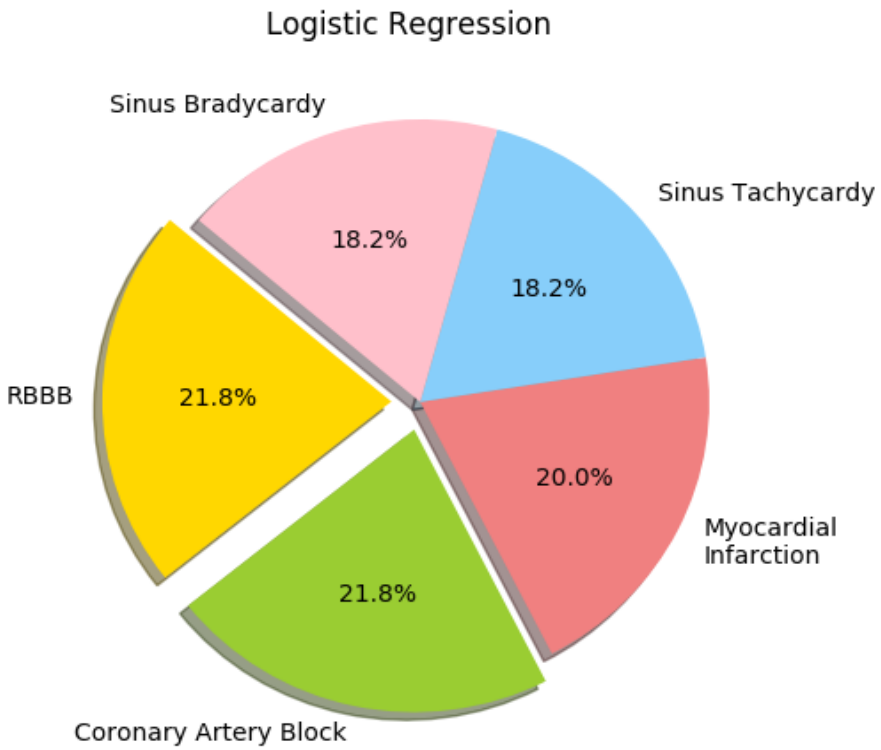

Fig. 5.7 Prediction for all diseases by Logistic Regression

5.3.2 Result on Decision Tree

In fig 5.8, We can see that it predicts the disease to be RBBB with an maximum score of 26.3 percent followed by Coronary Artery Block which has a score of 21.1 percent which is the second highest score.

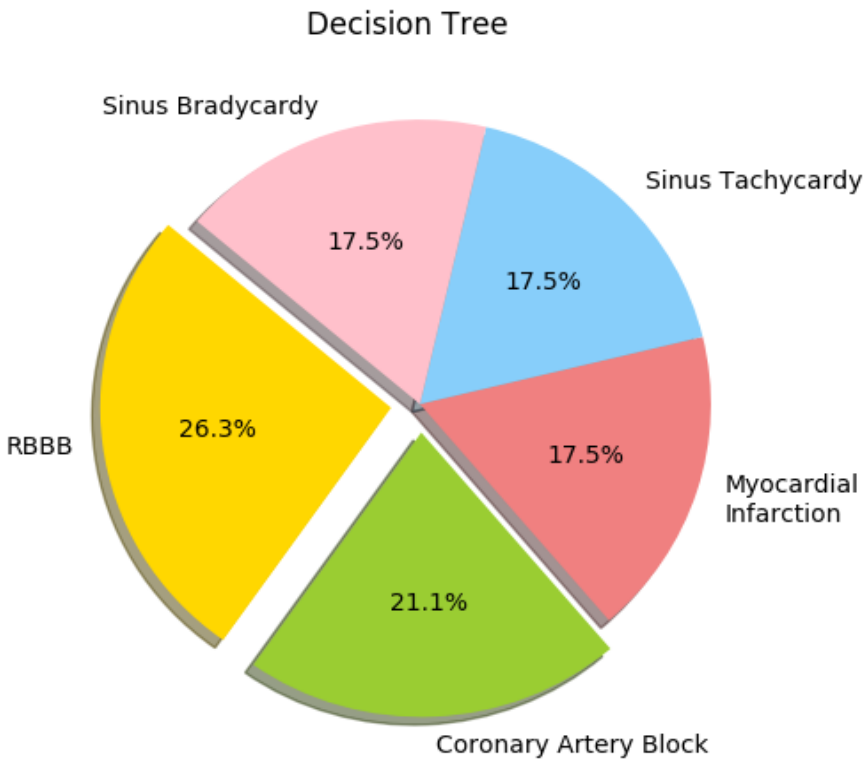

Fig. 5.8 Prediction for all diseases by Decision Tree

5.3.3 Result on Nearest Neighbour

In fig 5.9, We can see that it predicts the disease to be RBBB with an maximum score of 27.1 percent followed by Sinus Tachycardy which has a score of 22.0 percent which is the second highest score.

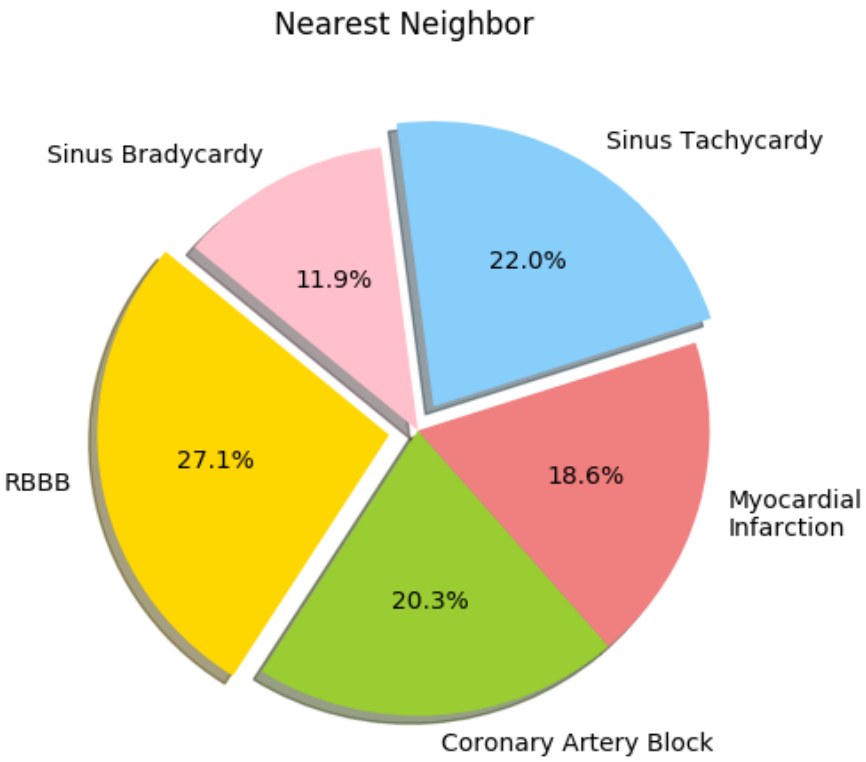

Fig. 5.9 Prediction for all diseases by Nearest Neighbour

5.3.4 Result on Naive Bayes

In fig 5.10, We can see that it predicts the disease to be RBBB with an maximum score of 27.9 percent followed by Coronary Artery Block which has a score of 19.7 percent which is the second highest score.

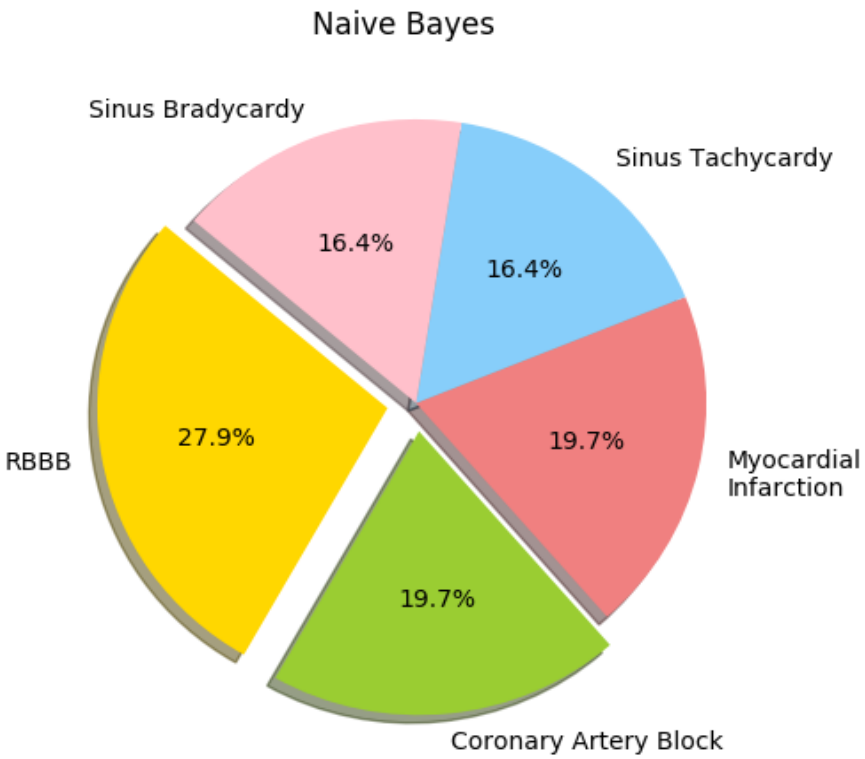

Fig. 5.10 Prediction for all diseases by Naive Bayes

5.3.5 Result on SVM

In fig 5.11, We can see that it predicts the disease to be RBBB with an maximum score of 23.6 percent followed by Coronary Artery Block which has a score of 21.8 percent which is the second highest score.

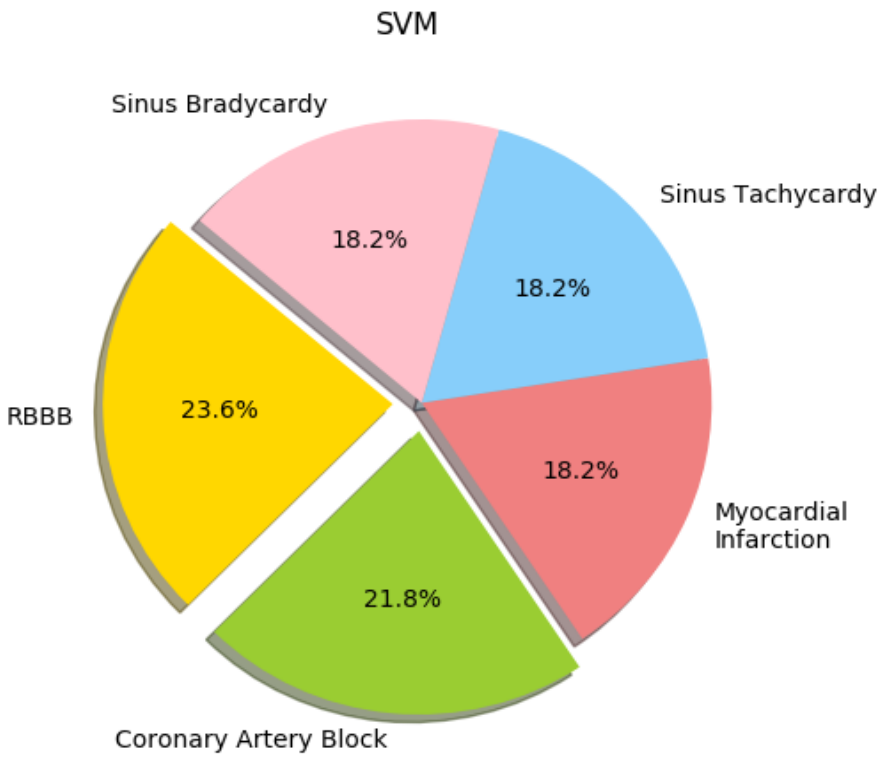

Fig. 5.11 Prediction for all diseases by SVM

5.3.6 Result on Artificial Neural Network

In fig 5.12, We can see that it predicts the disease to be RBBB with an maximum score of 25.5 percent percent followed by Myocardial Infarction which has a score of 21.8 percent which is the second highest score.

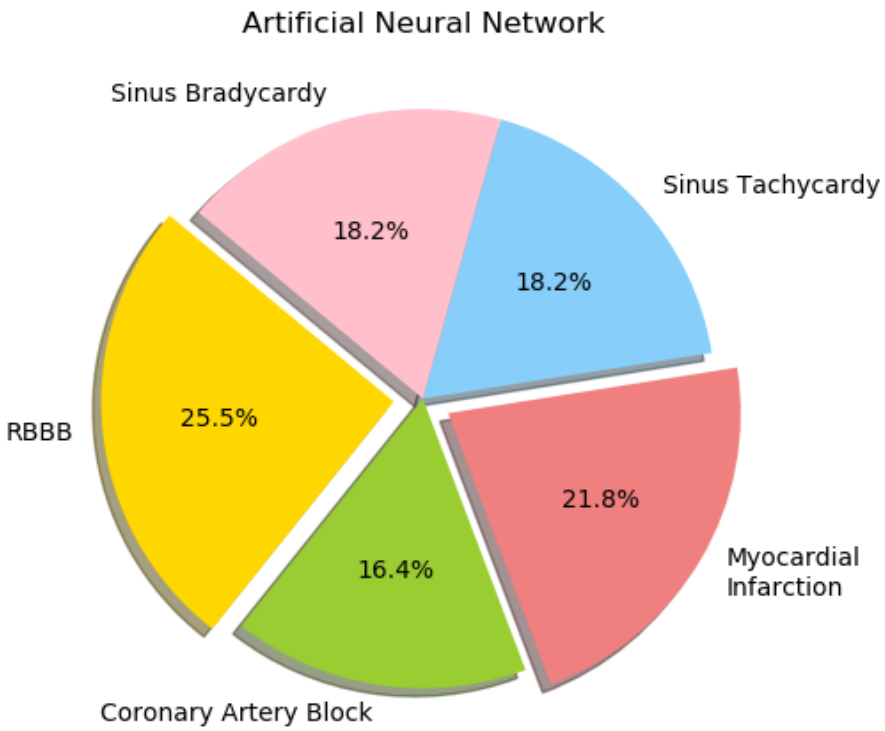

Fig. 5.12 Prediction for all diseases by Artificial Neural Network

5.3.7 Decision from these charts

Considering the result we found from the graph in fig 5.7 to 5.12, this can be concluded that the input data set is a sufferer from Right Bundle Branch Block since all the algorithms give a maximum score for RBBB. However, in real life, when we tried to figure out the real disease, we found out that this input data was indeed a data set of RBBB (the person whose data we used was a sufferer of RBBB).

5.4 Prediction of disease with a second input dataset

Another unknown data set is used as an input for all the six algorithms to figure out the possible disease for this input. The result for each algorithm is shown below in a different pie-chart below respectively. Each pie-chart shows the best two possible results similar to the previous sub section.

5.4.1 Result on Logistic Regression

In fig 5.13, We can see that it predicts the disease to be Coronary Artery Block with an maximum score of 20.8 percent followed by Myocardial Infarction which has a score of 22.6 percent which is the second highest score in fig.5.3.

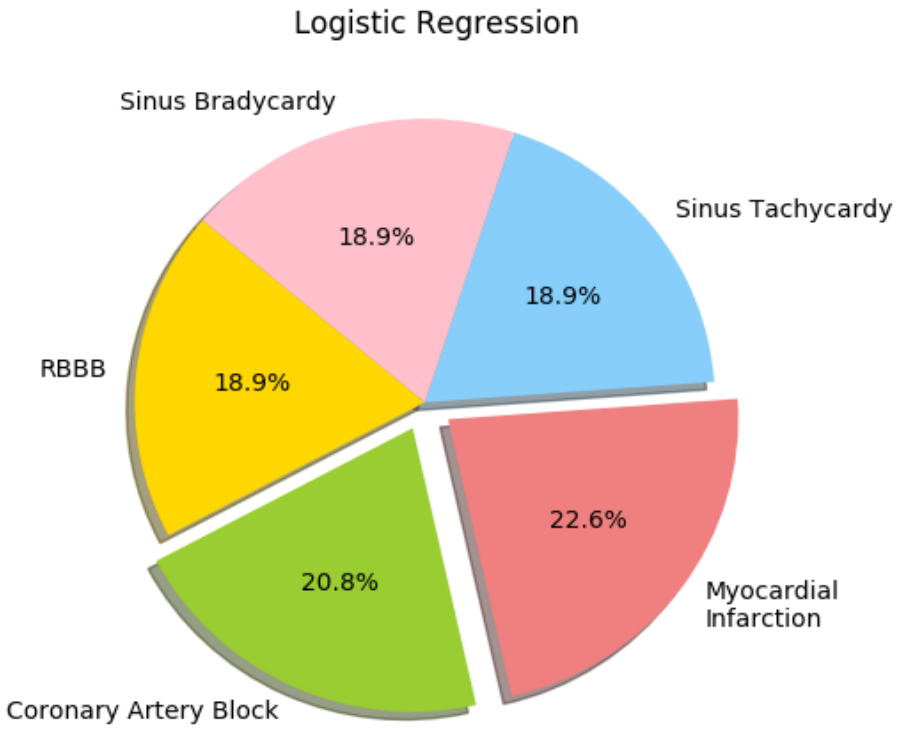

Fig. 5.13 Prediction for all diseases by Logistic Regression

5.4.2 Result on Decision Tree

In fig 5.14, We can see that it predicts the disease to be Coronary Artery Block with an maximum score of 25.9 percent followed by Myocardial Infarction which has a score of 20.4 percent which is the second highest score.

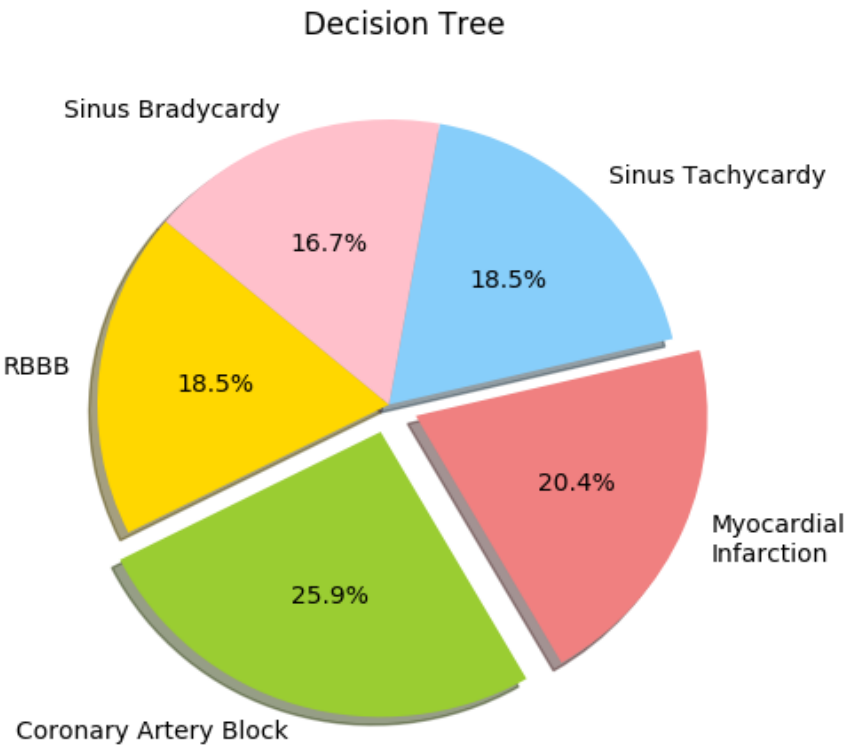

Fig. 5.14 Prediction for all diseases by Decision Tree

5.4.3 Result on Nearest Neighbour

In fig 5.15, We can see that it predicts the disease to be Sinus Tachycardy with an maximum score of 27.8 percent followed by Coronary Artery Block which has a score of 25.9 percent which is the second highest score.

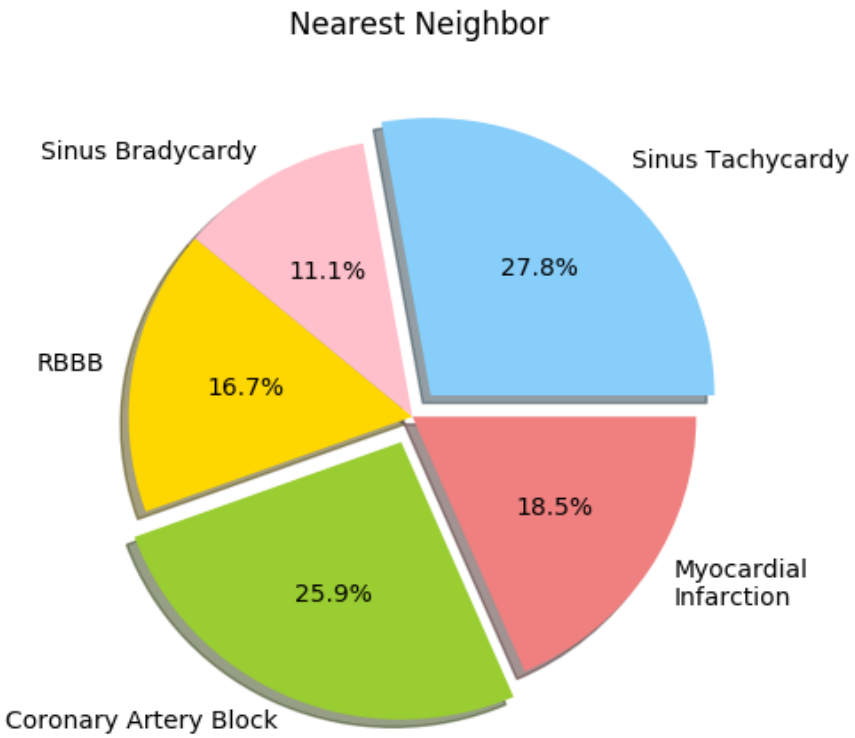

Fig. 5.15 Prediction for all diseases by Nearest Neighbour

5.4.4 Result on Naive Bayes

In fig 5.16 (on page 47), We can see that it predicts the disease to be Coronary Artery Block with an maximum score of 25.5 percent followed by Myocardial Infarction which has a score of 20.0 percent which is the second highest score.

5.4.5 Result on SVM

In fig 5.17 (on page 47), We can see that it predicts the disease to be RBBB with an maximum score of 22.2 percent followed by Coronary Artery Block which has a score of 20.4 percent which is the second highest score.

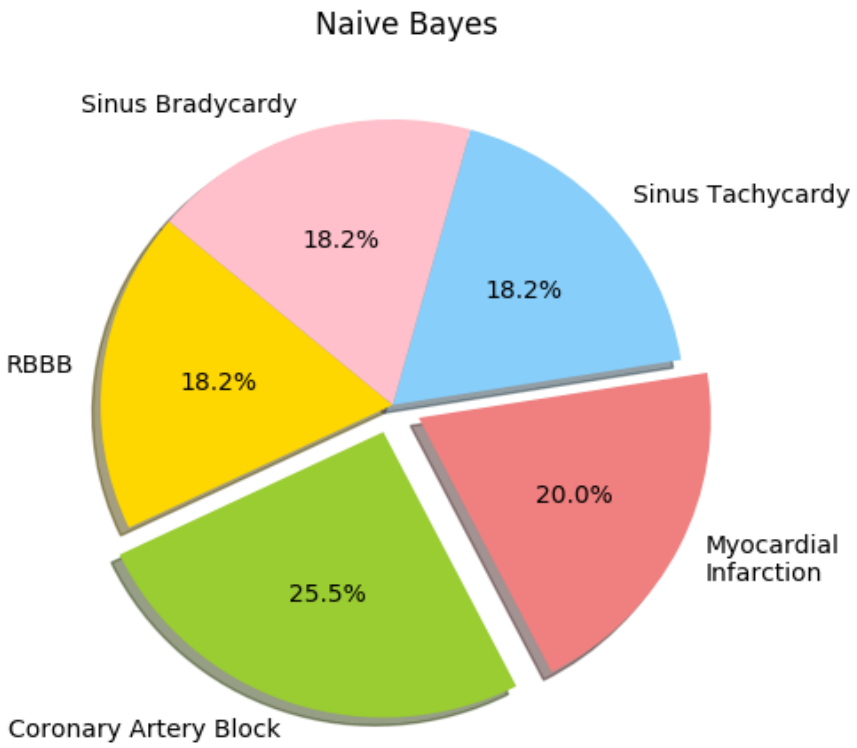

Fig. 5.16 Prediction for all diseases by Naive Bayes

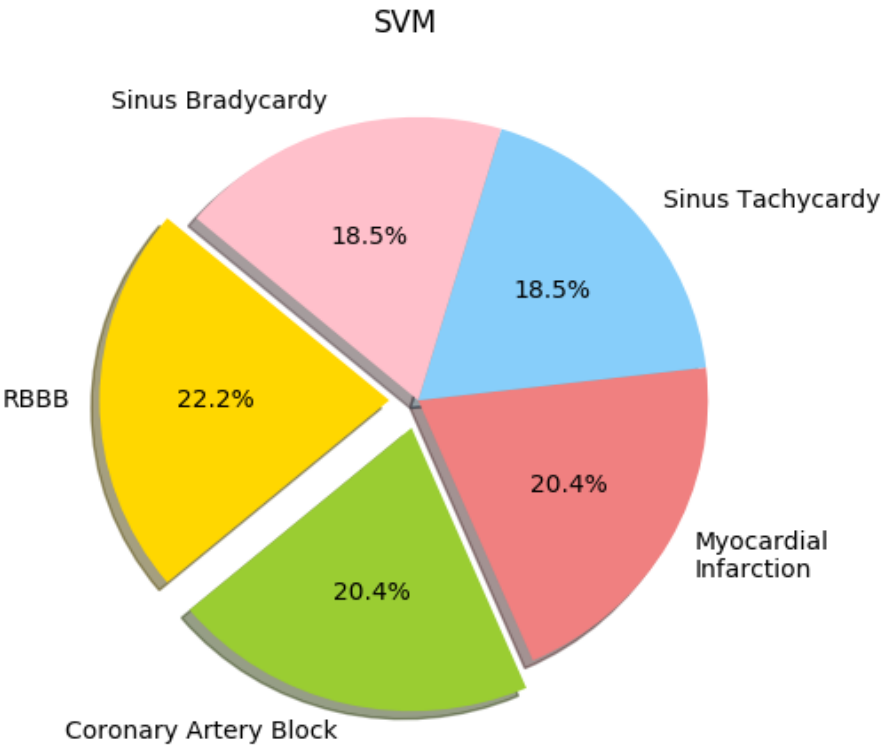

Fig. 5.17 Prediction for all diseases by SVM

5.4.6 Result on Artificial Neural Network

In fig 5.18, We can see that it predicts the disease to be Coronary Artery Block with an maximum score of 25.9 percent followed by Myocardial Infarction which has a score of 20.7 percent which is the second highest score.

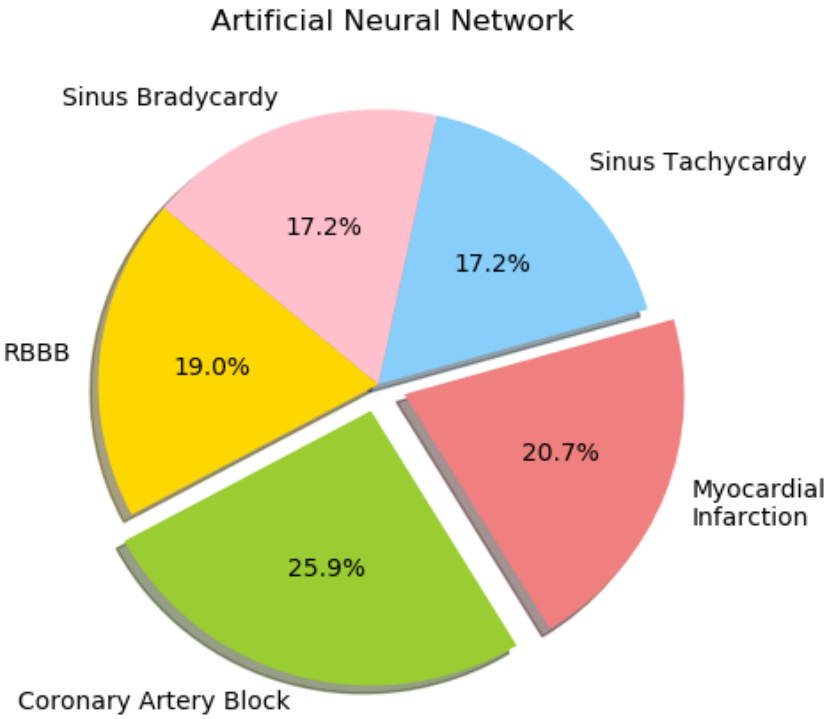

Fig. 5.18 Prediction for all diseases by Artificial Neural Network

5.4.7 Decision from these charts

Considering the result we found from the graph in fig 5.13 to 5.18, this can be concluded that the input data set is a sufferer from Coronary Artery Block since all the algorithms gave the highest score for CAD. However, in real life, when we tried to figure out the real disease, we found out that this input data was indeed a data set of CAD (the person whose data we used was a sufferer of CAD).

5.5 Result Analysis

Therefore, by examining the results given in table 5.4, we can see that for predicting CAD it is best to use Naive Bayes Classifier since it has the highest score of 94% accuracy among all the algorithms. For predicting Myocardial Infarction, Decision Tree Classifier worked the best. The score is 96%. For Sinus Tachycardia all the algorithm has a score of 95% except Nearest Neighbour. Decision Tree Classifier also worked well for Sinus Bradycardia with a score of 95%. And Lastly for Right Bundle Branch Block, Logistic Regression Classifier scored the highest, 96%.

Table 5.4 Best algorithm to use for individual disease and their score

Disease Name	Best Algorithm	Score
Right Bundle Branch Block	Logistic Regression	96%
Myocardial Infarction	Decision Tree	96%
Sinus Tachycardia	All except NN	95%
Sinus Bradycardia	Decision Tree	95%
CAD	Naive Bayes	94%

5.6 Comparison

In order to compare our models, we used the same Cross Validation setting which was used by the Bilkent University [15]. We used 10-fold cross validation technique. By this technique we have divided the data set into 10 subsets. We used one set as testing data and rest of the 9 sets as training set. This process was repeated 10 times once for each subset being the test set. Then we found the average of the 10 results.

In the experiment done by the Bilkent University, they used a algorithm called VFI5. The accuracy they found was only by using this algorithm which was 62% They found an accuracy of 68% using genetic algorithm and VFI5. Their goal was to classify every disease. In contrast to their experiment, we further did 2 more experiments. Firstly we divided the data set into two groups. One group contain all the normal arrhythmia instances and other contains all the abnormal arrhythmia instances. We took all the disease instance into one class. We ran 6 different algorithms. The lowest accuracy that we found is from Nearest Neighbour which is 68% and the highest we found is from Support Vector Machine which is 78%. On the other hand, when we isolated a single disease with the normal arrhythmia, our accuracy increases drastically. We got an accuracy result between 85% to 92% by classifying individual disease with different algorithms.

The cross validation accuracy scores were found automatically by our algorithm, however we had to use the equation no 5.1 to calculate the accuracy score for in case of random train-test split whenever needed.

$$ACC = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN} \tag{5.1}$$

Chapter 6

Conclusion and Future Works

This chapter gives a conclusion on our research and has information about our future work plan.

6.1 Conclusion

The main aim of our research was to distinguish between normal and abnormal heartbeat by means of machine learning. In addition we figured out which algorithm works best for predicting a particular disease. We tried collecting data set from different mediums a couple of times, but this was exceptionally hard. Therefore, we used the data set we found in UC Irvine Machine Learning Repository database to do our thesis. We faced a lot of trouble in some particular parts while doing our research, however we are happy that we managed to overcome them and reach at this level. From our research, we found out that Logistic regression is the best algorithm to be used for Right Bundle Branch Block which gives an accuracy of 96%. Decision Tree gives the best result for Myocardial Infarction which is also 96%. All the algorithm except Nearest Neighbour can be used in case of Sinus Tachycardy as all of them have a score of 95% equally. Decision tree is the best algorithm for Sinus Bradycardy and it also has a score of 95%. Lastly, Naive Bayes gives the best score for Coronary Artery Disease which is 94%.

6.2 Future Work Plan

As we have been dealing with few diseases that results from a slight change in any part of the ECG graph, we hope to use this model for all the possible diseases that results from such abnormality. We want to figure out which algorithm gives the most accurate result in

predicting a particular disease. Most importantly, we aim to build a health monitoring system which will contain all these information. A patient will be able to do his ECG from his house and his data will be stored in the cloud. His doctor will also be connected to the cloud and he will have access to all the data of his patient. The doctor can monitor his patient by the aid of these data from a distant place and can advice the patient in case of need. We think this can be a really helpful innovation for both the doctor and patient. Early actions can be taken in case of emergency and even when the patient can not visit the doctor for personal purposes. This will save time and will also be of great help in case of emergency.

References

[1] Abdullah, A. (2014). *ECG in Medical Practice*. JP Medical Ltd.

[2] Asfaqul Islam, A. (2015). A real-time ecg warning system on myocardial infarction, hyperkalemia and atrioventricular block.

[3] Benarroch, E. E. (2012). Postural tachycardia syndrome: a heterogeneous and multifactorial disorder. In *Mayo Clinic Proceedings*, volume 87, pages 1214–1225. Elsevier.

[4] Bishop, C. M. (2012). Pattern recognition and machine learning, 2006. , 60(1):78–78.

[5] Breiman, L. (2017). *Classification and regression trees*. Routledge.

[6] Brownlee, J. (May 23, 2018). A gentle introduction to k-fold cross-validation.

[7] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

[8] Defazio, A., Bach, F. R., and Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *CoRR*, abs/1407.0202.

[9] Dheeru, D. and Karra Taniskidou, E. (2017). UCI machine learning repository.

[10] Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

[11] Gillam, P. (2015). Cardiac cycle and the human heart: A* understanding for igcse biology 2.63 2.64.

[12] $\mathrm{gk}_{(}Jan25,2017).Textclassificationusingneuralnetworks.$

[13] Goldberger, A. L. (1991). Is the normal heartbeat chaotic or homeostatic? *Physiology*, 6(2):87–91.

[14] Grzymala-Busse, J. W. (1993). Selected algorithms of machine learning from examples. *Fundam. Inform.*, 18:193–207.

[15] Guvenir, H. A., Acar, B., Demiroz, G., and Cekin, A. (1997). Supervised machine learning algorithm for arrhythmia analysis. *Computers in cardiology*, pages 433–436.

[16] Hastie, T., Tibshirani, R., and Friedman, J. (2009). Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer.

[17] Immanuel, J. J. R., Prabhu, V., Christopheraj, V. J., Sugumar, D., and Vanathi, P. (2012). Separation of maternal and fetal ecg signals from the mixed source signal using fastica. *Procedia Engineering*, 30:356–363.

[18] Jambukia, S. H., Dabhi, V. K., and Prajapati, H. B. (2015). Classification of ecg signals using machine learning techniques: A survey. In *Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in*, pages 714–721. IEEE.

[19] James Heilman, M. (2012). File:sinus bradycardia lead2.svg.

[20] Kishor, N. (Apr 2, 2018). K-nearest neighbors - the laziest machine learning technique.

[21] Klabunde, R. E. (2016). Electrophysiological changes during cardiac ischemia.

[22] Lewis, T. (2015). Human heart: Anatomy, function & facts. *LiveScience. TechMedia Network*, 7.

[23] Mader, S. S. (1997). *Understanding human anatomy & physiology*. WCB/McGraw-Hill.

[24] Madhero88 (unknown). File:normal ecg 2.svg.

[25] Mahajan, S. (2014). Reinforcement learning: A review from a machine learning perspective. *International Journal*, 4(8).

[26] MD, C. S. (unknown). Myocardial infarction (heart attack).

[27] Mitchell, T. M. et al. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877.

[28] Njeri, R. (Sep 2, 2017). What is a decision tree algorithm?

[29] Νεοφύτου, N. (2012). *ECG event detection & recognition using time-frequency analysis*. PhD thesis.

[30] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[31] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

[32] Rad, A. (2006). This is a file from the wikimedia commons file:right bundle branch block ecg characteristics.png.

[33] Raschka, S. (2014). Naive bayes and text classification i-introduction and theory. *arXiv preprint arXiv:1410.5329*.

[34] Ray, S. (2017). Essentials of machine learning algorithms (with python and r codes). *Analytics Vidhya. Re*.

[35] Ray, S. (2018). Improve your model performance using cross validation (in python and r). *Analytics Vidhya. Re*.

[36] Richards, G. (2 July 2008). File:logistic-curve.svg.

[37] Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge university press.

[38] Roopa, C. and Harish, B. (2017). A survey on various machine learning approaches for ecg analysis. *Int J Comput Appl*, 163(9).

[39] Salzberg, S. L. (1994). C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, 16(3):235–240.

[40] Sarle, W. S. (1996). Stopped training and other remedies for overfitting. *Computing science and statistics*, pages 352–360.

[41] Schmidt, M., Le Roux, N., and Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112.

[42] Scoville, H. (2018). Evolution of the human heart.

[43] Soman, T. and Bobbie, P. O. (2005). Classification of arrhythmia using machine learning techniques. *WSEAS Transactions on computers*, 4(6):548–552.

[44] Taggart, A. J., DeSimone, A. M., Shih, J. S., Filloux, M. E., and Fairbrother, W. G. (2012). Large-scale mapping of branchpoints in human pre-mrna transcripts in vivo. *Nature structural & molecular biology*, 19(7):719.

[45] Tan, D. C. (2016-2018). Transport in animals (igcse biology syllabus).

[46] Thygesen, K., Alpert, J. S., White, H. D., et al. (2007). Universal definition of myocardial infarction. *Journal of the American College of Cardiology*, 50(22):2173–2195.

[47] unknown (unknown). Normal coronary blood flow in heart.

[48] VA, M. and on behalf of the U.S. Preventive Services Task Force* (2012). Screening for coronary heart disease with electrocardiography: U.s. preventive services task force recommendation statement. *Annals of Internal Medicine*, 157(7):512–518.

[49] Vecht, R., Gatzoulis, M. A., and Peters, N. (2009). *ECG diagnosis in clinical practice*. Springer Science & Business Media.

[50] Weinhaus, A. J. and Roberts, K. P. (2005). Anatomy of the human heart. In *Handbook of cardiac anatomy, physiology, and devices*, pages 51–79. Springer.

[51] Zimmerman, M. W. (2004). *Classification of ECG ST Events as Ischemic or Non-Ischemic Using Reconstructed Phase Spaces*. PhD thesis, Marquette University.