AN EFFECTIVE TECHNIQUE ON EARLY HEART DISEASE PREDICTION

TANZINA RAHMAN HERA

ROLL: BKH1711MSc111F

SESSION: 2016-2017

MASTER OF SCIENCE (INFORMATION AND COMMUNICATION ENGINEERING)

NOAKHALI SCIENCE AND TECHNOLOGY UNIVERSITY

AN EFFECTIVE TECHNIQUE ON EARLY HEART DISEASE PREDICTION

TANZINA RAHMAN HERA

This thesis proposal is submitted in fulfilment of the requirements for the award of the degree of Master of Science in Information and Communication Engineering

Department of Information and Communication Engineering
NOAKHALI SCIENCE AND TECHNOLOGY UNIVERSITY

December 2020

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Master of Science in "An Effective Technique On Early Heart Disease Prediction".

____

Professor Dr. Md. Ashikur Rahman Khan
Chairman
Department of Information and Communication Engineering
Noakhali Science and Technology University,
Noakhali- 3814, Bangladesh.
Department of ICE, NSTU

CO-SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Master of Science in "An Effective Technique On Early Heart Disease Prediction".

____

Zayed- Us- Salehin
Assistant Professor
Department of Information and Communication Engineering
Noakhali Science and Technology University,
Noakhali- 3814, Bangladesh.

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is my won except for quotations and summaries, which has been duly acknowledged. The thesis has not been accepted for any degree and is not concurrently submitted for award of other degree.

Tanzina Rahman Hera

Department of ICE, NSTU

Roll No: BKH1711MSc111F

ACKNOWLEDGMENTS

Firstly I am expressing my heartiest thanks and gratefulness to Almighty for divine blessings, which has gave me the courage to begin this research successfully. I feel grateful to and wish to acknowledgement my profound indebtedness to my Supervisor Professor Dr Md. Ashikur Rahman Khan, Chairman of the Department of Information & Communication Engineering, Noakhali Science and Technology University. Deep knowledge and keen interest of Professor Dr Md. Ashikur Rahman Khan in the field of Artificial Intelligence influenced me to carry out to start this thesis. His endless patience, scholarly guidance, continual encouragement, constructive criticism and constant supervision have made it possible to start this thesis. I would like to convey gratitude to all my course teachers whose teaching helps me a lot to start this thesis work. Lastly I also grateful to my family and classmates for giving me continuous support.

ABSTRACT

Heart disease is the most threatened issue in a human body. Most of the people in the world are affected by heart disease which increases the death rate of humans considerably. The detection of heart disease is the most trivial task for medical researchers which cannot be done more accurately. The early detection, manual prediction and prevention is a complex task and troublesome. So there is an urgent need for a well-designed method for the early detection of heart disease. Thus it is required to implement the automated system by considering the recent computer technologies which can help medical researchers to diagnosis heart disease firstly and accurately. Now a day's artificial neural network has been widely used as a tool for solving many decision modelling problems. This research enlightened a number of structures in Artificial Neural Network by varying the configurations of neural network algorithms like Radial Basis Function Neural Network, Recurrent Neural Network, Decision tree classifier and Support Vector Machine Neural Network. This research also compared those Neural Network models on the early detection of heart disease and chose the best neural network model among them to detect the heart disease early. The Network Structure that showed more accuracy and efficiency is chosen among them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ADA	American Diabetes Association
ANN	Artificial Neural Network
ASFT	Abdominal Subcutaneous Fat Thickness
BMI	Body Mass Index
BP	Back Propagation
BPN	Back Propagation network
CBFCM	Case Based Fuzzy Cognitive Map
GA	Genetic Algorithm
GP	Genetic Programming
MAE	Mean Absolute Error
MSD	Mean Squared Deviation
SVM	Support Vector Machine
RBF	Radial Basis Function

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

The most complicated and complex task in the field of medical sciences is the prediction of heart disease. Heart is the most vital organ of the human body. There is an intense need of the prediction level and seriousness of heart disease that will provide an accurate treatment to the patients. Heart disease can be referred to various conditions that lead to abnormal functioning of heart, which may involve blood vessels, arteries etc. Effective diagnosis of heart disease results in an appropriate treatment to a patient. This requires a deep study of cardiovascular analysis of the patient that includes symptoms such as chest pain, chest tightness, chest pressure, and discomfort in breathing, numbness etc. The cardiovascular diagnosis involves certain decisions to take based upon the health history and the clinical test results of a person. The process of decision-making is a challenging task to the medical practitioners, which has to be done accurately, and efficiently where a mere negligence may lead to the life risk of a patient. According to the World Health Organization (WHO), 17.9 million people die each year from Cardiovascular Disease, an estimated 31% of all deaths worldwide and 85% are due to heart attack and stroke [1]. In 2018 WHO has marked Heart Disease as one of the biggest causes of death all around the World. Populations most affected are from low in addition, middle-income countries like Bangladesh, where 80% of these deaths occur.

1.2 PROBLEM STATEMENT

A major challenge facing healthcare organizations (hospitals, medical centres) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. The diagnosis of disease is a difficult and tedious task in medical field. Diagnosis of heart disease from various factors or symptoms is a multi-layered issue which may lead to false presumptions and unpredictable effects. Only human intelligence alone is not enough for proper diagnosis. A number of difficulties will arrive during diagnosis, such as less accurate results, less experience, time dependent performance, knowledge up gradation difficulties as shown in Figure 1.

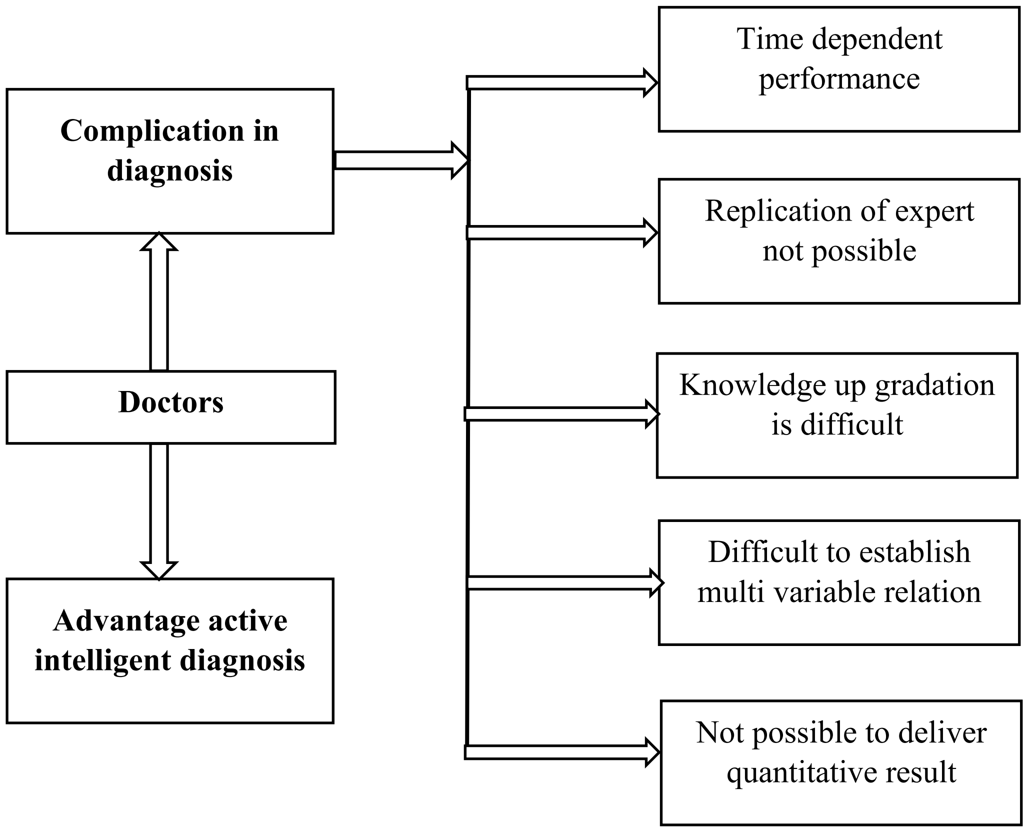

Figure 1.1 Complexity in diagnosis with doctors

One of the major topics in artificial intelligence is the discovery of interesting models of artificial neural network by applying various configurations based on a dataset. A majority of areas related to medical services such as prediction of effectiveness of surgical procedures, medical tests, medication and the discovery of relationships among clinical and diagnosis data effectiveness of medical treatments can be estimated by developing the neural network models. Disease diagnosis is usually based on signs, symptoms and physical examination of a patient. Almost all the doctors are predicting heart disease by learning and experience. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and decision support systems. Hospitals and clinics accumulate a huge amount of patient data over the years. These data provide a basis for the analysis of risk factors for many diseases. For example, we can predict the level of heart attack to find patterns associated with heart disease. The fact of diagnosis process of detecting heart disease is comparatively time consuming and laboured job for a person to go to the hospital and do all the tests and wait for getting the test report to know whether he/she has heart disease or not. For this reason more research is needed to find the most effective way of screening for heart disease. Therefore, there is an urgent need to adopt a cost-effective screening and diagnostic method of heart disease to reduce further complications. This research includes the method of analysing heart disease by using various Artificial Neural Network models like Multilayer perceptron neural network (MLPNN), Radial Basis Function Neural Network (RBFNN), Support Vector Machine (SVM) Network, Decision tree classifier, K-nearest neighbor classifier and exploring the most efficient model among them.

1.3 OBJECTIVES

The objectives of this research are as follows:

I. To develop cost-effective screening and diagnostic methods for analysing heart disease and to establish different Artificial Neural Network (ANN) models and structures for the early detection of heart disease.
II. To study the performance of these neural network models by applying the same dataset for a specific number of patients.
III. To conduct a comparison among those neural network structures on basis of finding the best network that meets higher accuracy.
IV. Finally, to explore the most efficient ANN structure for the early prediction of heart disease.

1.4 SCOPE OF THE STUDY

Heart disease is one of the most hazardous diseases to human, which shows the way to death all over the world since 15 years. Many researches have been done with the techniques of knowledge discovery in various fields for Heart Disease prediction and have shown the acceptable levels of accuracy. By investigating the survey of those accuracy levels, this research paper is proposed to help doctors not only to diagnose and predict the heart disease by achieving accuracy levels but also helps to prescribe the medicine successfully according to the predicted disease. In the paper, assessment is done by the methodology of ANN (Artificial neural network) by training the dataset, testing and make the production dataset. The result of the evaluation of ANN implemented for prescribing medicine from the history of old patients with Generalized Regression Neural Network and Radial basis function successfully. To accomplish the task of constructing different neural network structures, training, testing and validating those structures a neural network development environment called NEUROSOLUTIONS is used. MATLAB software is also being used. They combines a modular, icon-based (component-based) network design interface with an implementation of advanced learning procedures such as conjugate gradients, Levenberg-Marquardt and backpropagation through time. The software is used to design, train and deploy neural network (supervised learning and unsupervised learning) models to perform a wide variety of tasks such as data mining, classification, function approximation, multivariate regression and time-series prediction.

1.5 THESIS OUTLINE

The paper consists of five sections. The arrangement is as follows. In chapter 1, it provides a detailed overview, the problem statement of the work and the reason behind selecting the topic for thesis. In the chapter 2, it showed some related works. It gives the information about the contribution of several researches in this field. Next is chapter 3, which introduces the dataset and the methods and explains deep understanding about the algorithms and evaluating criteria. Then the chapter 4 describes experimental plans and prediction test results on evaluation criteria. Finally, in Chapter 5, paper will end up as allowing knowing why this study is a valuable for effective techniques in medical diagnosis in the conclusion.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

The need for an accurate predictor for the heart disease is highly required. Not only this but also a predictor that is extremely automated and with less human interference is obligatory. A heart disease predictor should meet the following specification; efficient modelling, applicability, accuracy and be trusted. It should be compatible with various diagnostic techniques. Many neural network prediction techniques like the Multi-layer perceptron (MLP), Radial basis function (RBF), Support vector machine (SVM), Recurrent neural network, Self-organizing feature map network are the most commonly used in the early detection of heart disease. Here are the discussion for some of the related works.

2.2 RELATED WORKS

Kaan Uyara and Ahmet Ilhana proposed a genetic algorithm (GA) based trained recurrent fuzzy neural networks (RFNN) to diagnosis of heart diseases [2]. The University of California Irvine (UCI) Cleveland heart disease dataset was used in that study. Out of total 297 instances of patient data, 252 were used for training and 45 of them were chosen to be the testing. The results showed that 97.78% accuracy was obtained from testing set.

Zeinab Arabasadi et al., proposed a highly accurate hybrid method for the diagnosis of coronary heart disease [3]. The method was able to increase the performance of neural network by approximately 10% through enhancing its initial weights using genetic algorithm which suggested better weights for neural network. Making use of such methodology, they achieved accuracy, sensitivity and specificity rates of 93.85%, 97% and 92% respectively on Z-Alizadeh Sani dataset.

Sagir and Sathasivam used ANFIS Matlab's built-in model and an ANFIS model with Levenberg-Marquardt algorithm to predict the potentials of the heart diseases [4]. This paper reported the design and evaluation of the classification performances of two discrete Adaptive Neuro Fuzzy Inference System models, ANFIS Matlab's built-in model (ANFIS_ LSGD) and a newly ANFIS model with Levenberg-Marquardt algorithm (ANFIS_LSLM). An attempt was also done to specify the effectiveness of the model's performance measuring accuracy, sensitivity and specificity. A comparison of the two models in terms of training and testing with the Statlog-Cleveland Heart Disease dataset have also been done.

Jayshril S. Sonawane and D. R. Patil presented a prediction system for heart disease using multilayer perceptron neural network [5]. The neural network in that system accepts 13 clinical features as input and it was trained using back-propagation algorithm to predict that there was a presence or absence of heart disease in the patient with highest accuracy of 98% comparative to other systems. The accuracy thus obtained with this system showed that it was better and efficient than other systems.

Aditi Gavhane et al., developed an application which can predict the vulnerability of a heart disease given basic symptoms like age, sex, pulse rate etc [6] . The machine learning algorithm neural networks had proven to be the most accurate and reliable algorithm and hence used in that proposed system.

Tulay karawulanet al., proposed a heart disease prediction system which used artificial neural network backpropagation algorithm [7] . 13 clinical features were used as input for the neural network and then the neural network was trained with backpropagation algorithm to predict absence or presence of heart disease with accuracy of 95%.

Sarath Babu et al., used K-means algorithm, MAFIA algorithm and Decision tree classification in heart disease prediction, applied the data mining technique to heart disease treatment [8]. It provided as reliable performance as that achieved in diagnosing heart disease.

Jyoti Soni et al., intended to provide a survey of current techniques of data extraction from databases using data mining techniques that were used in Heart Disease Prediction [9]. The techniques used there were Naive Bayes, Decision List and KNN. There the Classification based on clustering was not performing well.

P.K Anooj et al., presented a weighted fuzzy rule-based system for the diagnosis of heart disease, the system automatically retrieved knowledge from the patient's data [10]. The proposed system for the prediction of heart disease consisted of two phases: automated approach for the generation of weighted fuzzy rules and developing a fuzzy rule-based decision support system. The weighted fuzzy rules were used to build the system using Mamdani fuzzy inference system.

Nidhi Bhatla et al., proposed various data mining techniques used to analyse heart disease prediction [11] . The observations revealed that neural networks with 15 attributes had outperformed over all other data mining techniques. Another conclusion from the analysis was that decision tree had also shown good accuracy with the help of genetic algorithm and feature subset selection.

Aditya Methaila et al., desired to use data mining Classification Modelling Techniques such as Decision Trees, Naive Bayes and Neural Network in addition to weighted association Apriori algorithm and MAFIA algorithm in Heart Disease Prediction [12].

Shimpy goyal et al., discussed data mining techniques to predict heart disease based on k-means and apriori algorithm [13]. The researchers also presented the challenges in detecting and diagnose the diseases and analysed results of research.

M. Akhil jabbar et al., presented an efficient associative classification based genetic algorithm used for heart disease prediction [14]. The main reason for using genetic algorithm to predict disease from large dataset was that the actual size of data is to get best attribute set. There was certain limitation in the prediction of heart disease using data mining approach.

R Sethukkarasi et al., provided novel neuro fuzzy techniques and pre-processed by genetic algorithm [15] . There were four layered fuzzy neural networks used. A radial basic function neural network was constructed with five inputs, training and normalisation in hidden layer and output layer with one node.

Mohammed Abdul Khaleel et al., presented a method to find the diseases with Apriori data mining technique [16] . A graphically representation was also used to visualize the techniques. A prototype was produced to demonstrate the efficiency of that method. It resulted that the prototype could be useful in real world.

Boshra Bahrami et al., checked the different classification techniques in diagnosis heart disease [17]. Classifiers such as Decision Tree, KNN, and Naive Bayes were used to divide dataset. After the classification and performance evaluation the Decision tree was considered as the best classifier for heart disease diagnosis from the dataset.

Deepika N et al., used Pruning Classification Association Rule (PCAR) for detecting heart disease [18]. Pruning Classification Association Rule comes from Apriori algorithm. The proposed method deleted minimum frequency item with minimum frequency item sets and deleted infrequent items from item sets then the frequent item set was discovered.

Sonam Nikhar and A.M. Karandikar intended to provide a detailed description of Naive Bayes and decision tree classifier that were applied particularly in the prediction of Heart Disease [19]. Some experiment had been conducted to compare the execution of predictive data mining technique on the same dataset and the consequence revealed that Decision Tree outperforms over Bayesian classification.

Niti Guru et al., introduced the diagnosis of Heart Disease, Blood Pressure and diabetes with the aid of neural networks [20]. Experiments were carried out on a sampled data set of patient's records. The Neural Network was trained and tested with 13 input variables such as Blood Pressure, Age, Angiography's report and the like. The supervised network had been advised for diagnosis of heart diseases. Training was carried out with the help of back propagation algorithm. Whenever unfamiliar data was inserted by the doctor, the system identified the unknown data from comparisons with the trained data and produced a catalogue of probable diseases that the patient was vulnerable to.

Nishara Banu et al., published a research paper "Disease Forecasting System Using Data Mining Methods [21]. In that article, the pre-processed data was clustered using clustering algorithms as K-means to gather relevant data in a database. Maximal Frequent Item set Algorithm (MAFIA) was applied for mining maximal frequent model in heart disease database. The regular patterns could be classified into different classes using the C4.5 algorithm as training algorithm using the concept of information entropy. The result demonstrated that the designed prediction system was capable of predicting the heart attack successfully.

T. John Peter and K. Somasundaram presented a paper, "An Empirical Study on Prediction of Heart Disease using classification data mining technique" [22] . In that research paper, the use of pattern recognition and data mining techniques were used for the prediction of risk in the medical domain of heart disease.

Shamsher Bahadur Patel., Pramod Kumar Yadav and Dr. D. P.Shukla presented a research paper, "Prediction of the Diagnosis of Heart Disease Patients Using Classification Mining Techniques" [23]. In that research paper, the health care industry, the data mining was mainly utilized for the prediction of heart disease. The objective of their work was to predict the diagnosis of heart disease with a reduced number of attributes using Naive Bayes, Decision Tree.

SP Rajamhoana et al., analyzed various research works done on heart disease prediction and classification using various machine learning and deep learning techniques and to conclude which techniques were more effective and accurate [24].

Jayshril S. Sonawane and D. R. Patil presented a prediction system for heart disease using multilayer perceptron neural network [5]. The neural network in that system accepts 13 clinical features as input and it was trained using back-propagation algorithm to predict that there was a presence or absence of heart disease in the patient with highest accuracy of 98% comparative to other systems. The accuracy thus obtained with this system shows that it was better and efficient than other systems.

George et al. had proposed decision support system for dementia patients using support vector machines to define and detect agitation transition [25]. In that system two new SVM architectures were presented, which were applied to the detection of agitation and agitation transition. This approach gave the accuracy of 91.4%, which was higher as compared with 90.9% for the traditional SVM.

The automated recognition of obstructive sleep apnea syndrome by using SVM classifier was proposed by Haitham and Alan [26]. In that approach, for the detection of the syndrome they computed the features from the phase and magnitude of the signals obtained from thoracic and abdominal respiratory effort and evaluated the classification of whole night normal and apneic epochs. The highest accuracy of this system was 95%.

A system for diagnosis of heart disease that was based on support vector machine along with sequential minimal optimization algorithm was presented [27]. In that system the network structure of Radial Basis Function was also used and it was trained using Orthogonal Least Square algorithm and applied to the dataset based on Indian patients. The result showed that the SVM was equivalently as good as compared to Radial basis function in detection of heart disease with accuracy of 86.42%.

Tsai and Watanabe had classified myocardial heart disease from ultrasonic images by optimizing the fuzzy membership functions by using genetic algorithm based method [28]. In that method by using the texture features obtained from ultrasonic images, the Gaussian distributed membership function was constructed and the genetic algorithm based fuzzy classifier was used in classification. In that technique 96% of classification accuracy was achieved.

Genetic algorithm was used in another approach by Anbarasi et al. where number of tests that were to be conducted by patient was reduced by determining the attributes that involved in the prediction of heart disease [29]. In that approach three classifiers were used and these classifiers were fed with reduced attributes, but the system took more time for model construction.

Haung and Wang proposed a system using Support Vector Machine where feature selection and optimization of Support Vector Machine parameter was done using genetic algorithm [30]. This system uses less number of input parameters for support vector machine and it was evaluated on 11 real world datasets with improved accuracy of 89.6% .

Rajkumar and Sophia proposed the use of Data mining algorithm in diagnosis of heart disease with an accuracy of 52.33% [31]. They combined the ECG attributes and clinical symptoms to detect the heart disease. The algorithms used by this system were Naive Bayes algorithm, Decision tree algorithm and KNN algorithm.

Kumaravel et al., proposed automatic diagnosis system for heart diseases using neural network. In that system ECG data of the patients was used to extract features and 38 input parameters were used to classify 5 major types of heart diseases with accuracy of 63.6 - 82.9% [32].

Anchana Khemphila et al., introduced a classification approach using Multi-Layer Perceptron (MLP) with Backpropagation learning algorithm and a feature selection algorithm along with biomedical test values to diagnose heart disease [33]. The Artificial neural networks was used to classify the diagnosis of patients. Thirteen attributes were reduced to 8 attributes. The accuracy differed between 13 features and 8 features in training data set is 1.1% and in the validation data set was 0.82%.

A. T. Sayad and P. P. Halkarnikar implemented a diagnostic system of heart disease detection [27]. For diagnosis of heart disease significantly 13 attributes were used in proposed system as per the medical literature. They employed the Multi-layer Perceptron Neural Network with Back-propagation as the training algorithm. The results thus obtained illustrated that the designed diagnostic system was capable of predicting the risk level of heart disease effectively. This system acted as promising tool for diagnosis of heart disease.

Shahid Mehmood Awan et al., used data and attributes from the UCI repository [34]. In this paper, they enlightened the number of techniques in Artificial Neural Network (ANN). The accuracy was calculated and visualized such as ANN gives 94.7% but with Principle Component Analysis (PCA) accuracy rate improved to 97.7% [35].

2.3 TABLE OF PREVIOUS WORK

In this section the previous works related to this research is summarized by a table 2.1

Table 2.1 Table of related work

Sl No.	Title	Author	Year	Considered Parameters	Source of Data	Implemented Methods/ Techniques	Result
1	Diagnosis of heart disease using genetic algorithm based trained	Kaan Uyara, Ahmet Ilhana	2017	Oldpeak, Slope, Thal, ECG, Ca	Kaggle	Recurrent fuzzy neural networks (RFNN)	97.7%
2	Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm	Zeinab Arabasadi	2019	Sex, Cp, Chol, Trestbps, Fbs	Z-Alizadeh Sani dataset	Genetic algorithm	92%
3	Diagnosis of heart disease using genetic algorithm based trained	Sagir, Sathasivam	2017	Chol, Trestbps, Fbs,	Kaggle	Adaptive Neuro Fuzzy model	88%
4	Prediction of heart disease using multilayer perceptron neural network	Jayshril S. Sonawane and Patil	2014	Oldpeak, Slope, Thal, ECG, Ca	Kaggle	Multilayer perceptron	98%
5	Prediction of Heart Disease Using Machine Learning	Aditi Gavhane	2018	Chol, Trestbps, Fbs	Kaggle	Support Vector Machine	82%
6	Analysis of heart disease prediction using neural network	Tulay karawulan et		Age, Sex,Chol	UCI repository	Backpropagation algorithm	95%.

Sl No.	Title	Author	Year	Considered Parameters	Source of Data	Implemented Methods/ Techniques	Result
7	Heart disease diagnosis using data mining technique	Sarath Babu	2017	Oldpeak, Slope, Thal, ECG, Ca	UCI repository dataset	K-means algorithm, MAFIA algorithm and Decision tree	81%
8	Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction	Jyoti Soni	2017	Oldpeak, Slope, Thal, ECG, Ca	Kaggle	Naive Bayes, Decision List and KNN	88%
9	Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules	P.K Anooj	2019	Trestbps, Fbs, Thalach, Ca	Kaggle	Weighted fuzzy rule-based system	85%
10	An Analysis of Heart Disease Prediction using Different Data Mining Techniques	Nidhi Bhatla	2012	Age, Sex, Ca,Fbs	UCI repository dataset	Descision tree	87%
11	Early Heart Disease Prediction Using Data Mining Techniques	Aditya Methaila	2014	Sex, Age, Ca, Bp	UCI repository dataset	Decision Trees, Naive Bayes	91%
12	Heart disease diagnosis using data mining technique	Shimpy goyal	2019	Age, Sex, Cp, Ca, ECG	Kaggle	K-means and apriori algorithm	88%
13	Heart disease analysis using genetic algorithm	M. Akhil jabbar	2018	Age, Sex, Cp, Ca, ECG	Kaggle	Genetic algorithm	

Sl No.	Title	Author	Year	Considered Parameters	Source of Data	Implemented Methods/ Techniques	Result
14	Heart disease detection using neural network	R Sethukkarasi	2018	Age, Sex, Ca,Fbs	Kaggle	Neuro fuzzy techniques	
15	Association Rule Mining based on a Modified Apriori Algorithm in Heart Disease Prediction	Mohammed Abdul Khaleel	2018	Age, Sex, Ca,Fbs	UCI repository	Apriori data mining technique	
16	A Literature Survey on Applications of Data Mining Techniques to Predict Heart Diseases	Boshra Bahrami	2015	Age, Sex, Ca,Fbs, ECG	UCI repository	Decision Tree, KNN, and Naive Bayes	
17	A Literature Review on Heart Disease Prediction Based on Data Mining Algorithms	Deepika N	2018	Age, Sex, Ca,Fbs	UCI repository	Pruning Classification Association Rule (PCAR)	
18	Review of heart disease analysis using neural network	Sonam Nikhar and A.M. Karandikar	2019	Age, Sex, Bp, Bs,ECG, Ca	UCI repository	Naive Bayes and decision tree classifier	
20	Disease Forecasting System Using Data Mining Methods	Nishara Banu	2017	Sex, Age, Ca, Cp, Fbs, ECG, Bp, slope	UCI repository	K-means algorithm	92%

Sl No.	Title	Author	Year	Considered Parameters	Source of Data	Implemented Methods/ Techniques	Result
21	Neural Network Based Intelligent System for Predicting Heart Disease	V. B. K. Subhadra	2019	Sex, Age, Ca, Cp, Fbs, Bp	Kaggle	Multilayer perceptron	95%
22	Heart Disease Classification Using Neural Network and Feature Selection	V. B. Anchana Khemphila	2011	Age, Sex, Cp, Ca, Fbs	Kaggle	Genetic algorithm	85%
23	Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection	M.Akhil Jabbar	2013	Oldpeak, Slope, Thal, ECG, Ca	UCI repository	Decision trees	88%

2.4 SUMMARY

By summarizing these research paper it is shown that authors mainly used the dataset from the UCI repository or from kaggle. They had find a good accuracy though but this research uses the real dataset collected from different hospitals. So this research is more preferable comparing to them and it will be a good system to apply it in the next future.

CHAPTER 3

METHODOLOGY

3.1 INTRODUCTION

Artificial Neural Network (ANN) model is becoming more popular in detecting various diseases so that the risk of the specific disease decreases. The early detection or classification of any disease helps anyone to decrease the risk of the disease. When c1assification is the goal, the artificial neural network (ANN) models often deliver close to the best fit. This paper enlightened a number of techniques in Artificial Neural Network (ANN). A model of analysing the heart disease with high accuracy, less complex and has efficient performance is urgently needed. To address this need, the method identified in this study offers every person the opportunity to know his or her risk early. This chapter will discuss about the proposed model, workflow diagram and about dataset.

3.2 PROPOSED NEURAL NETWORK MODEL

Heart disease is increasing rapidly due to number of reasons. Heart disease diagnosis plays a more important role in real world where most of people including both men and women of all ages affected lot. If the prediction of cardiac arrest (dangerous conditions of heart) in the early stages is possible, it will be very helpful to cure this disease. Although doctors and health centres collect data daily, but mostly are not using artificial neural network techniques to extract the knowledge that can be very useful in prediction. Heart disease would cause more serious threat to the people which needs to be diagnosed more efficiently. It is required to diagnosis the heart disease early in order to ensure the early treatment procedure thus the unwanted health illness can be avoided. The main goal of this research is to introduce the model which can predict the heart disease in the automated manner as soon as possible with increased accuracy level. The sample a neural network is shown by the figure 3.1 given below. It starts with the input parameters of data from various sources. Some processing are done to the parameters. After the pre-processing, the dataset is fed to the neural network model. There may be several hidden layers. The input parameters are combined with several weight vectors. By considering the result of each model, this study will move forward to the next step called analysis.

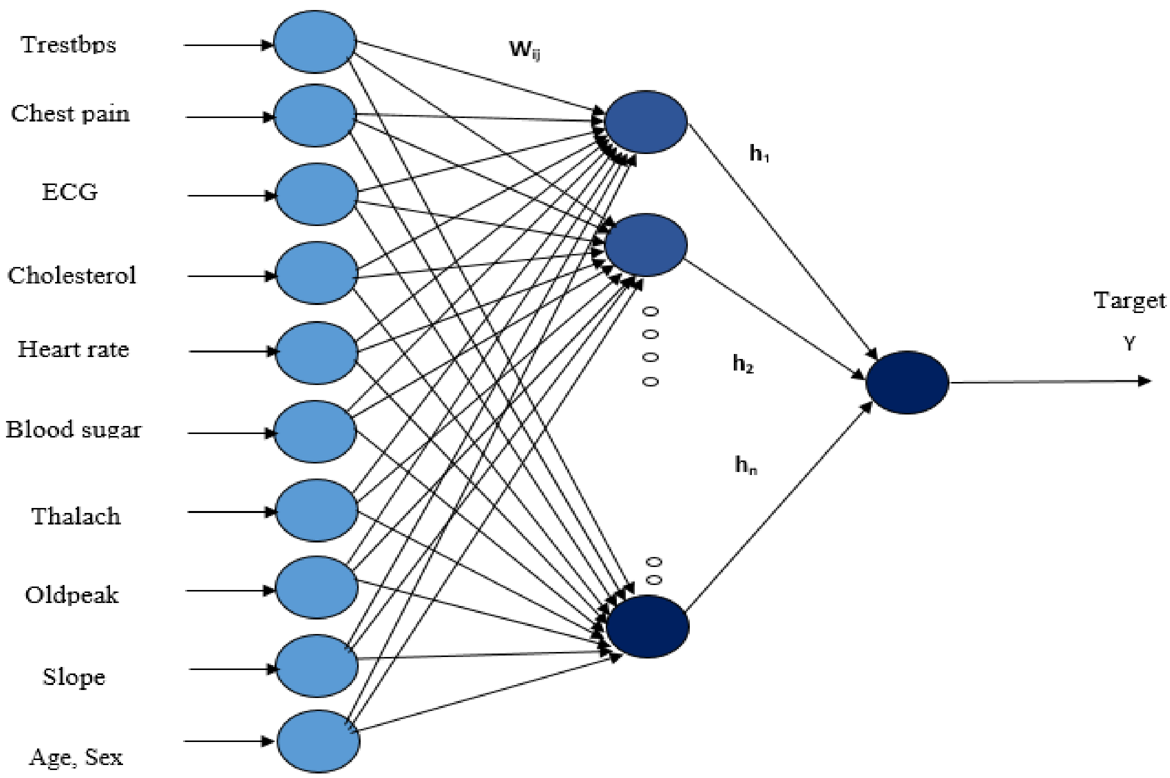

Figure 3.1 A conventional neural network structure

3.2.1 DATA COLLECTION

Data of heart disease risk factor is collected from patients of hospitals, diagnostic centre and clinic centre. The efficiency of the proposed model largely depends upon the quantity and quality of the data that is collected. The more the data could be collected, the more efficient model would be. But there were some limitations of collecting data from different sources. It was be a tough challenge. The data used to make the dataset, collected from Dhaka Medical College Hospital and two other private hospitals. The dataset have instances of patient data but 11 of them contain missing values. Table 1 shows the dataset attributes with their definition, value and ranges. In this study, 11 of the instances of the dataset containing missing entries are omitted. The diagnosis of heart disease attribute was categorized into two classes denoted as absence (num = 0) and presence (num = 1) of the heart disease. The dataset separated into three sub-sets for training (100 instances) and testing (100 instances) and validation (50 instances).

Table 1: Dataset Description

Attributes Name	**Description**	**Value**	**Range**
Sex	Gender	1 = Male 0 = Female	
Cp	Chest pain type	1: typical_angina 2: atypical_angina 3: non_anginal_pain 4: asymptomatic	
Trestbps	Resting blood pressure	in mm Hg on admission to the hospital	90 -120: normal 120-140: unusual 140 – 160: high >160: very high
Chol	Serum Cholesterol	in mg/dl	110-200: normal 200-240: borderline_high 240-250: high >250: very_high
Fbs	Fasting blood sugar > 120 mg/dl (fbs)	1 = true, 0 = false	1-true (> 120 mg/dl) 0-false (< 120 mg/dl)
Thalach	Maximum heart rate achieved	Positive integer	(60 - 202)
Exang	Exercise_induced_angina	1: yes 0: no	
Oldpeak	ST depression induced by exercise relative to rest	Positive integer	(-2.6 – 6.2)

Attributes Name	Description	Value	Range
Slope	The slope of the peak exercise ST segment	1: Upsloping 2: Flat 3:Downsloping	
Thal	Thalassemia	3: Normal 6: fixed_defect 7: reversible_defect	
ECG	Rhythm Q wave ST elevation T inversion	1: yes 2: no	
Ca	Number of major vessels (0-3) colored by flourosopy	Values 0, 1, 2, 3	
Target	Type of heart disease	0: No Heart disease 1, 2, 3, 4: Different levels of heart disease	

3.2.2 IMPLEMENTATION

The implementation module is consist of

I. **Data Pre-processing**

This is one of the most important states of the implementation module. A large amount of data have been collected from several sources but all of them would not be valid. There was unclear, duplicate data's and some instances have missing values. Considering these this study had to follow the following steps for data pre-processing.

a. Removing the unclear value
b. Removing the Duplicate value

c. Replacing the missing value
d. Normalize the dataset

II. **Training different neural network models**

This study have trained the neural network models with the dataset using six neural network classification algorithms.

III. **Testing neural network models**

After training the models, it evaluated the model performance for those classification algorithms. It have recorded these performance results for the analysis.

3.2.3 ANALYSIS

In this step, validation is performed to the dataset. Validation gave the performance characteristics and efficiency of the implemented models. The analysis module consisted of two states.

a. Compared the performance result: For each classification algorithm the accuracy is calculated and then compared it with others.
b. Identified the conclusion: Finally, after calculating all the selected neural network structure results, it found out a more efficient algorithm that outperforms all the others performance.

3.3 DIAGRAM OF PROPOSED METHODOLOGY

This research of analysing heart disease is summarized by the following figure 3.2 given below. It starts with the collection of data like collection of the patient detail. The data of the patients is then is analysed. Data was normalized to get better result. Then the input parameters were normalized. After preparing the training dataset the different artificial neural network was applied. This study used six neural network structures. The obtained results from them were then be compared. The network that gave the more accurate result is chosen. And heart disease detection with the selected network configuration is accomplished.

The fundamental steps in the implementation of heart disease detection that should be followed to apply ANN for the purposes of heart disease diagnosis with sufficient confidence are shown in figure 3.2

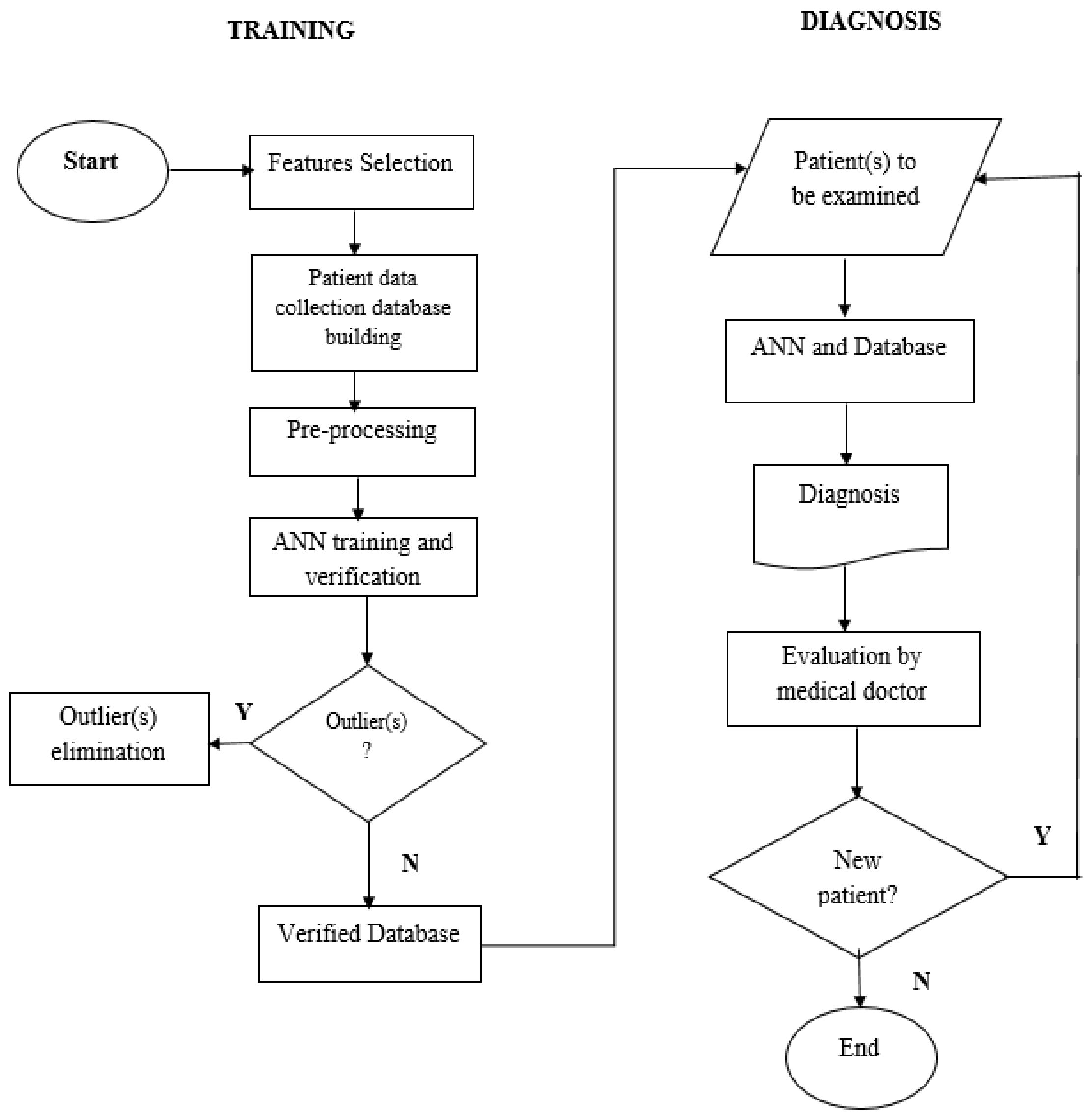

Figure 3.2 Workflow diagram of the implemented model

3.4 CLASSIFICATION ALGORITHMS

For the purpose of comparative analysis, seven neural network classification algorithms is used. The different algorithms predicted heart disease at different accuracy. The reason to choose these algorithms is based on their popularity.

1. Radial Basis Function (RBF)
2. Recurrent Network

3. Support Vector Machine (Linear SVM)
4. Ensemble classifier (Boosted tree)
5. Nearest neighbour classifiers (Medium KNN)
6. Decision tree (Simple tree)

3.4.1 RADIAL BASIS FUNCTION NEURAL NETWORK (RBFNN)

Radial basis function (RBF) networks are nonlinear hybrid networks typically containing a single hidden layer of processing elements. This layer uses Gaussian transfer functions, rather than the standard sigmoidal functions employed by MLPs. The centres and widths of the Gaussian's are set by unsupervised learning rules, and supervised learning is applied to the output layer.

ALGORITHM

Step 1: Define the number of hidden neurons 'K'

Step 2: Set the positions of RBF centers using K-means clustering algorithm.

Step 3: Calculate σ

Step 4: Calculate actions of RBF node

Step 5: Train the output

CODE

```
iw=cell2mat(net.iw(1));        %input weight

lw=cell2mat(net.lw(2,1));       %layerweight

ib=cell2mat(net.b(1));          %input bias

lb=cell2mat(net.b(2));          %layer bias

a= radbas(netprod(dist(iw,pn),ib));       %radial basis transfer function

a2=dotprod(net.lw{2,1},a);                 %product of first layer output with layer weights

a2=gadd(a2,net.b{2});                      %add with layer bias

a2=mapminmax('reverse',a2,ts);             %reverse the normalization as followed in the newrb
which was from 0.1 to 0.9
```

3.4.2 RECURRENT NETWORK

A recurrent neural network (RNN) is a class of artificial neural network where connections between nodes form a directed graph along a sequence. This allows it to exhibit dynamic temporal behaviour for a time sequence. Unlike feed-forward neural networks, RNNs can use their internal state (memory) to process sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition.

ALGORITHM

Step 1: A single time step of the input is provided to the network.

Step 2: Then calculate its current state using set of current input and the previous state.

Step 3: The current ht becomes ht-1 for the next time step.

Step 4: One can go as many time steps according to the problem and join the information from all the previous states.

Step 5: Once all the time steps are completed the final current state is used to calculate the output.

Step 6: The output is then compared to the actual output i.e the target output and the error is generated.

Step 7: The error is then back-propagated to the network to update the weights and hence the network (RNN) is trained.

CODE

```
[X,T] = simpleseries_dataset;
net = layrecnet(1:2,10);
[Xs,Xi,Ai,Ts] = preparets(net,X,T);
net = train(net,Xs,Ts,Xi,Ai);
view(net)
Y = net(Xs,Xi,Ai);
perf = perform(net,Y,Ts)
```

3.4.3 SUPPORT VECTOR MACHINE (SVM)

Support vector machine (SVM) is a relatively new form of supervised machine learning. The Support Vector Machine is implemented using the kernel Adatron algorithm. The kernel Adatron maps inputs to a high-dimensional feature space, and then optimally separates data into their respective classes by isolating those inputs which fall close to the data boundaries. Therefore, the kernel Adatron is especially effective in separating sets of data which share complex boundaries. SVM can only use for classification, not for function approximation.

ALGORITHM

Step 1: Data Pre-processing

Step 2: Importing datasets

Step 3: Extracting Independent and dependent Variable

Step 4: Splitting the dataset into training and test set.

Step 5: Feature Scaling

Step 6: Predicting the test set result

Step 7: Creating the confusion matrix and AUC-ROC curve

CODE

```
function [trainedClassifier, validationAccuracy] = trainClassifier(datasetTable)
predictorNames = {'Age', 'Sex', 'Chestpaincp', 'BloodpressuremmHg', 'Cholesterolmgdl', 'Bloodsugarmgdl', 'ElectocardiogramECG', 'Heartrate', 'Angina', 'Oldpeak', 'Slope', 'Vessels', 'Thalassemia'};
predictors = datasetTable(:,predictorNames);
predictors = table2array(varfun(@double, predictors));
response = datasetTable.Target;
trainedClassifier = fitcsvm(predictors, response, 'KernelFunction', 'linear', 'PolynomialOrder', [], 'KernelScale', 'auto', 'BoxConstraint', 1, 'Standardize', 1, 'PredictorNames', {'Age' 'Sex' 'Chestpaincp' 'BloodpressuremmHg' 'Cholesterolmgdl' 'Bloodsugarmgdl' 'ElectocardiogramECG' 'Heartrate' 'Angina' 'Oldpeak' 'Slope' 'Vessels' 'Thalassemia'}, 'ResponseName', 'Target', 'ClassNames', [0 1]);
```

```
partitionedModel = crossval(trainedClassifier, 'KFold', 10);
validationAccuracy = 1 - kfoldLoss(partitionedModel, 'LossFun', 'ClassifError');
```

3.4.4 ENSEMBLE CLASSIFIER

In artificial intelligence, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Unlike a statistical ensemble in statistical mechanics, which is usually infinite, a machine-learning ensemble consists of only a concrete finite set of alternative models, but typically allows for much more flexible structure to exist among those alternatives. The main challenge is not to obtain highly accurate base models, but rather to obtain base models, which make different kinds of errors. For example, if ensembles are used for classification, high accuracies can be accomplished if different base models misclassify different training examples, even if the base classifier accuracy is low.

ALGORITHM

Step 1: Input, Training data D

Step 2: Output, Enseble classifier H

Step 3: Learn base-level classifiers

Step 4: for t=1 to T do

Learn H_t base on D

Step 5: end for

Step 6: Construct new data for predictions

Step 7: Return H

CODE

```
function [trainedClassifier, validationAccuracy] = trainClassifier(datasetTable)
predictorNames = {'Age', 'Sex', 'Chestpaincp', 'BloodpressuremmHg', 'Cholesterolmgdl', 'Bloodsugarmgdl', 'ElectocardiogramECG', 'Heartrate', 'Angina', 'Oldpeak', 'Slope', 'Vessels', 'Thalassemia'};
predictors = datasetTable(:,predictorNames);
predictors = table2array(varfun(@double, predictors));
```

```
response = datasetTable.Target;
trainedClassifier = fitctree(predictors, response, 'PredictorNames', {'Age' 'Sex' 'Chestpaincp' 'BloodpressuremmHg' 'Cholesterolmgdl' 'Bloodsugarmgdl' 'ElectocardiogramECG' 'Heartrate' 'Angina' 'Oldpeak' 'Slope' 'Vessels' 'Thalassemia'}, 'ResponseName', 'Target', 'ClassNames', [0 1], 'SplitCriterion', 'gdi', 'MaxNumSplits', 4, 'Surrogate', 'off');
partitionedModel = crossval(trainedClassifier, 'KFold', 10);
validationAccuracy = 1 - kfoldLoss(partitionedModel, 'LossFun', 'ClassifError');
```

3.4.5 NEAREST NEIGHBOUR CLASSIFIERS

The k-nearest neighbours (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. K-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, normalizing the training data can improve its accuracy dramatically.

ALGORITHM

Step 1: Select the number K of the neighbors.

Step 2: Calculate the Euclidean distance of K number of neighbors.

Step 3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step 4: Among these k neighbors, count the number of the data points in each category.

CODE

```
function [trainedClassifier, validationAccuracy] = trainClassifier(datasetTable)
predictorNames = {'Age', 'Sex', 'Chestpaincp', 'BloodpressuremmHg', 'Cholesterolmgdl', 'Bloodsugarmgdl', 'ElectocardiogramECG', 'Heartrate', 'Angina', 'Oldpeak', 'Slope', 'Vessels', 'Thalassemia'};
predictors = datasetTable(:,predictorNames);
predictors = table2array(varfun(@double, predictors));
response = datasetTable.Target;
trainedClassifier = fitcknn(predictors, response, 'PredictorNames', {'Age' 'Sex' 'Chestpaincp' 'BloodpressuremmHg' 'Cholesterolmgdl' 'Bloodsugarmgdl' 'ElectocardiogramECG' 'Heartrate' 'Angina' 'Oldpeak' 'Slope' 'Vessels' 'Thalassemia'}, 'ResponseName', 'Target', 'ClassNames', [0
```

```
1], 'Distance', 'Euclidean', 'Exponent', '', 'NumNeighbors', 1, 'DistanceWeight', 'Equal', 'StandardizeData', 1);
partitionedModel = crossval(trainedClassifier, 'KFold', 10);
validationAccuracy = 1 - kfoldLoss(partitionedModel, 'LossFun', 'ClassifError');
```

3.4.6 DECISION TREE

Decision tree learning is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

ALGORITHM

Step 1: Begin the tree with the root node, says S, which contains the complete dataset.

Step 2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).

Step 3: Divide the S into subsets that contains possible values for the best attributes.

Step 4: Generate the decision tree node, which contains the best attribute.

Step 5: Recursively make new decision trees using the subsets of the dataset created in step 3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

CODE

```
function [trainedClassifier, validationAccuracy] = trainClassifier(datasetTable)
predictorNames = {'Age', 'Sex', 'Chestpaincp', 'BloodpressuremmHg', 'Cholesterolmgdl', 'Bloodsugarmgdl', 'ElectocardiogramECG', 'Heartrate', 'Angina', 'Oldpeak', 'Slope', 'Vessels', 'Thalassemia'};
predictors = datasetTable(:,predictorNames);
predictors = table2array(varfun(@double, predictors));
response = datasetTable.Target;
```

```
trainedClassifier = fitcsvm(predictors, response, 'KernelFunction', 'linear', 'PolynomialOrder', [], 'KernelScale', 'auto', 'BoxConstraint', 1, 'Standardize', 1, 'PredictorNames', {'Age' 'Sex' 'Chestpaincp' 'BloodpressuremmHg' 'Cholesterolmgdl' 'Bloodsugarmgdl' 'ElectocardiogramECG' 'Heartrate' 'Angina' 'Oldpeak' 'Slope' 'Vessels' 'Thalassemia'}, 'ResponseName', 'Target', 'ClassNames', [0 1]);
partitionedModel = crossval(trainedClassifier, 'KFold', 10);
validationAccuracy = 1 - kfoldLoss(partitionedModel, 'LossFun', 'ClassifError');
```

3.5 PERFORMANCE METRICS

Applying different artificial neural network structures for any specific study this study have got different results. But it had to find the best network structure considering different parameters like Mean Squared Error (MSE), Minimum Absolute Error (MAE), Maximum Absolute Error and Linear Correlation Coefficient (r).

3.5.1 MEAN SQUARE ERROR (MSE)

In neural network, the mean squared error (MSE) measures the average of the squares of the errors that is, the average squared difference between the estimated values and what is estimated. The MSE will be obtained by Eq. (1) respectively.

$$\text{MSE} = 1/\text{N} \sum_{o=1}^{N}(t_T - y_o)^2 \quad (1)$$

Where, N is the total number of training pattern.

3.5.2 MEAN ABSOLUTE ERROR (MAE)

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. The MAE can be obtained by Eq. (3), respectively,

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_i - \dot{y}_j| \quad (2)$$

3.5.3 MIN ABSOLUTE ERROR AND MAX ABSOLUTE ERROR

Min Abs Error and Max Abs Error can be calculated by the following equations respectively,

$$\text{Min Abs Error} = \min |y_i - t_j| \quad (3)$$

Max Abs Error = max $|y_i - t_j|$ (4)

Where, y_i is desired output, t_j is computed output.

3.5.4 NORMALIZED MEAN SQUARE ERROR (NMSE)

The NMSE (Normalized Mean Square Error) is an estimator of the overall deviations between predicted and measured values. It is defined as:

NMSE = RMSE / $(y_{max} - y_{min})$ (5)

NMSE generally shows the most striking differences among models. If a model has a very low NMSE, then it is well performing both in space and time. On the other hand, high NMSE values do not necessarily mean that a model is completely wrong.

3.5.5 LINEAR CORRELATION COEFFICIENT, r

R-value, the coefficient of correlation or determination are used to measure the correlation between actual and predicted value. It measures the direction and strength of the linear relationship between actual and predicted value. The r value is always between -1and +1.The highest value of r is 1. The r value is a measure of how well the variation in the output is explained by the target. There is perfect correlation between target and output when this value is equal to 1. A positive r indicates a positive relationship and a negative r indicates a negative relationship.

3.5.6 CONFUSION MATRIX

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix, a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa). The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabelling one as another). It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions (each combination of dimension and class is a variable in the contingency table).

3.5.7 ROC CURVE

In Machine Learning, performance measurement is an essential task. So when it comes to a classification problem, it can count on an AUC - ROC Curve. When it needs to check or visualize the performance of the multi - class classification problem, it uses AUC (Area under the Curve) ROC (Receiver Operating Characteristics) curve. It is one of the most important evaluation metrics for checking any classification model’s performance. It is also written as AUROC (Area under the Receiver Operating Characteristics).

3.6 TRAINING, TESTING, VALIDATION

Once a network has been structured for a particular study, that network is ready to be trained. To start this process the initial weights are chosen randomly. There are two approaches to training - supervised and unsupervised. Supervised training involves a mechanism of providing the network with the desired output either by manually "grading" the network's performance or by providing the desired outputs with the inputs. Unsupervised training is where the network has to make sense of the inputs without outside help. Unsupervised training is used to perform some initial characterization on inputs. In supervised training, both the inputs and the outputs are provided. The network then processes the inputs and compares its resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the weights which control the network. The set of data which enables the training is called the "training set." During the training of a network the same set of data is processed many times as the connection weights are ever refined. The other type of training is called unsupervised training. In unsupervised training, the network is provided with inputs but not with desired outputs. The system itself must then decide what features it will use to group the input data. This is often referred to as self-organization or adaption. The study of detecting heart disease will use 60% of total dataset for training approach.

Testing of the data is used to confirm the expected result, i.e. when test data is entered the expected result should come and some test data is used to verify the software behaviour to invalid input data. Test data is generated by testers or by automation tools which support testing. Data validation is intended to provide certain well-defined guarantees for fitness, accuracy, and consistency for any of various kinds of user input into an application or automated system. This study will use 30% of total dataset for the testing purpose.

Data validation is the process of ensuring data have undergone data cleansing to ensure they have data quality, that is, that they are both correct and useful. It uses routines, often called "validation rules" "validation constraints" or "check routines", that check for correctness, meaningfulness, and security of data that are input to the system. The rules may be implemented through the automated facilities of a data dictionary or by the inclusion of explicit application program validation logic. This study will use 10% data from the dataset.

CHAPTER 4

RESULT AND DISCUSSION

4.1 INTRODUCTION

Prediction of occurrences of heart diseases in medical field is significant and complex job. Data analytics is useful for prediction from more information and it helps medical centre to predict of various disease. This research analyses the data of various patients with different parameters. The stored data can be useful for source of predicting the occurrence of future disease. Some of the artificial neural network techniques were used to predict the heart disease, such as Radial Basis Function Neural Network (RBFNN), Fuzzy Neural Network and Support Vector Machine (SVM) etc. This research provides an insight of the existing algorithm and it will give an overall summary of the existing works. The accuracy level is different from structure to structure. This research presents a novel approach to predict heart disease. It presents a comparative study conducted on some heart patient and non-patient based on artificial neural network (ANN) models like Radial Basis Function, Self-Organizing Feature Map, Recurrent Network, and Support Vector Machine and Fuzzy logic. Among them the Radial Basis Function Neural network (RBFNN) shows the most promising result. After that this research conducted a comparison with the result got from support vector machine running by two software. The software that shows more accuracy is adopted.

4.2 RADIAL BASIS FUNCTION

The radial basis function neural network configuration for this research shows the more accurate result than the other network configuration. Figures and tables are showing the training, testing and validation results below.

4.2.1 TRAINING RESULT

The training result of radial basis function neural network structure for this research is shown for the epoch number of 30000. In the following table the computed minimum mean square (MSE) error and final MSE for the best network configuration structure is given.

Table 4.1 Training result of the best network of RBFNN

Best Network	Training
Epoch	29990
Minimum MSE	0.034358199
Final MSE	0.033498587

In the following fig 4.1 training mean square error vs. epoch is shown for the specific epoch. The solid line is showing the training MSE with respect to the epoch number.

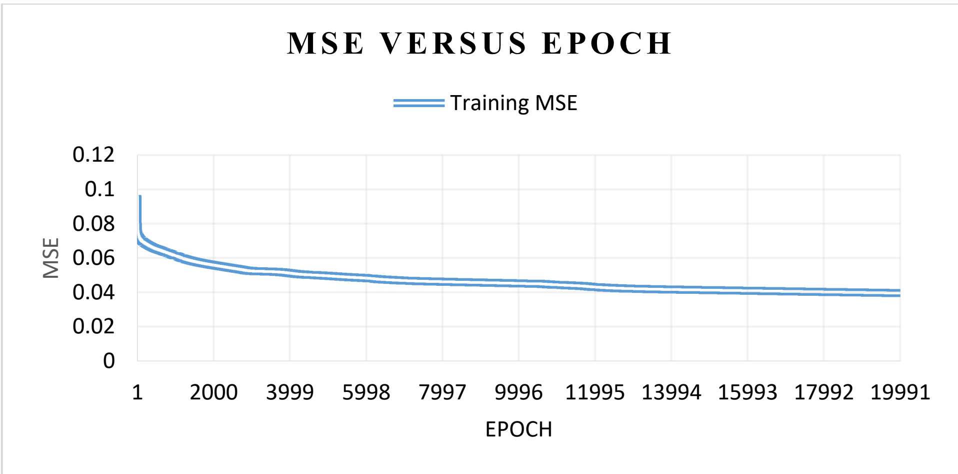

Figure 4.1 Training result of radial basis function

In the fig training epoch 20000 is used. The solid line is showing the training MSE with respect to the epoch number.

4.2.2 TESTING RESULT

The testing result of the Radial Basis Function (RBFNN) is shown by the figure given below. The desired output and actual radial basis function neural network output is shown concurrently. Table of testing performance parameters like mean square error (MSE), normalized mean square error (NMSE), mean average error, min absolute error, max absolute error, linear correlation coefficient (r) and percent of correct output values and their calculated outcome is given below. Output is computed in case of outcome 0 and outcome 1. Total no. of correctly identified instances are 238 and incorrectly identified instances are 50.

Table 4.2 Performance matrices of radial basis function

Output / Desired	Outcome(0)	Outcome(1)
Outcome(0)	92	35
Outcome(1)	15	146

Table 4.3 shows the performance criteria and their outcome. It shows the MSE, NMSE, MAE, r value.

Table 4.3 Testing result with performance and output of RBFNN

Performance	Outcome
MSE	0.148367498
NMSE	0.782761984
MAE	0.31359339
Min Abs Error	0.004629013
Max Abs Error	1.050074899
R	0.51565647

The value r which is close to 1 is a good value for the accuracy of the neural network.

4.2.3 VALIDATION RESULT

Validation result of this study by using Radial Basis Function neural network (RBFNN) has achieved 82.5% accuracy. This figure shows the graphical view in basis of the comparison between the desired output and Radial Basis Function neural network (RBFNN) output.

4.3 RECURRENT NETWORK

The network structure by using the recurrent network for this study uses two hidden layer with six processing element and 60000 epoch. The train, test and validation result is given by the following tables and figures. 4.2.1 Training result The following table shows the best network configuration of training result with the epoch number, minimum MSE, final MSE.

4.3.1 TRAINING RESULT

The following table shows the best network configuration of training result with the epoch number, minimum MSE, final MSE. Table 4.2 Training result of recurrent neural network.

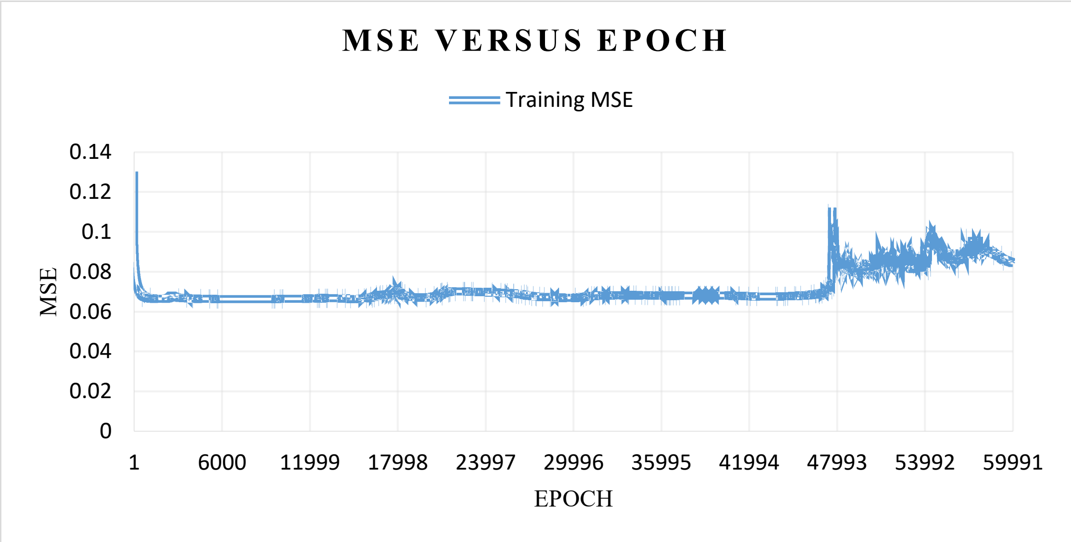

Figure 4.2 Training recurrent network

In the fig training epoch 60000 is used. The solid line is showing the training MSE with respect to the epoch number.

4.3.2 TESTING RESULT

This table shows the numerical values of testing performance parameters like MSE, NMSE, MAE, Min Absolute Error, Max Absolute Error and linear correlation coefficient (r) value and their outcome's. Output is computed in case of outcome 0 and outcome 1. Total no. of correctly identified instances are 229 and incorrectly identified instances are 59.

Table 4.4 Testing result of recurrent neural network.

Output / Desired	**Outcome(0)**	**Outcome(1)**
Outcome(0)	96	31
Outcome(1)	28	133

Table 4.5 Testing result of recurrent neural network

Performance	Outcome
MSE	0.15074796
NMSE	0.691308831
MAE	0.310484452
Min Abs Error	0.000834192
Max Abs Error	0.984280028
R	0.467301842

The value r which is close to 1 is a good value for the accuracy of the neural network.

4.2.3 VALIDATION RESULT

Validation result of this study by using recurrent neural network has achieved 79.5% accuracy on detecting heart disease. This figure shows the graphical view in basis of the comparison between the desired output and recurrent neural network output.

4.4 SUPPORT VECTOR MACHINE (MEDIUM GAUSSIAN SVM)

For applying the SVM classifier on dataset, have to go to the App part of the software. After A 10-fold cross validation the resultant accuracy obtained was 84.4%. The time taken to build cross-validation was 26.09 seconds.

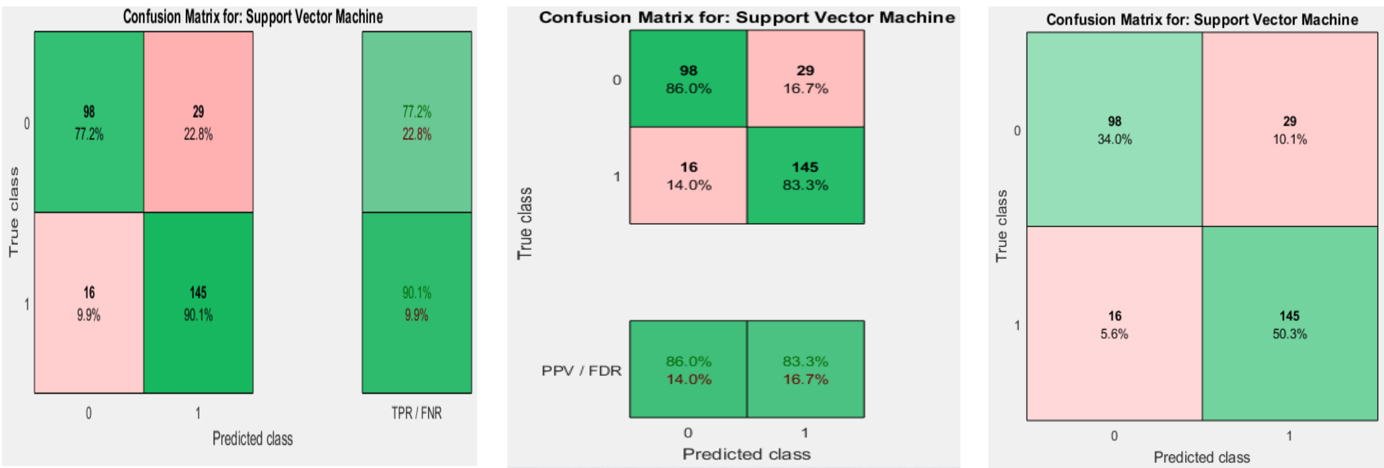

Figure 4.3 Confusion matrices of SVM classifier

The above figure stands for confusion matrix by SVM Classifier where the green cell is shown where the predicted result is matched with the true class. And the red cell is shown where the predicted result is not matched with the true class. The confusion matrix generated from the 10-fold cross validation by using SVM classifier is shown in Table.

Table 4.6 Accuracy table of SVM classifier

Total no. of Instances	**No. of Instances**	**Correctly predicted instances**	**Incorrectly classified instances**	**Percentage of error**	**Percentage of Accuracy**
Class ‘0’	127	98	29	22.8%	84.4%
Class ‘1’	161	145	16	9.9%	15.6%
Total	288	243	45	32.7%	100%

After performing 10-fold Cross-Validations Confusion matrix is given. In the above table, correctly classified instances are 243 and incorrectly classified instances are 47, which leads 84.4% and 15.6% respectively. From the above table get the values of confusion matrix for (SVM) where class 0 means negative value and class 1 means positive value. Left side column is for the actual value and the first row is for predicted value. Following curve shown in figure is for ROC curve for measuring performance. The figure given below is showing the ROC curve of SVM classifier.

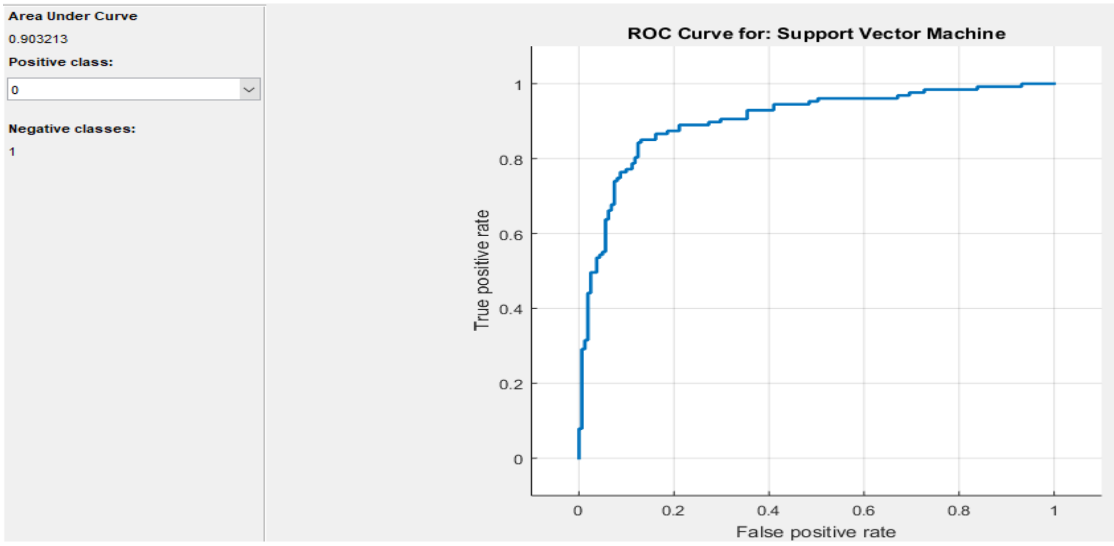

Figure 4.4 ROC curve for measuring performance of SVM

Fig shows Receiver Operating Characteristic (ROC) Curve of SVM classifier. The Y axis is stand for true positive value and X axis is for false positive value. Area under ROC (AUC) for both Class 0 and Class 1 is 0.90. The closer point of the true positive rate is (0.36, 0.88). build.

4.5 ENSEMBLE CLASSIFIER (BOOSTED TREE)

For applying the Ensemble classifier on dataset, have to go to the App part of the software. After A 10-fold cross validation the resultant accuracy obtained was 84.4%. The time taken to build cross-validation was 26.09 seconds.

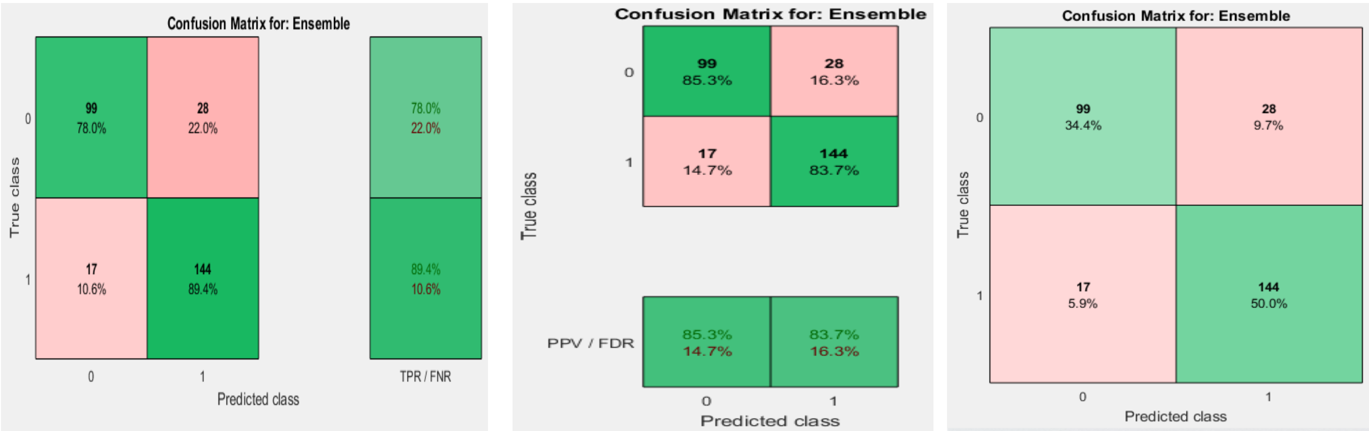

Figure 4.5 Confusion matrices of Ensemble classifier (Boosted tree)

The above figure is stand for confusion matrix by Ensemble Classifier where the green cell is shown where the predicted result is matched with the true class and the red cell is shown where the predicted result is not matched with the true class. The confusion matrix generated from the 10-fold cross validation by using ensemble classifier is shown in Table 4.8.

Table 4.8 Accuracy table of Ensemble classifier (Boosted tree)

Total no. of Instances	**No. of Instances**	**Correctly predicted instances**	**Incorrectly classified instances**	**Percentage of error**	**Percentage of Accuracy**
Class ‘0’	127	99	28	22.6%	84.4%
Class ‘1’	161	144	17	10%	15.6%
Total	290	243	45	32.6%	100%

After performing 10-fold crossValidations Table 4.8 of Confusion matrix is given. In the above table, correctly classified instances are 243 and incorrectly classified instances are 47, which leads 84.4% and 15.6% respectively. From the above table get the values of confusion matrix for Boosted tree where class 0 means negative value and class 1 means positive value. Left side column is for the actual value and the first row is for predicted value. Following curve shown in figure is for ROC curve for measuring performance.

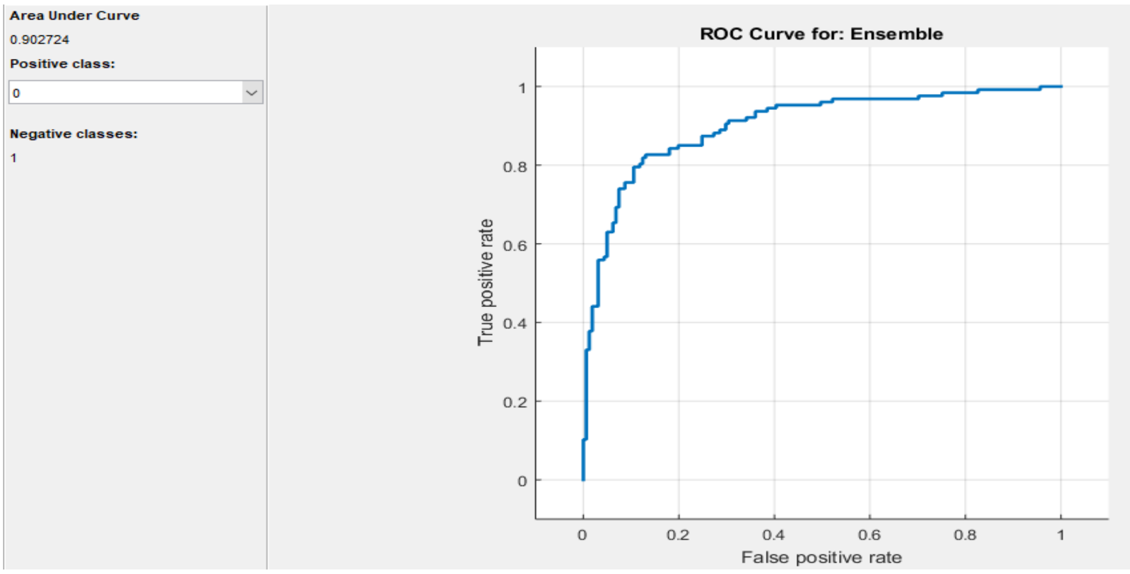

Figure 4.6 ROC curve for measuring performance of Boosted tree

Fig shows Receiver Operating Characteristic (ROC) Curve of boosted tree classifier. The Y-axis is stand for true positive value and X-axis is for false positive value. Area under ROC (AUC) for both Class 0 and Class 1 is 0.90. The closer point of the true positive rate is (0.36, 0.88).

4.6 DECISION TREE (SIMPLE TREE)

For applying the Decision tree classifier on dataset, have to go to the App part of the software. After A 10-fold cross validation the resultant accuracy obtained was 83.0%. The time taken to build cross-validation was 25.09 seconds.

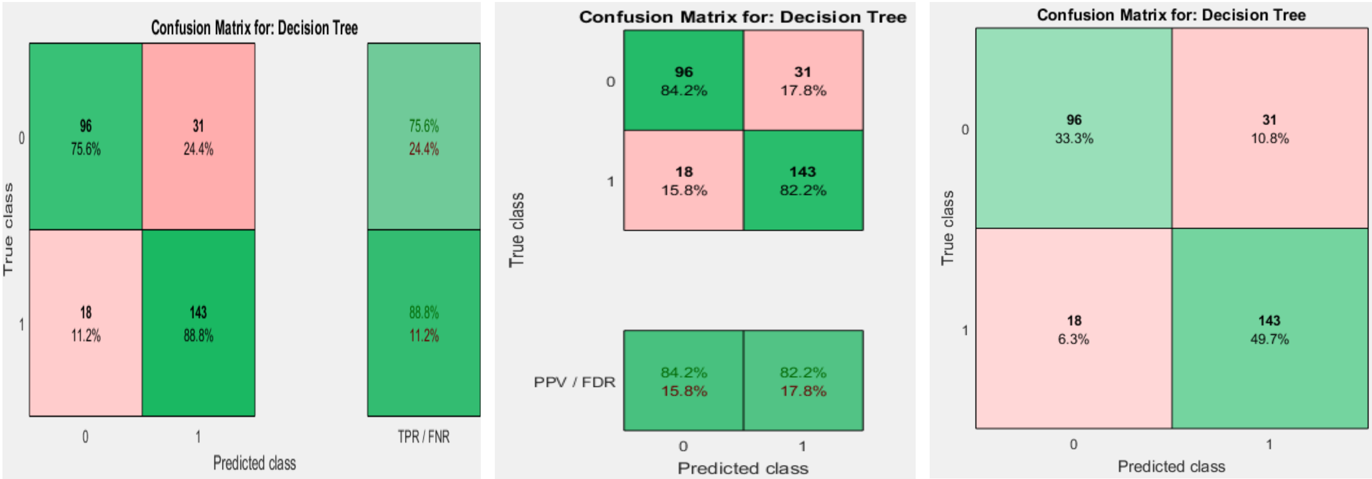

Figure 4.8 Confusion matrices of Decision tree

The above figure is stand for confusion matrix by Decision tree Classifier where the green cell is shown where the predicted result is matched with the true class and the red cell is shown where the predicted result is not matched with the true class. The confusion matrix generated from the 10-fold cross validation by using ensemble classifier is shown in Table 4.8.

Table 4.10 Accuracy table of Decision tree classifier (Simple tree)

Total no. of Instances	**No. of Instances**	**Correctly predicted instances**	**Incorrectly classified instances**	**Percentage of error**	**Percentage of Accuracy**
Class '0'	127	96	31	22.6%	83.0%
Class '1'	161	143	18	10%	17.0%
Total	290	239	51	32.6%	100%

After performing 10-fold Cross-Validations Table 4.10 of Confusion matrix is given. In the above table, correctly classified instances are 239 and incorrectly classified instances are 51, which leads 83.0% and 17% respectively. From the above table get the values of confusion matrix for Boosted tree where class 0 means negative value and class 1 means positive value. Left side column is for the actual value and the first row is for predicted value. Following curve shown in figure is for ROC curve for measuring performance.

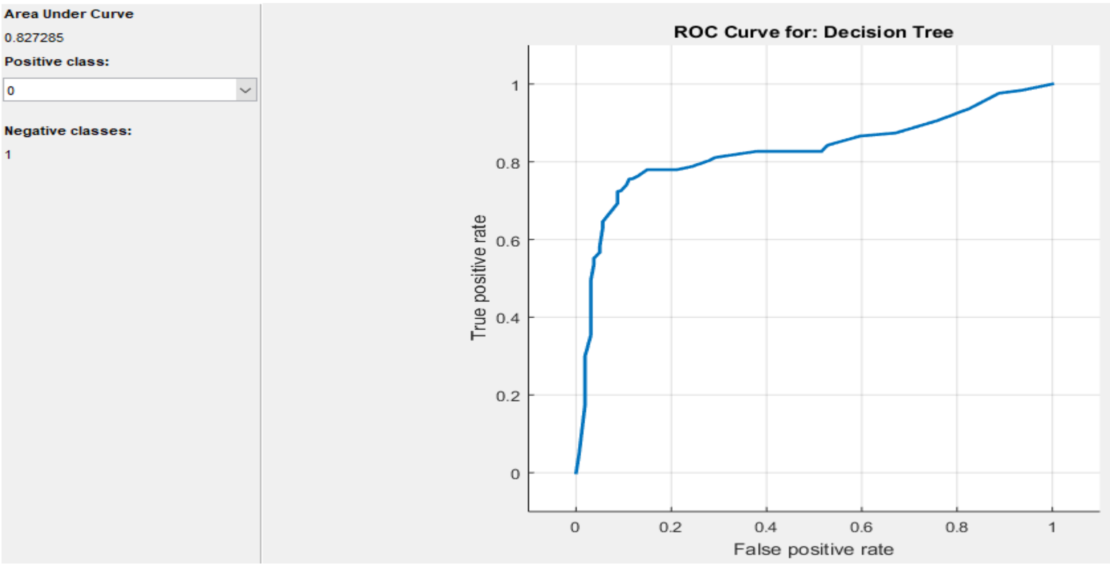

Figure 4.9 ROC curve for measuring performance of decision tree

Fig shows Receiver Operating Characteristic (ROC) Curve of SVM classifier. The Y-axis is stand for true positive value and X-axis is for false positive value. Area under ROC (AUC) for both Class 0 and Class 1 is 0.827. The closer point of the true positive rate is (0.36, 0.88).

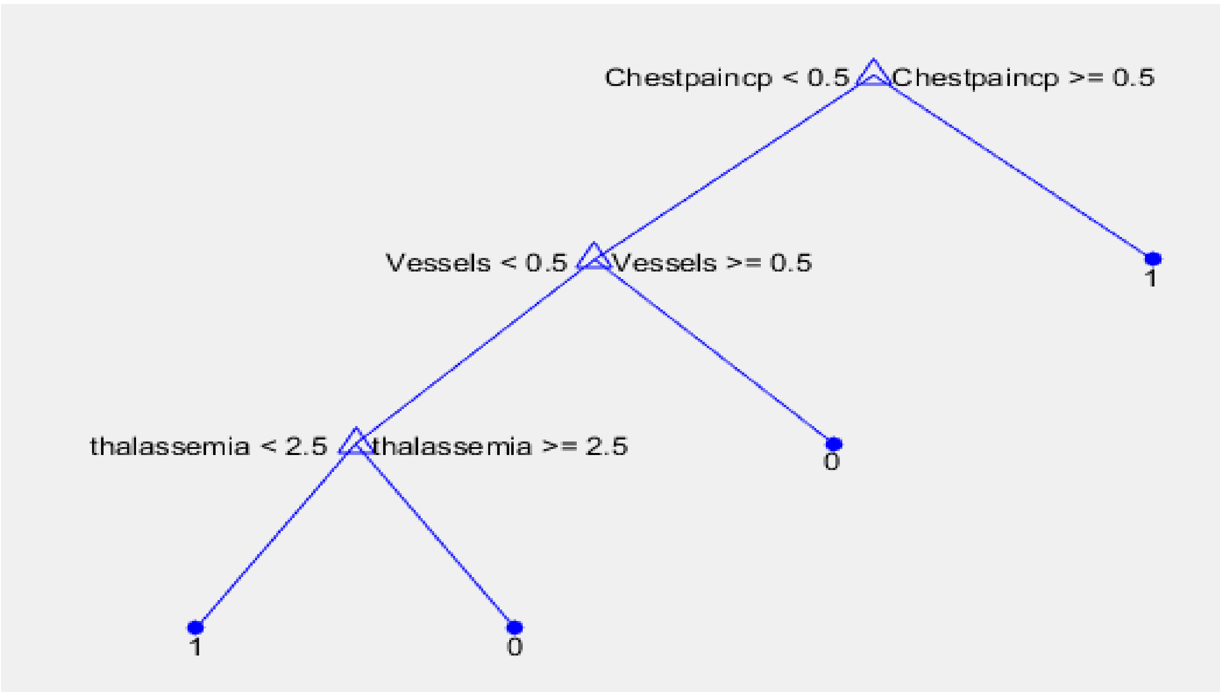

Figure 4.10 Graphical view of the Simple Tree model

For generating the tree view of the Simple Tree, the whole code and the compact model both have to be exported in the work space. The function to generate the graph of the Tree is written. In the command line, for graphical view have to write the line that is view (M, 'Mode', 'Graph') where the M stands for the model name which is exported after model build.

4.7 K-NEAREST NEIGHBOUR CLASSIFIER (WEIGHTED KNN)

For applying the KNN classifier on dataset, have to go to the App part of the software. After A 10-fold cross validation the resultant accuracy obtained was 83.7%. The time taken to build cross-validation was 29.09 seconds.

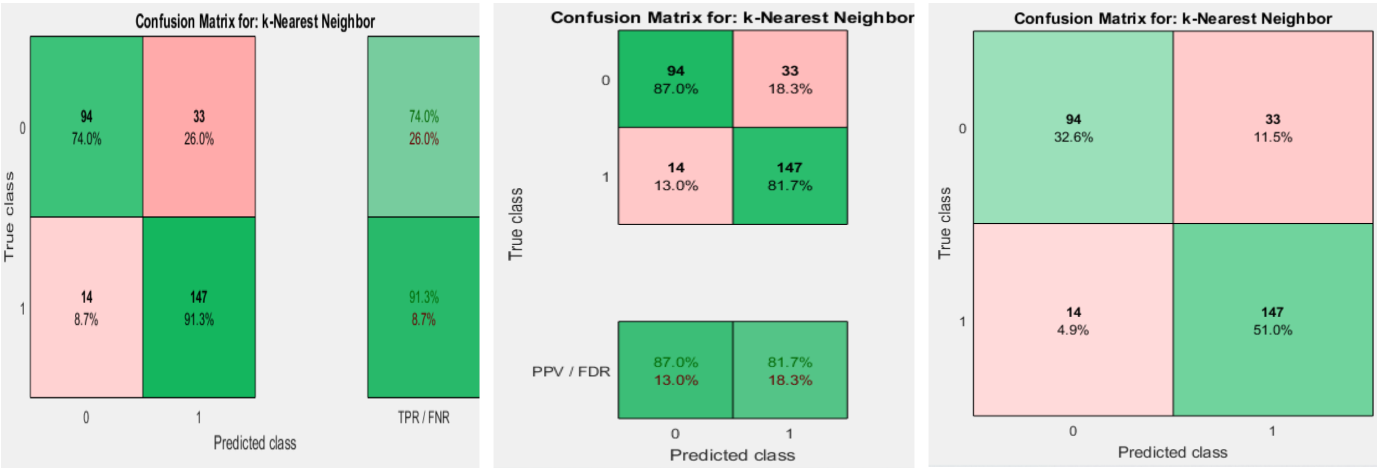

Figure 4.11 Confusion matrix from KNN classifier

The above figure is stand for confusion matrix by Ensemble Classifier where the green cell is shown where the predicted result is matched with the true class and the red cell is shown where the predicted result is not matched with the true class. The confusion matrix generated from the 10-fold cross validation by using ensemble classifier is shown in Table 4.12.

Table 4.12 Accuracy result of KNN (Weighted KNN)

Total no. of Instances	No. of Instances	Correctly predicted instances	Incorrectly classified instances	Percentage of error	Percentage of Accuracy
Class '0'	127	94	33	26%	83.7%
Class '1'	161	147	14	8.7%	16.3%
Total	290	241	47	34.7%	100%

In the above table, correctly classified instances are 241 and incorrectly classified instances are 49, which leads 83.7% and 16.3% respectively. From the above table get the values of confusion matrix for Weighted KNN where class 0 means negative value and class 1 means positive value. Left side column is for the actual value and the first row is for predicted value. Following curve shown in figure is for ROC curve for measuring performance.

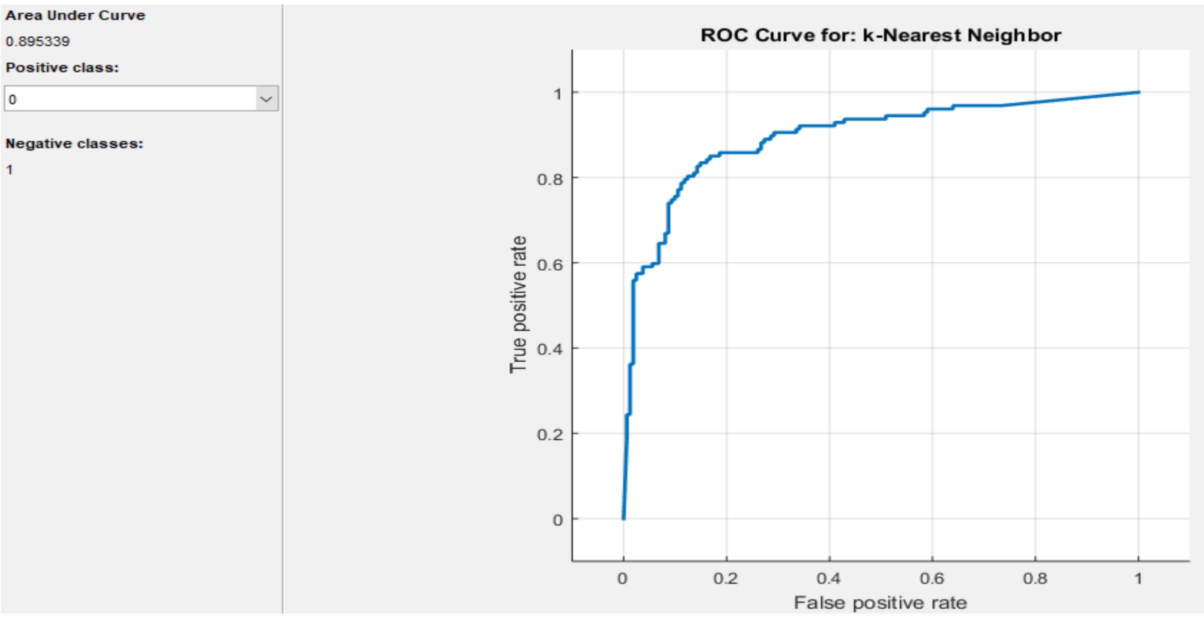

Figure 4.12 ROC curve for measuring performance of Weighted KNN

Fig shows Receiver Operating Characteristic (ROC) Curve of boosted tree classifier. The Y-axis is stand for true positive value and X-axis is for false positive value. Area under ROC (AUC) for both Class 0 and Class 1 is 0.89. The closer point of the true positive rate is (0.36, 0.88).

4.8 PREDICTION PROCESS

After whole model development and performance analysis with different classification algorithm, the values of different models are investigated with the performance on the classification methods. Comparison is made among these classification algorithms and finally linear SVM and boosted tree classifier showed the better performance and they provides the accuracy of 84.4%. Now this models are used for prediction process. The steps of the prediction process are given bellow.

I. Test the dataset with new 10 instances.

II. For prediction process, after training process export the selected whole model from software APP in the working space.

Table 4.13 Sample dataset for prediction process

Age	Sex	Chest pain	Blood pressure	Cholesterol	Blood sugar	Electrocardiogram	Heart rate	Angina	Old peak	Slope	Vessels	Thalassemia	Target
51	1	2	94	227	0	1	154	1	0	2	1	3	1
39	0	2	95	199	0	1	179	0	0	2	0	2	1
51	1	2	100	222	0	1	143	1	1.2	1	0	2	1
58	0	0	100	248	0	0	122	0	1	1	0	2	1
67	1	0	100	299	0	0	125	1	0.9	1	2	2	0
55	1	0	101	234	0	1	156	0	0.1	2	1	3	0
46	1	1	101	197	1	1	156	0	0	2	0	3	1
60	0	2	102	318	0	1	160	0	0	2	1	2	1
42	0	0	102	265	0	0	122	0	0.6	1	0	2	1
45	1	0	104	208	0	0	148	1	3	1	1	2	1

III. For prediction process, after training process export the selected whole model from software APP in the working space.

IV. Then import the new sample dataset which is also normalized. In the dataset, all the attribute field will be as same as the previous full dataset for training purpose. Just the values of the target class are not being included.

V. In the working window, have to write a specific function for all the trained model which is exported and it is
‘X = table2array(varfun(@double, T(:,TrainedClassifier2.PredictorNames)));
yfit = predict(trainedClassifier2, X)
Here trainedmodel is the compact model name and T is the name of the test dataset.

VI. Run the test dataset in MATLAB. Then apply different classifier algorithms for testing purpose. As a result, different algorithm shows different predictive result. As for example, in figure includes output of algorithm which is Bagged Trees.

4.9 PREDICTION RESULTS

One of the main task of the research is to show the prediction with new samples. As the main goal of the research is to find out the best algorithm which will give better accuracy in early prediction of heart disease to minimize the death rate. In this purpose, after follow the 4.3 section process, get the output of the prediction. In the following figure shows the prediction results of the all of the experimental models.

Table 4.14 Output comparison for prediction process

Target Output	Predicted Output			
	Linear SVM	**Boosted Tree**	**Medium K-NN**	**Decision Tree**
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
0	0	0	0	0
0	1	0	1	0
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	0	0	0	0

The table given above states two part of targeted output and predicted output. In the predicted output part there are four parts as Linear SVM, Boosted tree, Medium K-NN, Decision tree.

4.10 COMPARISON BETWEEN THE NEURAL NETWORK STRUCTURES

The different neural network structures varies with their varying epoch number, processing elements, hidden layer, transfer function, activation function. The result for a specific study varies by altering them and they build different network structures that shows different results. Some neural network structures shows more accurate result than the others. In every research, the best network is chosen considering the parameters like mean square error

(MSE), standard deviation, linear correlation coefficient (r) and many others. In this study there is developed such a comparison among the built neural network structure. There is a table showing all the attributes that differs one network from another in this study. Table 4.8 is showing the parameters and their values.

Table 4.14 Comparison among performance matrices of neuro solution

Neural network structure s	**Training**	**Testin g**	
	MSE	MAE	r
Recurrent network	0.067774138	0.130747968	0.467301842
Radial Basis Function	0.034358199	0.148367498	0.51565647

On the other hand a network structure is considered to be better than other when its r value is near to 1. From the table of comparison it can be seen that the radial basis function neural network shows the lowest MSE of 0.034358199 in training but in testing its increases to 0.148367498. On the other hand the recurrent neural network structure shows the lowest MSE in testing that is 0.130747968. From the validation average percentage of accuracy is found out. Accuracy of a neural network finally decised whice network structure will be more reliable for this research. Recurrent network shows 78.64%. The tabuler form of the comparison between them is given by the following table.

Table 4.15 Comparison of the neural network structures by validation results

Neural network structure	**Accuracy**
Recurrent network	78.64%
Radial Basis Function	82.5%

RBFNN shows 82.5% which is near to approximately 83%. From the previous research of detecting heart disease the rate of accuracy obtained by radial basis function neural network is only 81%. So the radial basis function neural network structure for detecting heart disease is much better than other research's. So this research of detecting heart disease selects the network structure of Radial Basis Function Neural Network (RBFNN) as it shows the highest accuracy.

The table given below shows the performance measurements and accuracy paercentage of different classifiers.

Table 4.16 Performance measurement of all classifiers

Evaluation Criteria	**Classifiers**					
	SVM	Boosted tree	Decision tree	KNN	RBF	Recurrent network
Timing to build model (in Sec)	55.07	15.188	12.42	20.27	30.3	42.1
Correctly classified instances	243	243	239	241	238	229
Incorrectly classified instances	47	47	51	49	50	59
Accuracy (%)	84.4%	84.4%	83.0%	83.7%	82.5%	78.64%

In the table above table shows the values in some evaluation criteria- time for building model, number of correctly classified instances, incorrectly classified instances and accuracy of the entire model. The following figure shows in bar chart form of accuracy of all the models.

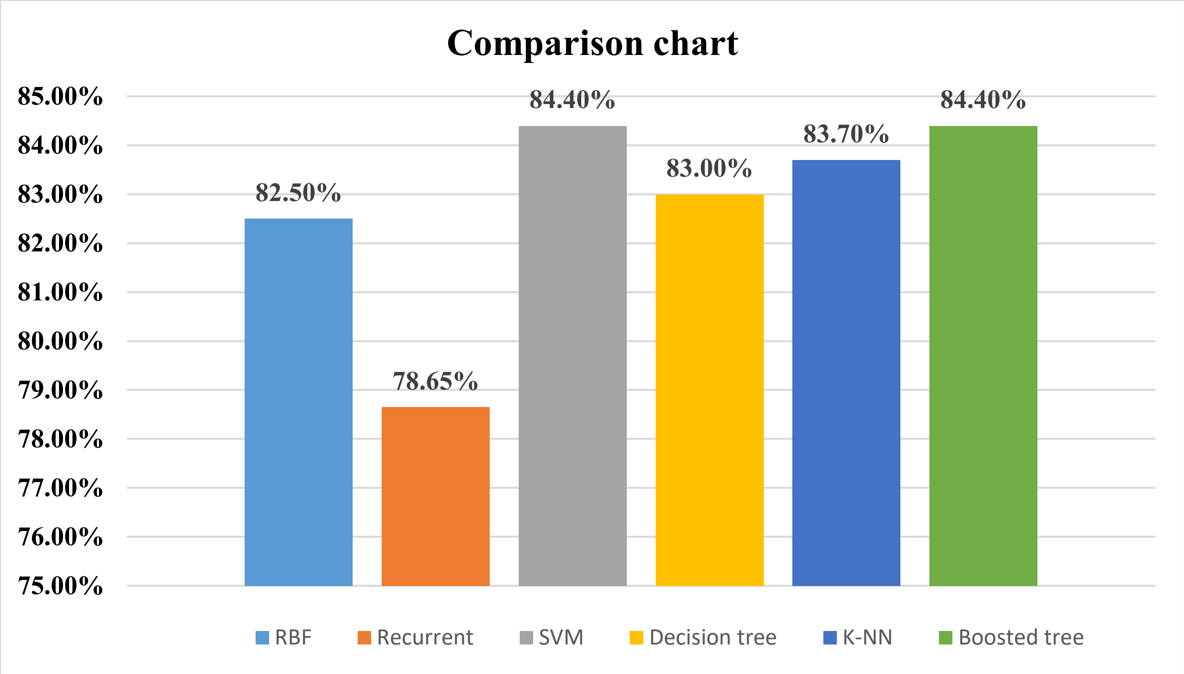

Figure 4.26 Comparison chart between different neural network structures

From the above figure, it is seen that Support vector machine and boosted tree shows the highest accuracy. The others two classifiers decision tree and K-NN shows 83% and 83.70% accuracy respectively. Although, all the time measurement of the model creation is varied from machine to machine.

4.11 COMPARISON WITH PREVIOUS RELATED WORK

A system for diagnosis of heart disease that was based on support vector machine along with Radial basis function algorithm was presented [27]. The result showed that the SVM was equivalently as good as compared to Radial basis function in detection of heart disease with accuracy of 86.42% using the indian dataset but this research found an acuuracy of about 84.4% from SVM and 82.5% from RBF using the real dataset. Haung and Wang proposed a system using Support Vector Machine where feature selection and optimization of Support Vector Machine parameter was done using genetic algorithm [30] with improved accuracy of 89.6% using online based dataset. Rajkumar and Sophia proposed the use of Data mining algorithm in diagnosis of heart disease with an accuracy of 52.33% [31]. The algorithms used by this system were Naive Bayes algorithm, Decision tree algorithm and KNN algorithm. Kumaravel et al., proposed automatic diagnosis system for heart diseases using neural network with an accuracy of 63.6 - 82.9% [32].

Table 4.16 Comparison with previous related work

Sl No.	Title	Author	Year	Methods Techniques	Accuracy (%)	Accuracy of this work (%)
1	Diagnosis of heart disease using neural network approach	P. P. H. A. T. Sayad	2014	Support Vector Machine, Radial Basis Function	86.42%	84.4% (SVM), 82.5% (RBF)
2	Diagnosis Of Heart Disease Using Datamining Algorithm	Rajkumar and Sophia	2009	K-Nearest Neighbor, Decision tree, Naive Bayes algorithm	52.33%	83.7% (K-NN), 83.0% (Decision tree)
3	Neural Network Based Intelligent System for Predicting Heart Disease	Kumaravel, Subhadra	2019	Neural network	63.6-82.9%	84.40% (SVM)
4	A GA-based feature selection and parameters optimization for support vector machines	Haung, Wang	2009	Support Vector Machine, Genetic algorithm	89.6%	84.4% (SVM)

From the comparison table it is clearly stated that those research used the dataset from the various sites but this research uses a dataset of real patients. So this research is more reliable comparing to those.

4.12 COMPARATIVE ANALYSIS AND FINAL SELECTION

The paper carried out experiments in order to evaluate the performance and usefulness of different classification algorithms for predicting heart disease. After performing different classification algorithms, value of different term for each classification algorithm are listed in the previous table 4.16 to measure and investigate the performance on the classification models.

CHAPTER 5

CONCLUSION AND FUTURE WORK

By using different artificial neural network, algorithms to predict the occurrence of heart disease will be summarised. Determining the prediction performance of each algorithm and applying the proposed system for the area it needed. This study will use more relevant feature selection methods to improve the accurate performance of algorithms. There are several treatment methods for patient, if they once diagnosed with the particular form of heart disease. Artificial neural network models can be of very knowledge form such suitable dataset. In conclusion, as identified through the literature survey, believe only a marginal success is achieved in the creation of predictive model for heart disease patients and hence there is a need for combinational and more complex models to increase the accuracy of the predicting the early onset of heart disease. With the more amount of data being fed into the database the system will be very intelligent.

There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. Due to time limitation, the following research / work need to be performed for the future. Would like to make use of testing different discretization techniques, multiple classifier voting technique and different decision tree types namely information gain and gain ratio. This study is willing to explore different optimizing techniques such as genetic algorithm and clustering algorithms.

REFERENCES

[1] "Cardiovascular Disease," World Health Organization (WHO), 2019.

[2] A. İ. Kaan Uyara, "Diagnosis of heart disease using genetic algorithm based trained," ScienceDirect, vol. 120, pp. 588-593, 2017.

[3] Zeinab Arabasadi et al., "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm," Computer Methods and Programs in Biomedicine, vol. 141, pp. 19-26, 2017.

[4] S. S. A.M Sagir, "Diagnosis of heart disease using genetic algorithm based trained," Procedia Computer Science, vol. 120, pp. 588-593, 2017.

[5] D. R. P. Jayshril S. Sonawane, "Prediction of heart disease using multilayer perceptron neural network," IEEE, vol. 112, pp. 412-420, 2014.

[6] Aditi Gavhane et al., "Prediction of Heart Disease Using Machine Learning," IEEE, vol. 100, pp. 1275-1278, 2018.

[7] O. K. Tulay Karayolan, "ANALYSIS OF HEART DISEASE PREDICTION SYSTEM USING ARTIFICIAL NEURAL," Journal of Emerging Technologies and Innovative Research (JETIR), vol. 5, no. 11, pp. 336-345, 2018.

[8] Sarath Babu et al., "Heart disease diagnosis using data mining technique," International conference of Electronics, Communication and Aerospace Technology (ICECA), vol. 5, no. 9, pp. 311-320, 2017.

[9] Jyoti Soni et al., "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," International Journal of Computer Applications, vol. 5, no. 7, pp. 43-48, 2011.

[10] A. P. K, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," Journal of King Saud University - Computer and Information Sciences , vol. 25(1), pp. 27-40, 2012.

[11] K. J. Nidhi Bhatla, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques," International Journal of Engineering Research & Technology (IJERT), vol. 1, no. 8, pp. 1-4, 2012.

[12] Aditya Methaila et al., "Early Heart Disease Prediction Using Data Mining Techniques," Conference: Fourth International Conference on Computational Science, Engineering and Information Technology, vol. 2, no. 4, pp. 23-30, 2014.

[13] Shimpy Goyal et al., "Heart disease diagnosis using data mining technique," IEEE, vol. 2, no. 5, pp. 30-38, 2011.

[14] B. D. Jabbar Akhil, "Heart Disease Prediction System using Associative Classification and Genetic Algorithm," International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, vol. 3, no. 4, pp. 5-10, 2012.

[15] R. Sethukkarasi et al., "An intelligent neuro fuzzy temporal knowledge representation model for mining temporal patterns," Journal of Intelligent and Fuzzy Systems (J INTELL FUZZY SYST), vol. 1, no. 1064-1246, pp. 322-330, 2001.

[16] M. A. Khaleel, "Association Rule Mining based on a Modified Apriori Algorithm in Heart Disease Prediction," International Journal of Scientific & Engineering Research Volume 9, Issue 9, September-2018 , vol. 9, no. 9, pp. 1025-1031, 2018.

[17] S. G. Boshra Bahrami, "A Literature Survey on Applications of Data Mining Techniques to Predict Heart Diseases," International Journal of Engineering Sciences Paradigms and Researches (IJESPR), vol. 20, no. 01, pp. 1-5, 2015.

[18] A. P. e. a. Deepika N, "A Literature Review on Heart Disease Prediction Based on Data Mining Algorithms," International Journal for Research Trends and Innovation , vol. 3, no. 6, pp. 100-104, 2018.

[19] ,. A. K. Sonam Nikhar, "Prediction of Heart Disease Using Machine Learning Algorithms," International Journal of Advanced Engineering, Management and Science (IJAEMS) , vol. 2, no. 6, pp. 617-621, 2016.

[20] E. Olaniyi, "Heart Diseases Diagnosis Using Neural Networks Arbitration," International Journal of Intelligent Systems Technologies and Applications , vol. 7, no. 12, pp. 75-82, 2015.

[21] M. A. Nishara Banu et al., "Disease Forecasting System Using Data Mining Methods," International Conference on Intelligent Computing Applications (ICICA), vol. 1, no. 4, pp. 5-10, 2014.

[22] K. S. T. John Peter, "An empirical study on prediction of heart disease using classification data mining techniques," International Conference on Intelligent Computing Applications, 2012.

[23] Samsher Bahadur Patel et al., "Role of Different Data Mining Techniques for Predicting Heart Disease," International Research Journal of Engineering and Technology (IRJET) , vol. 6, no. 9, pp. 2141-2145, 2019.

[24] D. C. A. U. K. Rajamohana Sp, "Analysis of Neural Networks Based Heart Disease Prediction System," International Conference on Human System Interaction (HSI), 2018.

[25] J. S. George, "Survey on Decision Support System For Heart Disease," International Journal of Advancements in Technology , vol. 4, no. 1, pp. 89-96, 2013.

[26] H. A.-A. Haitham Al-Angari, "Automated Recognition of Obstructive Sleep Apnea Syndrome Using Support Vector Machine Classifier," Transactions on information

technology in biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society , vol. 3, no. 8, 2012.

[27] P. P. H. A. T. SAYAD, "DIAGNOSIS OF HEART DISEASE USING NEURAL NETWORK APPROACH," International Journal of Advances in Science Engineering and Technology, vol. 2, no. 3, pp. 88-92, 2014.

[28] S. W. Du-Yih Tsai, "A method for optimization of fuzzy reasoning by genetic algorithms and its application to discrimination of myocardial heart disease," Transactions on Nuclear Science, 2000.

[29] N. C. S. N. I. Anbarasi Masilamani, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm," International Journal of Engineering Science and Technology , vol. 2, no. 10, 2010.

[30] C.-J. W. Cheng-Lung Huang, "A GA-based feature selection and parameters optimization for support vector machines," Expert Systems with Applications, vol. 2, no. 31, 2006.

[31] M. G. S. R. Asha Rajkumar, "Diagnosis Of Heart Disease Using Datamining Algorithm," IEEE, vol. 1, no. 9, 2009.

[32] V. B. K. Subhadra, "Neural Network Based Intelligent System for Predicting Heart Disease," International Journal of Innovative Technology and Exploring Engineering (IJITEE) , vol. 8, no. 5, pp. 484-487, 2019.

[33] V. B. Anchana Khemphila, "Heart Disease Classification Using Neural Network and Feature Selection," Systems Engineering (ICSEng), 2011.

[34] M. U. R. A. G. K. SHAHID MEHMOOD AWAN, "PREDICTION OF HEART DISEASE USING ARTIFICIAL NEURAL NETWORK," VFAST Transactions on Software Engineering, vol. 13, no. 3, pp. 102-112, 2018.

[35] D. K. U. Rani, "ANALYSIS OFHEARTDISEASESDATASET USING NEURALNETWORKAPPROACH," International Journal of Data Mining & Knowledge Management Process (IJDKP) , vol. 1, no. 5, pp. 1-8, 2011.

[36] B. D. t. P. C. M.Akhil Jabbar, "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection," Global Journal of Computer Science and Technology Neural & Artificial Intelligence , vol. 13, no. 3, pp. 4-14, 2013.

[37] M. D. O. Bhaskaru, "Accurate and Fast Diagnosis of Heart Disease using Hybrid Differential Neural Network Algorithm," International Journal of Engineering and Advanced Technology (IJEAT) , vol. 8, no. 3S, pp. 452-457, 2019.

[38] I. T. A. S. Resul Dasa, "Effective diagnosis of heart disease through neural networks ensembles," Expert Systems with Applications, vol. 36, no. 4, pp. 7676-7680, 2009.

[39] J. Engel, "Polytomous logistic regression," Statistica Neerlandica, vol. 42, no. 4, 1988.

[40] T. K. K. C. P. S. &. P. M. Gupta, " Dimensionality reduction techniques and its applications," Journal of Computer Science & Systems Biology, vol. 8, no. 3, 2015.

[41] K. R. &. K. E, "An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining," In Emerging Trends in Engineering, Technology and Science (ICETETS), pp. 1-5, 2016.

[42] A. a. G. D. M. Mosca, "Adapting resilient propagation for deep learning," arXiv preprint arXiv, vol. 1509, no. 4.

[43] A. M. &. K. Z, "Mobile Price Class prediction using Machine Learning Techniques".

[44] A. S. K. Y. S. &. M. T. Gandapur, "Study of risk factors in coronary heart disease," Pakistan Heart Journal, vol. 21, no. 4, 2012.

[45] C. A. K. U. e. a. SP Rajamhoana, "Analysis of neural networks based heart disease prediction system," IEEE, pp. 233-239, 2018.

[46] M. A. e. a. M. Marimuthu, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach," International Journal of Computer Applications , vol. 181, no. 18, pp. 20-25, 2018.

[47] R. A. Sellappan Palaniappan, "Intelligent Heart Disease Prediction System Using Data Mining Techniques," IEEE, 2008.

[48] X. D. V. S. S. a. T. A. V. I. SUMEET DUA, "NOVEL CLASSIFICATION OF CORONARY ARTERY DISEASE USING HEART RATE VARIABILITY ANALYSIS," Journal of Mechanics in Medicine and Biology , vol. 12, no. 4, pp. 1-22, 2012.

[49] V. H. V. V. A. M. S. P. Mohammad Ismail, "An Effective Heart Disease Prediction Method using Artificial Neural Network," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 8, pp. 1529-1532, 2019.

[50] D. R. D. A. K.Srinivas, "Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques," IEEE, pp. 1344-1349, 2010.

[51] V. B. Anchana Khemphila, "Heart disease Classification using Neural Network and Feature Selection," IEEE, pp. 406-409, 2011.

[52] [Online]. Available: https://en.wikipedia.org/wiki/NeuroSolutions.

[53] W.-S. K. K. Y. N. Heon Gyu Lee, "Coronary artery disease prediction method using linear and nonlinear feature of heart rate variability in three recumbent postures," Information Systems Frontiers, 2009.

[54] D. R. V. P. R. Anju Gulia, "Liver patient classification using Intelligent Techniques," International Journal of Computer Science and Information Technologies, vol. 5, no. 4, pp. 5110-5115, 2014.

[55] O. Gregory, "Prediction of liver disease (biliary cirrhosis) using datamining technique," International Journal of Engineering Technology and Research, vol. 1, no. 1, pp. 2347-6079, 2015.

[56] E. O. O. K. Aadnan, "Liver disease diagnosis based on neural networks," in Advances in Computational Intelligence, 2015.

[57] S. K. P. Tapas Ranjan Baitharua, "Analysis of datamining techniques for healthcare decision support system using liver disorder dataset," in International Conference on Computational Modeling and Security, 2016.

APPENDIX A

Sample of Collected Data

Age	Sex	Chest pain	Blood pressure	Cholesterol	Blood sugar	Electrocardiogram	Heart rate	Angina	Old peak	Slope	Vessels	Thalassemia	Target
54	0	1	132	288	1	0	159	1	0	2	1	2	1
45	0	1	112	160	0	1	138	0	0	1	0	2	1
53	1	0	142	226	0	0	111	1	0	2	0	3	1
62	0	0	140	394	0	0	157	0	1.2	1	0	2	1
52	1	0	108	233	1	1	147	0	0.1	2	3	3	1
43	1	2	130	315	0	1	162	0	1.9	2	1	2	1
53	1	2	130	246	1	0	173	0	0	2	3	2	1
42	1	2	120	240	1	1	194	0	0.8	0	0	3	1
50	1	2	129	196	0	1	163	0	0	2	0	2	1
68	0	2	120	211	0	0	115	0	1.5	1	0	2	1
69	1	3	160	234	1	0	131	0	0.1	1	1	2	0
45	0	0	138	236	0	0	152	1	0.2	1	0	2	0
50	0	1	120	244	0	1	162	0	1.1	2	0	2	0
50	0	0	110	254	0	0	159	0	0	2	0	2	0
64	0	0	180	325	0	1	154	1	0	2	0	2	0
57	1	2	150	126	1	1	173	0	0.2	2	1	3	0
64	0	2	140	313	0	1	133	0	0.2	2	0	3	0
43	1	0	110	211	0	1	161	0	0	2	0	3	0
55	1	1	130	262	0	1	155	0	0	2	0	2	0
37	0	2	120	215	0	1	170	0	0	2	0	2	0
41	1	2	130	214	0	0	168	0	2	1	0	2	0
56	1	3	120	193	0	0	162	0	1.9	1	0	3	1
46	0	1	105	204	0	1	172	0	0	2	0	2	1
46	0	0	138	243	0	0	152	1	0	1	0	2	1

APPENDIX B

Algorithms

Radial Basis Function

Step 1: Define the number of hidden neurons ‘K’

Step 2: Set the positions of RBF centers using K-means clustering algorithm.

Step 3: Calculate σ

Step 4: Calculate actions of RBF node

Step 5: Train the output

Recurrent Network

Step 1: A single time step of the input is provided to the network.

Step 2: Then calculate its current state using set of current input and the previous state.

Step 3: The current ht becomes ht-1 for the next time step.

Step 4: One can go as many time steps according to the problem and join the information from all the previous states.

Step 5: Once all the time steps are completed the final current state is used to calculate the output.

Step 6: The output is then compared to the actual output i.e the target output and the error is generated.

Step 7: The error is then back-propagated to the network to update the weights and hence the network (RNN) is trained.

Support Vector Machine

Step 1: Data Pre-processing

Step 2: Importing datasets

Step 3: Extracting Independent and dependent Variable

Step 4: Splitting the dataset into training and test set.

Step 5: Feature Scaling

Step 6: Predicting the test set result

Step 7: Creating the confusion matrix and AUC-ROC curve

Ensemble Classifier

Step 1: Input, Training data D

Step 2: Output, Enseble classifier H

Step 3: Learn base-level classifiers

Step 4: for t=1 to T do

Learn H_t base on D

Step 5: end for

Step 6: Construct new data for predictions

Step 7: Return H

K-Nearest Neighbor Classifier

Step 1: Select the number K of the neighbors.

Step 2: Calculate the Euclidean distance of K number of neighbors.

Step 3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step 4: Among these k neighbors, count the number of the data points in each category.

Decision Tree classifier

Step 1: Begin the tree with the root node, says S, which contains the complete dataset.

Step 2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).

Step 3: Divide the S into subsets that contains possible values for the best attributes.

Step 4: Generate the decision tree node, which contains the best attribute.

Step 5: Recursively make new decision trees using the subsets of the dataset created in step 3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

APPENDIX C

Code

Radial Basis Function

```
iw=cell2mat(net.iw(1));        %input weight

lw=cell2mat(net.lw(2,1));      %layerweight

ib=cell2mat(net.b(1));         %input bias

lb=cell2mat(net.b(2));         %layer bias

a= radbas(netprod(dist(iw,pn),ib));      %radial basis transfer function

a2=dotprod(net.lw{2,1},a);               %product of first layer output with layer weights

a2=gadd(a2,net.b{2});                    %add with layer bias

a2=mapminmax('reverse',a2,ts);           %reverse the normalization as followed in the newrb which was from 0.1 to 0.9
```

Recurrent Network

```
[X,T] = simpleseries_dataset;
net = layrecnet(1:2,10);
[Xs,Xi,Ai,Ts] = preparets(net,X,T);
net = train(net,Xs,Ts,Xi,Ai);
view(net)
Y = net(Xs,Xi,Ai);
perf = perform(net,Y,Ts)
```

Support Factor Machine

```
function [trainedClassifier, validationAccuracy] = trainClassifier(datasetTable)
predictorNames = {'Age', 'Sex', 'Chestpaincp', 'BloodpressuremmHg', 'Cholesterolmgdl', 'Bloodsugarmgdl', 'ElectocardiogramECG', 'Heartrate', 'Angina', 'Oldpeak', 'Slope', 'Vessels', 'Thalassemia'};
predictors = datasetTable(:,predictorNames);
predictors = table2array(varfun(@double, predictors));
response = datasetTable.Target;
```

```
trainedClassifier = fitcsvm(predictors, response, 'KernelFunction', 'linear', 'PolynomialOrder', [], 'KernelScale', 'auto', 'BoxConstraint', 1, 'Standardize', 1, 'PredictorNames', {'Age' 'Sex' 'Chestpaincp' 'BloodpressuremmHg' 'Cholesterolmgdl' 'Bloodsugarmgdl' 'ElectocardiogramECG' 'Heartrate' 'Angina' 'Oldpeak' 'Slope' 'Vessels' 'Thalassemia'}, 'ResponseName', 'Target', 'ClassNames', [0 1]);
partitionedModel = crossval(trainedClassifier, 'KFold', 10);
validationAccuracy = 1 - kfoldLoss(partitionedModel, 'LossFun', 'ClassifError');
```

Ensemble Classifier

```
function [trainedClassifier, validationAccuracy] = trainClassifier(datasetTable)
predictorNames = {'Age', 'Sex', 'Chestpaincp', 'BloodpressuremmHg', 'Cholesterolmgdl', 'Bloodsugarmgdl', 'ElectocardiogramECG', 'Heartrate', 'Angina', 'Oldpeak', 'Slope', 'Vessels', 'Thalassemia'};
predictors = datasetTable(:,predictorNames);
predictors = table2array(varfun(@double, predictors));
response = datasetTable.Target;
trainedClassifier = fitctree(predictors, response, 'PredictorNames', {'Age' 'Sex' 'Chestpaincp' 'BloodpressuremmHg' 'Cholesterolmgdl' 'Bloodsugarmgdl' 'ElectocardiogramECG' 'Heartrate' 'Angina' 'Oldpeak' 'Slope' 'Vessels' 'Thalassemia'}, 'ResponseName', 'Target', 'ClassNames', [0 1], 'SplitCriterion', 'gdi', 'MaxNumSplits', 4, 'Surrogate', 'off');
partitionedModel = crossval(trainedClassifier, 'KFold', 10);
validationAccuracy = 1 - kfoldLoss(partitionedModel, 'LossFun', 'ClassifError');
```

Decision Tree Classifier

```
function [trainedClassifier, validationAccuracy] = trainClassifier(datasetTable)
predictorNames = {'Age', 'Sex', 'Chestpaincp', 'BloodpressuremmHg', 'Cholesterolmgdl', 'Bloodsugarmgdl', 'ElectocardiogramECG', 'Heartrate', 'Angina', 'Oldpeak', 'Slope', 'Vessels', 'Thalassemia'};
predictors = datasetTable(:,predictorNames);
predictors = table2array(varfun(@double, predictors));
response = datasetTable.Target;
trainedClassifier = fitcsvm(predictors, response, 'KernelFunction', 'linear', 'PolynomialOrder', [], 'KernelScale', 'auto', 'BoxConstraint', 1, 'Standardize', 1, 'PredictorNames', {'Age' 'Sex'
```

'Chestpaincp' 'BloodpressuremmHg' 'Cholesterolmgdl' 'Bloodsugarmgdl' 'ElectocardiogramECG' 'Heartrate' 'Angina' 'Oldpeak' 'Slope' 'Vessels' 'Thalassemia'}, 'ResponseName', 'Target', 'ClassNames', [0 1]);
partitionedModel = crossval(trainedClassifier, 'KFold', 10);
validationAccuracy = 1 - kfoldLoss(partitionedModel, 'LossFun', 'ClassifError');

K-Nearest Neighbor Classifier

function [trainedClassifier, validationAccuracy] = trainClassifier(datasetTable)
predictorNames = {'Age', 'Sex', 'Chestpaincp', 'BloodpressuremmHg', 'Cholesterolmgdl', 'Bloodsugarmgdl', 'ElectocardiogramECG', 'Heartrate', 'Angina', 'Oldpeak', 'Slope', 'Vessels', 'Thalassemia'};
predictors = datasetTable(:,predictorNames);
predictors = table2array(varfun(@double, predictors));
response = datasetTable.Target;
trainedClassifier = fitcknn(predictors, response, 'PredictorNames', {'Age' 'Sex' 'Chestpaincp' 'BloodpressuremmHg' 'Cholesterolmgdl' 'Bloodsugarmgdl' 'ElectocardiogramECG' 'Heartrate' 'Angina' 'Oldpeak' 'Slope' 'Vessels' 'Thalassemia'}, 'ResponseName', 'Target', 'ClassNames', [0 1], 'Distance', 'Euclidean', 'Exponent', '', 'NumNeighbors', 1, 'DistanceWeight', 'Equal', 'StandardizeData', 1);
partitionedModel = crossval(trainedClassifier, 'KFold', 10);
validationAccuracy = 1 - kfoldLoss(partitionedModel, 'LossFun', 'ClassifError');