Towards Heart Disease Prediction using Hybrid Data Mining

A Thesis

submitted in partial fulfilment of the requirements for the award of degree of

Master of Engineering

in

Computer Science & Engineering

by

Meenal

Roll No. 801532030

Under the supervision of

Dr. Niyati Baliyan
Assistant Professor

Ms. Vineeta Bassi
Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

July 2017

Certificate

I hereby certify that the matter which is being presented in the thesis titled, **"Towards Heart Disease Prediction using Hybrid Data Mining"**, in partial fulfilment of the requirements for the award of degree of Master of Engineering in Computer Science and Engineering submitted in **Computer Science and Engineering Department** of Thapar University, Patiala, is an authentic record of my own work, under the supervision of **Dr. Niyati Baliyan** and **Ms. Vineeta Bassi** and refers other researchers' work which is duly listed in the reference section.

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

Meenal
801532030
M.E. (CSED)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Dr. Niyati Baliyan
Assistant Professor
CSED
Thapar University
Patiala

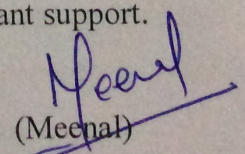
Ms. Vineeta Bassi
Assistant Professor
CSED
Thapar University
Patiala

Acknowledgement

First of all, I would like to thank the Almighty, who has always guided me to work on the right path of life. This work would not have been possible without the encouragement and able guidance of my supervisors - **Dr. Niyati Baliyan** and **Ms. Vineeta Bassi**. I thank my supervisors for their time, patience, discussions and valuable comments. Their enthusiasm and optimism made this experience both rewarding and enjoyable.

I am equally grateful to **Dr. Maninder Singh**, Head of Computer Science and Engineering Department, a nice person, an excellent teacher and a well-credited researcher, who always encouraged me to keep working well and always advised me with his invaluable suggestions. I will be failing in my duty if I do not express my gratitude to **Dr. S.S. Bhatia**, Dean of Academic Affairs, for making provisions of infrastructure such as library, computer labs equipped with Internet facilities, immensely useful for the learners to equip themselves with the latest in the field. I am also thankful to the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, love, and affection, which made my stay at Thapar University memorable.

Last but not least, I would like to thank my family whom I dearly love and without whose blessings none of this would have been possible. To my parents, I own thanks for their care and encouragement. I would also like to thank my brother since he insisted that I should do so. I would also like to thank my close friends for their constant support.



(Meenal)

Abstract

Nowadays, heart diseases are very common and one of the major causes of death across the globe. This calls for accurate and timely diagnosis of the heart disease. Although, the healthcare industry has come a long way to treat patients with several kinds of diseases, yet the prediction of heart diseases is a complicated task in the healthcare field. Therefore, it is essential to develop a decision support system for analysing the heart disease in a patient. A heart disease is a dangerous disease which is not visible to naked eyes. Bad decision making by the physician can even cause death of a patient. To avoid these kinds of decisions, numerous hospitals use the clinical information system to manage the data of patients' health. There is abundant data available with the health care systems; however, the knowledge about the data is rather poor. The accessibility of the enormous size of medical dataset hints towards the requirement of a tool which analyses data to extract valuable information. Unfortunately, this data is hardly used to support the healthcare decision making. There are huge amounts of hidden patterns in this data which are yet to be explored; this gives rise to the question that how we can extract useful information from these patterns. Thus, it is essential to form a model with the help of standard datasets to predict the heart disease of the patients even before it occurs. Data scientists have attempted several analytical methods in order to improvise the examination of heart diseases. Previously, various data mining techniques have been implemented in the healthcare systems, however, hybridization in addition to single technique in the identification of heart disease shows promising outcomes, and can be useful in further investigating the treatment of the heart diseases. Additionally, this can reduce the cost of treatment. This work attempts to survey some recent techniques applied towards knowledge discovery for heart disease and further proposes a novel prediction method using bagging and boosting to attain improved accuracy.

Keywords—Data Mining, Heart Disease, Classification, Bagging, Boosting

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACO	Ant Colony Optimization
ANN	Artificial Neural Network
ARF	Acute Rheumatic Fever
BN	Bayesian Net
BRNN	Bidirectional Recurrent Neural Network
CAD	Carotid Artery Disease
CKD	Chronic Kidney Disease
Ctree	Conditional Inference Tree
DT-GI	Decision Tree using Gini Index
DT-IG	Decision tree using Information Gain
ELM	Extreme Learning Machine
GAM	Generalised Adaptive Model
GLM	Generalized Linear Model
HCWV	Hybrid Classifier using Weighted Voting
HDPS	Heart Disease Prediction System
HMV	Hierarchical Majority Voting
KDD	Knowledge Discovery in Databases
k-NN	k-Nearest Neighbour
LASSO	Least Absolute Shrinkage and Selection Operator
LR	Linear Regression
MLP	Multilayer Perceptron
NB	Naïve Bayes
PAD	Peripheral Artery Disease
QDA	Quadratic Discriminant Analysis
RNN	Recurrent Neural Network
SVM	Support Vector Machine
TDNN	Time Delay Neural Network
UCI	University of California, Irvine

Chapter 1
Introduction

Heart disease is one of the main reasons of death in India and all across Asia. An estimated 17.7 million people died due to heart disease in 2015, which is 31% of all global deaths [1]. There are lifestyle factors such as alcohol, high blood pressure, physical inactivity, hypertension, obesity, diabetes etc. which established the major risk of heart disease. Therefore, researchers have been able to identify these factors that may cause heart disease [2]. With the pervasion of Artificial Intelligence in the healthcare industry, we can actually help in detecting and preventing heart diseases. For better decision making for the discovery of heart disease, data mining and machine learning help in extracting the useful data from huge clinical database available in the hospitals to enhance the quality of the patients' health. There are many classification techniques such as Support Vector Machine, Linear Model, and Decision Tree etc. which help in building an intelligent classifier for heart disease prediction, however, these classifiers are weak classifiers and they need bagging and boosting techniques to enhance their performance. The challenging task in the healthcare data is to choose the adequate technique which leads to best classification result [3]. This work aims at developing a model based on hybrid techniques which is implemented on different classifiers with weighted voting to obtain better accuracy [4].

Data mining discloses relations among large amounts of data across numerous data sets. The process of extracting the knowledge from data comprises an iterative order, i.e., cleaning, integration and selection of data, pattern recognition and information representation. The applications of data mining are in many fields such as risk assessment and crime detection etc. [5]. Data mining is also helpful in detection of gaining and deeper knowledge of a disease, making better health care policies, taking a useful decision for preventing a disease, minimizing the death rates in the hospitals, cost saving, fraud insurance, and others [6]. Diagnosis of various diseases, namely, cancer, diabetes, etc. can also be carried out with the help data mining approach [7]. Data mining is known for more than two decades now, however, its capability is being leveraged to a great extent only now. To extract the knowledge and hidden patterns from vast database, we can combine machine learning, statistical knowledge and database technology as one

hybrid model of data mining [8], [9].

1.1 Heart Diseases

There are several kinds of problems that can affect the functioning of heart, many of which include Coronary Artery Disease (CAD) also called atherosclerosis. Atherosclerosis is a condition where a substance called plaque forms on the walls of arteries, which are blood vessels that carry blood to the heart and other parts of the body. Plaque is composed of calcium, cholesterol, fat etc. With the passage of time, plaque may partially block the wall and narrow the arteries; it may decrease the flow of blood as shown in Figure 1.1 [10]. The flow of blood can be stopped as a result of blood clot formation, this can cause heart attack. Atherosclerosis can affect artery in any part of the body such as in the kidneys, heart, brain, and legs which results in different diseases in the part of the body which is affected.

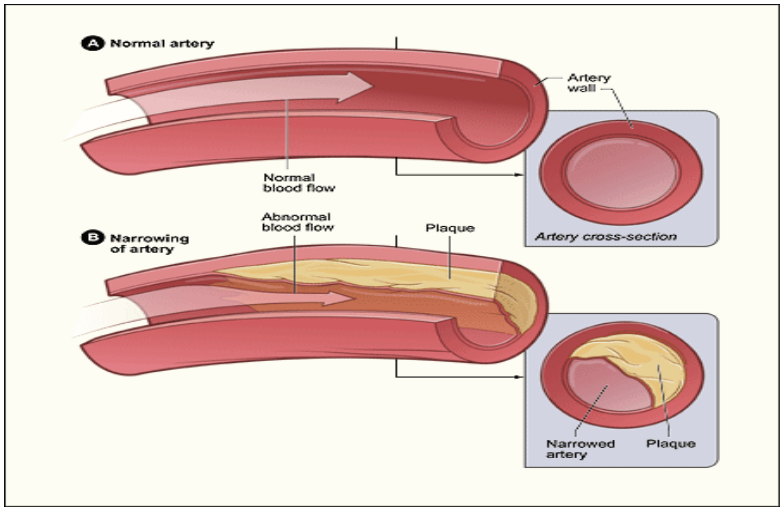


Figure 1.1 Atherosclerosis

1.1.1 Coronary Heart Diseases

Coronary Heart Disease (CHD) arises when plaque is deposited on the walls of coronary arteries. Plaque can harden with time. The hardened plaque reduces the flow of blood and narrows the coronary arteries which results in partial or complete block of the blood flow as shown in Figure 1.2 [11]. With the reduced or blocked flow of blood, chest pain, discomfort or heart attack may be caused.

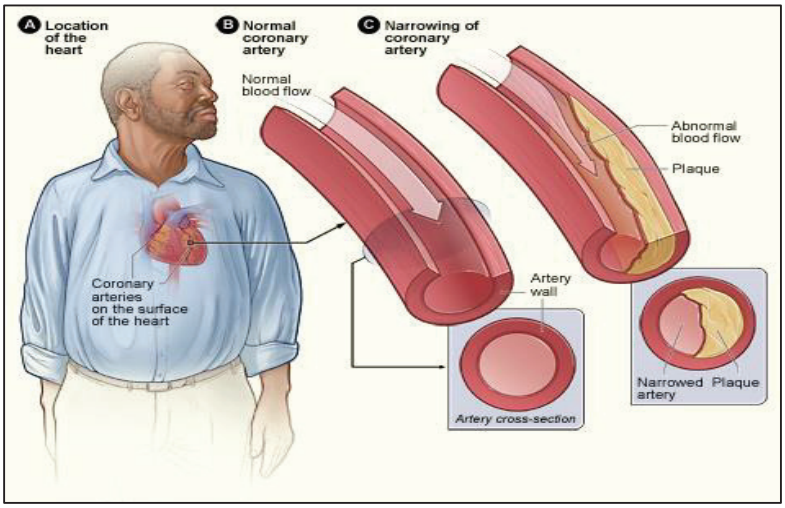


Figure 1.2 Coronary Heart Diseases

1.1.2 Carotid Artery Diseases

Carotid Artery Disease (CAD) is caused when plaque collects in the arteries on both sides of neck. Arteries bring oxygen-rich blood to the brain. The reduced or blocked flow of blood may cause stroke. This disease is serious because it can cause brain attack. It occurs when there is no flow of blood to brain, and within few minutes, the cells of brain start dying. There are two common types of carotid arteries, on both sides of neck. Each carotid artery divides into internal and external carotid arteries as shown in Figure 1.3 [12].

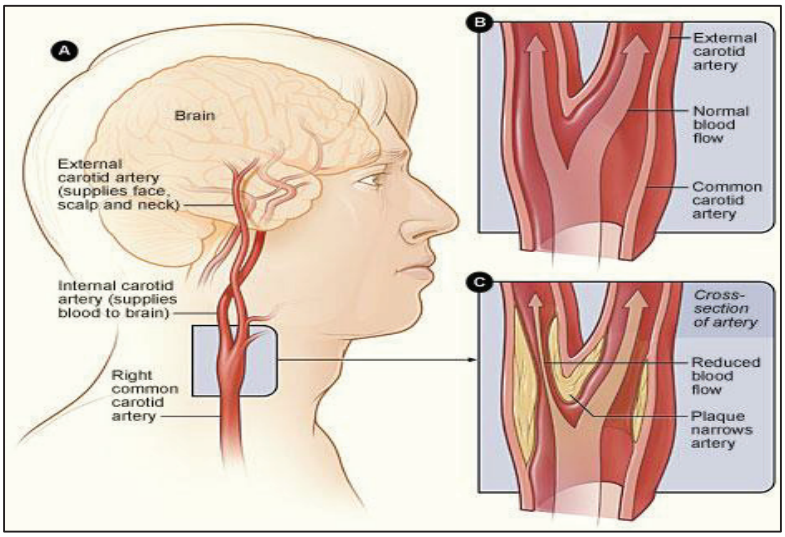


Figure 1.3 Carotid Artery Diseases

1.1.3 Peripheral Artery Diseases

Peripheral Artery Disease (PAD) occurs when plaque collects in the main arteries that supply oxygen-rich blood to pelvis, legs, and arms. Reduced or blocked flow of blood, may cause pain or risky infection. Plaque substance is prepared from calcium, fat, fibrous tissue, cholesterol, calcium, and other elements in the blood as shown in Figure 1.4 [13]. PAD usually disturbs the arteries in the legs.

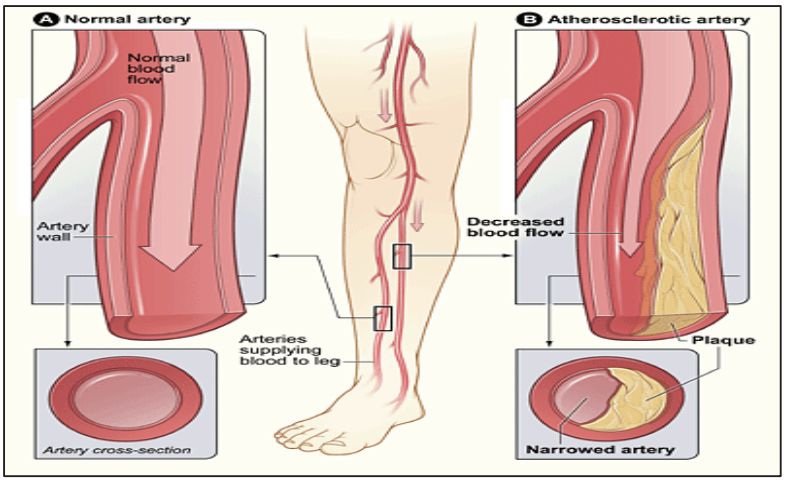


Figure 1.4 Peripheral Artery Diseases

1.1.4 Chronic Kidney Diseases

Chronic Kidney Disease (CKD) can occur when plaque collects in renal arteries. These arteries provide oxygen-rich blood to the kidneys. With the passing period, functioning of kidney slows down as shown in Figure 1.5 [14]. The major task of kidney is to removal wastage and unnecessary water out of the body.

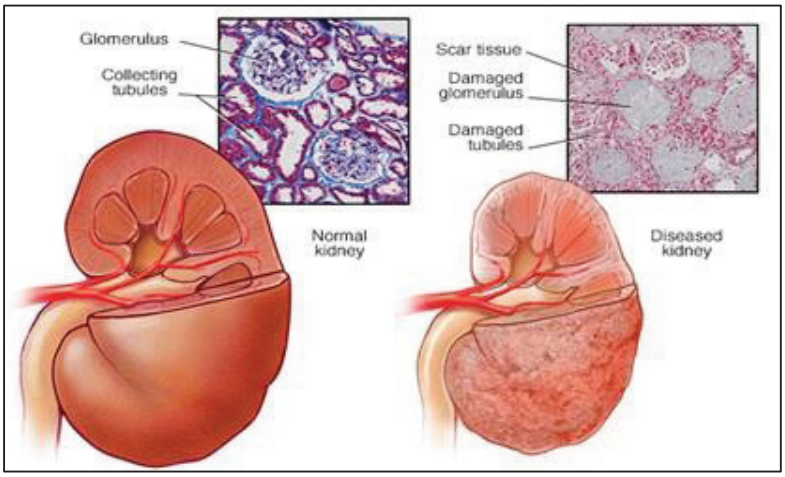


Figure 1.5 Chronic Kidney Diseases

1.1.5 Other Cardiovascular Diseases

- **Heart failure**

Heart failure is a state in which the heart is not pumping the blood that can meet the requirements of the body. The heart keeps working but the need for blood and oxygen does not meet, and heart failure occurs. Heart failure is a serious situation that requires medical care.

- **Arrhythmia**

Arrhythmia is the irregular beat of the heart. There are different kinds of arrhythmia, i.e., heart beat can be very slow, very fast or uneven. Bradycardia is the condition where the heart beat is lower than 60 beats every minute. Arrhythmia can affect the working of heart, as pumping of blood and oxygen does not meet the requirements of the body.

- **Heart valve problems**

Heart valve problem is caused if the heart valve does not work in the expected manner. The heart valve is at the end of four heart chambers and helps in maintaining one way flow of blood through heart. The four valves of heart chambers ensure that blood flows easily in forward direction and without any backward leakage. When the valve of heart is not sufficiently opened, the narrow opening of heart valve make it difficult to pump blood through it, this may lead to heart failure, a condition known as stenosis. When the heart valves do not close properly or tightly, some blood will leak backwards through the valve; this condition is known as regurgitation.

1.2 Data Mining

The discovery of information from the database consists of different steps, of which data mining is an essential step that brings about the valuable information from the hidden patterns of a huge database [15], [16]. Data mining is the process of extracting valuable information from the huge data repositories; otherwise useful patterns might remain unknown. There is large quantity of data obtainable in different industries like hospitals, healthcare industry etc. This data is useless unless it is transformed into valuable knowledge. Thus, it is essential to examine the large quantity of data and retrieve valuable knowledge from it. There are different processes involved in data mining such as cleaning of data, integration of data, and transformation of pattern analysis and

information visualization. When these procedures are completed, the extracted information provides the abilities to forecast the outcomes of future statement.

In the healthcare domain, the treatment is given based on the intuition or experience of the health expert, rather than using the valuable information gained from some relevant database. There are many determinants of diagnosis of heart disease, doctors usually decide on the basis of patient's current test result and also examine the decision made for the same kind of patient's condition earlier [17]. This kind of treatment consists of bias, errors or expensive medical treatment and also leads to inferior service and care of the patient, this may also result in an adverse effect on the patient's health. This task is not easy for physician to combine all the factors that are needed to evaluate the heart disease. Therefore, physician needs to be more experienced or highly skilled for diagnosis of heart disease. Wu, et al. [18] recommended ensemble of decision making with patient's real data record that can reduce health care errors, can update the security of the patients, reduce unwanted expense and can enrich treatment accuracy. This technique can help in data modelling and creating heart disease analysis tools, in order to produce information that enhances the decision making process in healthcare domain.

Recent advancements in the field of healthcare industry, artificially intelligent expert systems have been developed for the medical applications. Computational tools have been designed for the experts from last few decades, to improve the decision making ability of the physician for the patient [19]. Wu, et al.'s technique can help in data modelling and creating heart disease analysis tools, in order to produce information that enhances the decision making process in healthcare domain.

1.2.1 Knowledge Discovery

The Knowledge Discovery in Databases (KDD) is the wide procedure of extracting knowledge in the data, and emphasizes the high-level use of specific data mining approaches. With highly qualified physicians or expert staff who have good knowledge in healthcare domain, KDD may be developed for healthcare field. KDD may be helpful in working with huge amount of data which can extract the meaningful patterns and to improve the decision making. The steps of KDD are shown in Figure 1.6. Analysts of insurance companies, physicians, health care industry, personnel policy makers, pharmacological companies etc. and apply KDD to their domain [20].

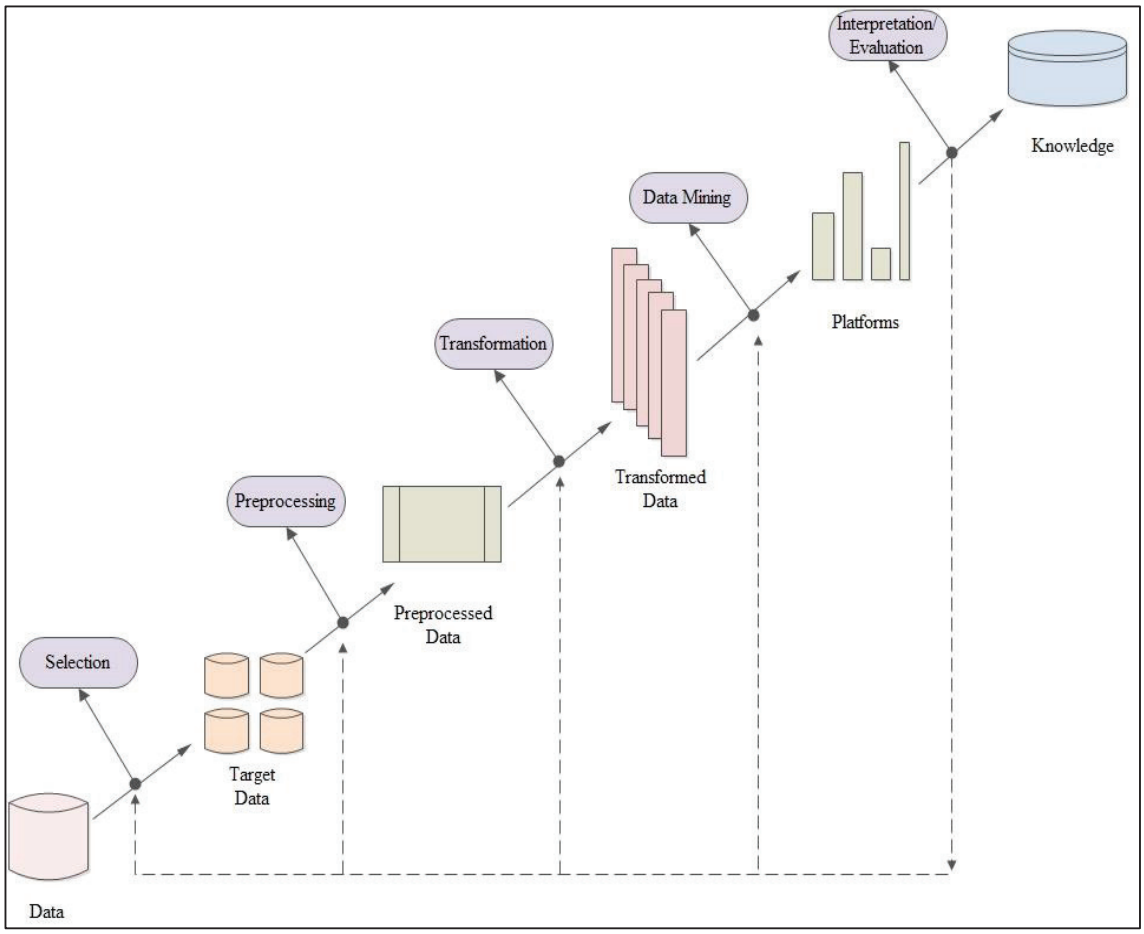


Figure 1.6 Steps of Knowledge Discovery in Databases

There are four major applications of KDD in the healthcare industry:

1. **Clinical Medicine**: Nowadays, hospitals and clinics exceed their traditional role of diagnosis of disease and treatment and are now acting as a huge source of data of complex clinics, equipment use, laboratory and drug management data which can help in the analysis of heart disease and decision making;
2. **Public Health**: It involves the early detection of disease and any disorder observed in the patients' health;
3. **Healthcare Text mining**: It includes the medical literature survey and extracting information from the healthcare clinical data such as patients' clinical records;
4. **Healthcare Policy and Planning:** It including the expensive treatments for the patients' disease. KDD helps the healthcare planners to solve the resource allocation problems and capacity problems.

1.2.2 Classification

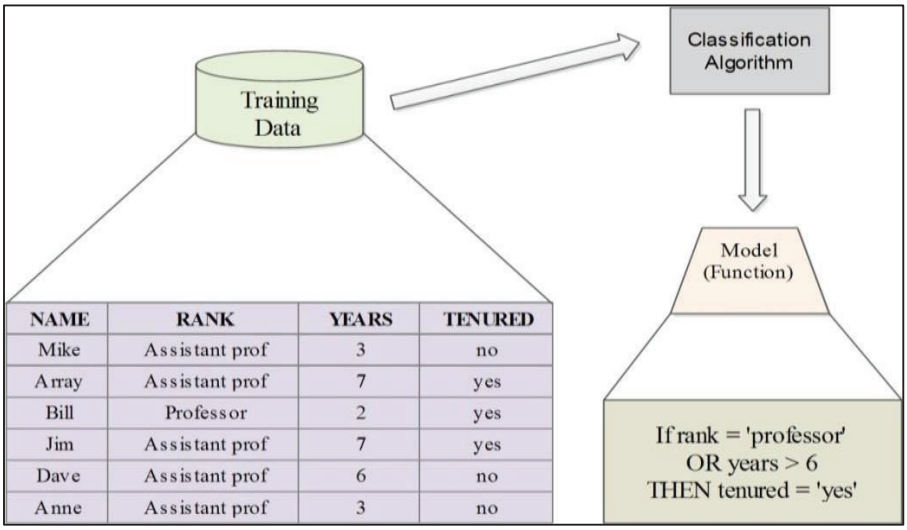
Classification is the process of assigning a label to the objects based on the analysis of their features. It includes features of recurrence in the same class or group to which the feature belongs. Any wrong identification and analysis may often lead to misclassification. In data mining, classification predicts the outcomes of the data on basis of input. To predict the outcomes, the technique is applied on the training dataset having a set of attributes and their outcomes, usually named as predictive attributes. The technique attempts to find out the relations among the attributes that would render it probable to predict the conclusion. Next, it predicts the class of a data not obtained earlier, called prediction set, which holds the same set of attributes, except the unknown attributes, i.e., prediction attribute. The technique examines the training data in the form of input and produces the output in form of predicted data. The accuracy of the prediction data states how 'good' the technique is.

Classification allocates data item in a group to target classes or groups. The main objective of classification is to predict the target class accurately for every instance in data. The task of classification initiates through data set in which classes are known. Classification tests the predicted values to the known target values by comparison between them on test data set. Classification is supervised machine learning [21].

Classification method is a two-step procedure, model building and using model for classification

- *Model building*

Building up the model is the knowledge creation stage of classification process in which model is built on the basis of classification algorithm. This can be understood by Figure 1.7. The dataset is bifurcated, i.e., training data and testing data. A model is built by using a training dataset having samples that belong to well-known class or category. Samples can also be called as objects, tuples, or data points. This training data is tested on the testing dataset.

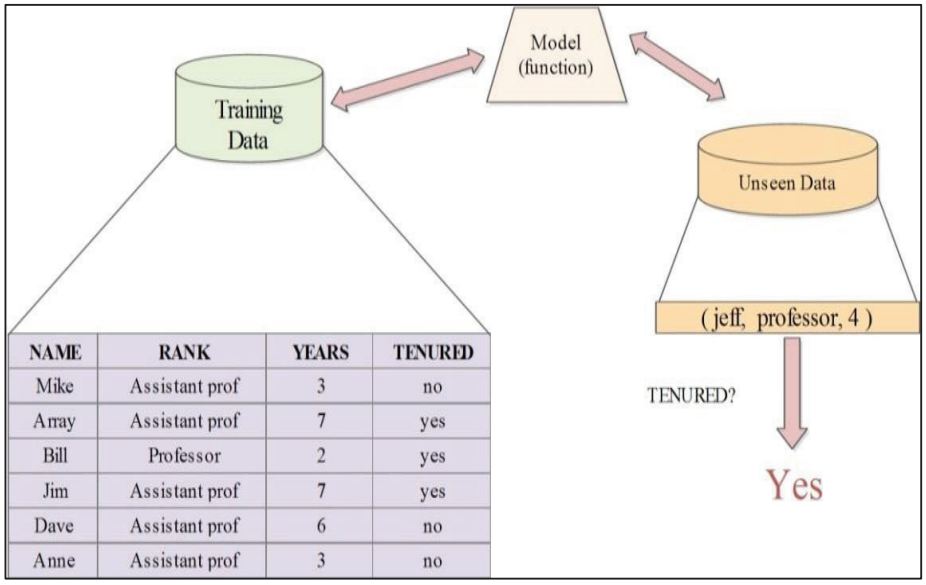


NAME	RANK	YEARS	TENURED
Mike	Assistant prof	3	no
Array	Assistant prof	7	yes
Bill	Professor	2	yes
Jim	Assistant prof	7	yes
Dave	Assistant prof	6	no
Anne	Assistant prof	3	no

Figure 1.7 Building the Classifier

- *Using model for classification*

In the next phase of classification, the model is used for classification process as shown in Figure 1.8. If the accuracy of the model is adequate then some classification rules may be applied to the unseen data.



NAME	RANK	YEARS	TENURED
Mike	Assistant prof	3	no
Array	Assistant prof	7	yes
Bill	Professor	2	yes
Jim	Assistant prof	7	yes
Dave	Assistant prof	6	no
Anne	Assistant prof	3	no

Figure 1.8 Classifications through Classifier

1.3 Issues in Data Mining

Data mining application depends on the dataset to supply the input. If the dataset is noisy, incomplete, dynamic and large then it can make further complications to the data mining application. The issues in data mining can be listed as below.

- *Limited information*

In the organizations, dataset is designed for the general purpose not for the data mining purpose. There is a possibility that some of the very essential information may be missing in the dataset. It will be difficult to retrieve the useful information from this limited dataset.

- *Noise and Missing data*

Incorrect data may lead the data mining application into the wrong result. Sometimes client's real data may not be collected in the dataset because client was not interested to give the real data, simply for the sake of giving he might have filled the incorrect data. Sometimes client may not fill the complete data that is called as missing data. To remove the noise and missing data problem, at the time of data pre-processing we have to do the data cleaning.

- *User interaction and prior knowledge*

To retrieve the hidden information from the dataset, the analyst must be familiar with the dataset domain knowledge otherwise he will not understand the dataset correctly and interpret it properly. The use of domain knowledge discovery is important in all steps of the knowledge discovery in database process. It should be both interactive and iterative to collect more domain knowledge.

- *Size and irrelevant fields*

In an organization because of very frequent transactions with their client, the dataset size may be added or become very big. The contents in the dataset keep changing as information is modified or removed. We need up-to-date and consistent information for knowledge recovery process.

- *The nature of the healthcare dataset*

The medicine datasets are hierarchical, complex, time series, heterogeneous, large, and of varying quality. Raw medical data are huge and mixed. Medical data are collected from several sources like physician's notes, interpretation and interviews with the patients. There is no specific mathematical formula to calculate the final prediction in the medicine data like other sciences. All the datasets are not going to be in the uniform format. Different people use different words and grammars to mention the same problem and conclusion.

- *High Sensitive health care data*

Because of moral, authorised and communal aspects, the medicine datasets are very sensitive ones. It is very difficult to get the medical dataset for the mining purpose. The patient's data are highly sensitive one.

- *Needed for Active Collaboration*

Due to the absence of knowledge in the medical field, the data miner needs active collaboration with the domain expert.

1.4 Motivation of the Research

This section discusses the motivation for the research in applying machine learning methods to medical diagnosis. Nowadays, data mining methods are used in the health care industry to detect the fraud claims in medical insurance sector. Even though applying data mining in medicine is a little bit risky, it can help the health care domain in various ways. The list below contains various uses that data mining brings to the health care industry [22].

- *Useful Knowledge from Database*

There is ample knowledge available in the computerized healthcare. Even though lot of computerized patient data is available, it is very complicated one and it is not easy to find the important knowledge from the electronic data.

- *Avoidance of Hospital Errors*

When healthcare organizations apply data mining on the past medical records, they can find fresh, valuable and possibly lifesaving information. They can find various errors committed in the past in the time of diagnosis or treatment that can be avoided in the future cases.

- *Policy Making in Public health*

When we apply data mining in the past medical data we can find the needed prevention methods in the future that help the higher policy makers to have some policies to prevent that type of diseases in the future.

- *Money value and saving cost*

Data mining facilitates the healthcare organizations to get more value from the current data and minimize the cost.

- *Early detection and prevention of diseases*

Classification algorithms in data mining can help us to detect future problems very early and try to prevent it in the patients' life. Early detection of heart disease, cancer diseases can make a huge impact in the people's life.

1.5 Thesis Outline

In Chapter 2, some frameworks for heart disease prediction developed so far have been discussed in brief, as part of the literature survey.

Chapter 3 discusses the problem statement.

Chapter 4 presents the basic methodology that was adopted to cater to heart disease prediction.

Chapter 5 describes the proposed work.

Chapter 6 presents the results and the discussion of the Heart Disease prediction model that has been built.

Chapter 7 concludes the work and outlines possible future work.

Chapter 2
Literature Review

In this Chapter, we reviewed the various researches conducted using heart disease dataset. From the literature review, we tried to understand the different machine learning algorithms applied in the heart disease dataset and compare their accuracy. In the machine learning methods, there are different kinds of algorithms that are used in the healthcare, such as Support Vector Machine (SVM), Neural Network (NN), Naïve Bayes (NB), Decision Tree (DT), Generalised Linear Model (GLM), Lasso, Quadratic Discriminant Analysis (QDA), Classification and Regression Tree (rpart2), Genetic Algorithm (GA), Multivariate Adaptive Regression Spline (earth), Bagging, Boosting technique and hybrid techniques.

2.1 Background

KDD and data mining is an interdisciplinary zone concentrating on the methods for retrieving useful patterns and information from the data. In the recent years, it comes out that the extraction of useful information from huge database and valuable decision making has been steadily increasing. KDD is the central process for transforming data into knowledge for valuable decision making, known as data mining. Machine Intelligence is used for medical data mining and it extracts biomedical and health care knowledge for clinical decision making and generates scientific hypothesis from large medical data. As a result, KDD becomes the most important tool in healthcare because it uses data mining technique.

Over the years, data mining is broadly used in the parts of computer and engineering, genetics and healthcare. Data mining plays a vital role in tackling overloaded data. There is huge database in the healthcare, which is used for extracting useful information.

With the progress of information skills and extensive medical data being available, medical data classification plays a crucial role in many medical applications. It is the procedure of converting reports of healthcare diagnosis and processes to general codes. Diagnosis codes are helpful in diagnosis of disease and healthcare conditions. Medical classification is widely used in hospitals for the statistical analysis of diseases and therapies. It addresses the problems of diagnosis, analysis and teaching purposes in

medicine. Medical data has made a great progress over the past decades in the development and use of classification algorithms. In healthcare, medical data can be transformed into aggregations, to calculate average values per patient and compare with other values, to group data into clusters of similar data etc. There is collection of heterogeneous tools and techniques used in the data mining. These heterogeneous techniques are based on machine learning, statistics, visualization etc. and these can be useful in discovering hidden patterns and providing additional knowledge for making better decisions.

The reasons why healthcare will benefit from the applications of data mining are:

- Health insurance companies try to reduce monetary loss due to fraud by using data mining.
- Large quantity of data is produced during healthcare transactions.
- Data mining improves decision-making since it is used for the discovery of trends and patterns in large amounts of different types of data.

The above mentioned benefits are due to the fact that healthcare organizations that implement data mining have improved predictions for mid and long-term requirements. There are algorithms which can automatically categorize the healthcare data built on the basis on similarities of rules and patterns attained between the training and testing dataset. Nowadays, the hybrid approach has become a popular and it gives better result as compared to others, during classification [23]. Machine learning provides technical basis for data mining which can be regarded as a confluence of statistics and machine learning.

2.2 Related Work

This section summarizes the data mining approaches employed for heart disease detection. Researchers have used various methods to advance the diagnosis of heart disease for many years. A particular diagnostic dataset by University of California, Irvine, and Cleveland is commonly popular and available online [24]. Few recent research works in the domain of interest are as follows:

In 2016, S. Rajathi and G. Radhamani proposed an integrated framework using k-Nearest Neighbor (k-NN) with Ant Colony Optimization (ACO) technique to forecast the

probability of having heart disease [25]. The study has been achieved in two stages. In the first stage, k-NN is used to classify the testing data. During the second stage, the classifier ACO is used for the population to initialize and examine for the optimized result. The Streptococcus Pyogenic bacteria dataset used for this work which causes Rheumatic Fever, also called Acute Rheumatic Fever (ARF). The objective of this research is to maximize the level of performance and to minimize the error rates by classification techniques. The classification approach is integrate with optimization methods like ACO. The outcomes are compared with four dissimilar algorithms and the analysis established that the integrated framework shows accuracy of 70.26%.

In 2016, authors proposed an ensemble framework using Hierarchical Majority Voting (HMV) and multi-layer classification for the classification of disease and analysis using data mining approach [26]. They evaluated their framework on a hepatitis dataset, two diabetes datasets, two heart disease datasets, two breast cancer datasets, two liver disease datasets and one Parkinson's disease dataset, these datasets are gained from communal sources. The proposed framework is examined by evaluation of outcomes with some recognized classifiers along with hybrid mode. The experimental estimate displays that the planned framework handled all forms of attributes and attained maximum diagnosis accuracy. HMV is based on three modules. The first module is data acquisition and preprocessing which obtains data from several data sources and preprocess them. Each classifier's training is then performed on the training set in second module and then they are used to predict unknown class labels for test set instances. The prediction and evaluation is the third module of the proposed ensemble framework which comprises three classification layers and gains an accuracy of 97%.

In 2016, combination method which is helpful in seeking the best combination for heart disease analysis was proposed [27]. The analysis of different ensemble techniques, bagging, boosting and stacking with six different classification algorithms which are Bayesian Net (BN), Sequential Minimal Optimization for the SVM, Multi-Layer Perceptron (MLP), NB, and Decision tree classifiers. The method used in assembling is majority vote based and it is designed for every data set that belongs to the heart disease domain. The accuracy of ensemble model is 90%. Experimental observation shows that the best combination is when one of its classifier is a NB model with an accuracy of 92%.

In 2015, the researchers improved bagging technique integrated with the weighted voting scheme and represented a novel classifier ensemble for the analysis and examination of heart disease [28]. The approach used five heterogeneous classifiers named as NB, SVM, LR, instance based learner, QDA. Their 5 heart diseases dataset is acquired from UCI data repositions to do experiments and evaluate result. Various parameters are used to demonstrate the evaluation result like accuracy, sensitivity, specificity, p value, f measure, ANNOVA statistics and confusion matrix. The hybrid model is equated with the single classifier as well as with hybrid classifier the substantial outcomes are obtained, i.e., the overall accuracy of the proposed model is 84.16%.

In 2015, Extreme Learning Machine (ELM) is used to perfect attributes like age, sex, blood sugar, cholesterol, etc. The technique can substitute expensive medical check-ups with a cautionary message for the patient which shows the probability of heart disease. In this proposed model, there are five outcomes (0-4), which give fine results when equated with other models of Bidirectional Recurrent Neural Network (BPNN) or with ELM with single outcomes. With the ELM model, all the prior information is used to examine the new patient condition. It resolves the problems of information retrieval time and rendering it more appropriate for the application on big data. It further resolves the problem of missing attributes by examining these attributes on the overall decision. This technique is applied on real world data where approximately 300 patients' data have been collected by Cleveland clinic foundation [29]. The accuracy obtained is 80%.

In 2014, models like SVM, Naïve Bayes and DT-GI were applied to analyze the patient's heart disease from the dataset of 13 attributes [30]. Before the model construction, variations and missing values were also determined. Furthermore, using the majority vote technique, the heart disease prediction is figured for their ensemble approach. From the observations it comes out that accuracy of their hybrid approach is greater than that of other techniques. This approach can further be extended to classify the strength of heart disease. Results reveal show that the ensemble gave 82% accuracy for the UCI heart disease dataset.

In 2014, the researchers developed intelligent disease prediction classifier for the prediction of heart disease and analysis [31]. By combining five different machine learning classifiers such as memory-based learner, NB, SVM, Decision Tree Gini Index (DT-GI) and DT, an ensemble model discovers the heart disease. Five different sets of

attributes were used from five different data sets. Experiments show that the framework merges all types of attributes and predicts the heart disease with comparatively higher accuracy then the other techniques. They were assembled by a majority voting method for training and testing. Experiments showed that MV5 predicts with a high accuracy, i.e., 88.52%.

In 2013, Syed Umar et al. proposed a hybrid model which uses major risk factors, these are helpful for healthcare experts in treatment and they also give caution about the possible existence of heart disease even before the patient visits to the hospitals. The hybrid model involves two data mining tools, one is neural network and the other one is genetic algorithm. Using global optimization, genetic algorithm initializes the weight of a neural network. Adaptive power of this model is fast and accurate as compared to other model and the prediction accuracy is 89%.

In 2011, the development of the Heart Disease Prediction System (HDPS) that can be used by healthcare experts for calculating heart disease on the basis of health care data of the patient that makes estimation results using Artificial Neural Network (ANN) techniques [33], has accuracy of 80%.

In 2010, utilization of three classifiers named as NB, DT and classification by clustering, evaluated the diagnosis of heart disease and obtained similar accuracy before the pre-processing of the attributes [34]. The predictions for accuracy are 96.5 %, 99.2 % and 88.3 % by NB, DT, and classification by clustering, respectively.

Table 2.1 Literature Review

Work	Research Question	Contribution	Methodology	Technique	Accuracy
[25]	Rheumatic Heart Disease analysis	Predict the presence of Streptococcus Pyogenes with minimal fault rate	Hybrid approach	k-NN integrated with ACO	70.26%

[26]	Decision support framework for disease analysis	Overpowers the limitations of conventional performance	Assembling approach	Majority voting and multi-layer classification	97%
[27]	For heart disease analysis, ensemble combination model	Generalized model with two benchmarks	Hybrid approach	Bagging, Boosting, Stacking	90%
[28]	Heart disease analysis	Overcome the conventional performance	Hybrid approach to five heterogeneous classifiers	Bagging with multi-objective weighted voting scheme	84.16%
[29]	Heart disease diagnosis	Substitute expensive medical check-ups with caution message	Extreme Machine learning (ELM)	NN, pattern classification	80%
[30]	Decision Support Framework for Heart Disease	Intensity of heart disease can be identified	Hybrid approach	NB, DT-GI and SVM	82%
[31]	Majority vote-based scheme for heart disease analysis	MV5 framework, it can deal with all kinds of attribute	Hybrid approach	NB, DT-GI, DT-IG, memory-based learner and SVM	88.52%

[32]	Major risk factors for analysis of heart disease	At an early stage major risk factors are identified.	Hybrid approach	NN and GA	89%
[33]	HDPS is developed	GUI developed for HDPS	Machine learning	ANN algorithm	80%
[34]	Heart disease prediction	Reduces number of tests	GA	NB	96.5 %
				Classification by clustering	99.2%
				DT	88.3%

Chapter 3
Problem Statement

Based on machine learning technique, new rules and patterns are retrieved from huge amount of data. Data mining plays a crucial role in disease prediction. Diagnosis of a disease requires the performance of a number of tests on the patient. However, data mining techniques reduce the number of tests in healthcare. With reduced number of tests, the wastage of time and money reduces. Data mining of heart disease is important as it provides convenience to the doctors to let them know which attribute or feature is important, which will help the doctors in diagnosis of heart disease more efficiently.

There are numerous techniques available for data mining in the healthcare industry, however, the research that has to be done is on the performance of the various classification techniques, to enable the choice of the best among them can be chosen. The research presented in the thesis is intended to address the challenge of improving the prediction model for heart disease patients and provide timely response. A prime challenge facing health care administration is the provision of valuable facilities at a reasonable price. Valuable facility implies analysis of the patient's disease properly and directing effective treatments. Clinical choices are usually made subject to a specialist's nature and expertise, rather than by using knowledge gained from applying data mining technique in the domain database. This repetition prompts undesirable results, bias, mistakes and extreme costs, which influences the nature of facility given to the heart disease patient. Poor decisions regarding the patient can lead to undesirable and irreversible consequences. Medicinal facilities should likewise reduce the expenses of healthcare tests. A data mining model can accomplish these outcomes via utilizing real world data and provide accurate decision making.

The primary objective of current work is to identify useful patterns from the healthcare data to build an artificially intelligent model with the help of classifier model. Therefore, more relevant data attributes for the diagnosis of heart disease are observed. This aids the researchers to find out the root cause of disease in depth. Today, many hospitals maintain records of their patient's report. This kind of framework generates an enormous amount of information that can be visualized. However, this information can be used in

the decision making of the disease. As the result of open research issues analysed and discussed in Chapter 2, we were motivated to work in the said domain.

Briefly, the important research functions of the work are therefore stated as:

1. How various data mining techniques can be used in the health care industry and how to identify their performance in prediction?
2. How do classification techniques help in developing the prediction model so as to predict accurately the risk of heart disease among patients?
3. How can data be turned into useful information that helps doctors to make intelligent decisions?

Chapter 4
Methodology

The research methodology followed for the present work is discussed in detail in this Chapter.

4.1 Data collection

Heart Disease dataset is appropriate to extract necessary features in order to form an intelligent decision support system. Generally, medical data is not easily released by many healthcare centres due to their privacy constraint. Therefore, this research work uses the standard heart disease dataset from data mining repository dataset of University of California, Irvine (UCI). This UCI dataset is composed of four databases, namely Cleveland dataset, Hungarian dataset, long-beach-via, and Switzerland. The Heart Disease Dataset can be freely downloaded from the available machine learning UCI repository. The dataset has 1370 instances and 76 attributes; however, only 14 of them have been used so that we obtain accurate results with reduced number of attributes [35].

The major benefit of using this dataset is that it can be easily communal with other makers or healthcare experts. The advice gained from other makers can be helpful in improving the result of the proposed framework. Table 4.1 describes the chosen attributes and their possible data types or values in heart disease dataset.

Table 4.1 Attribute Description

Attribute	Description	Possible values
Age	In years	Continuous
Sex	Gender	1 = male 2 = female
Cp	Chest Pain type	1 = typical angina 2 = atypical angina 3 = non-angina pain 4 = asymptomatic
trestbps	Resting blood pressure (in mm Hg)	Continuous
Chol	Serum cholesterol (in mg/dl)	Continuous

Fbs	Fasting blood sugar	1 = true 2 = false
restecg	Resting electrocardiographic	0 = normal 1 = having ST-T wave abnormality 2 = left ventricular hypertrophy
thalach	Maximum heart rate	>0
exang	Exercise induced angina	1 = yes 0 = no
oldpeak	ST depression induced by exercise relative to rest	Continuous
Slope	Slope of the peak exercise ST segment	1 = upsloping 2 = flat 3 = downsloping
Ca	Number of major vessels colored by flurosopy	0-3
Thal	maximum heart rate achieved	3 = normal 6 = fixed defect 7 = reversible defect

4.2 Data pre-processing

The collected data is commonly weakly organized and there are missing values, out-of-range values etc. The dataset which does not encounter these kinds of problems may give false results. Therefore, pre-processing of the dataset needs to be carried out as a prerequisite activity for the study. If there is large amount of incorrect or noisy data, then extracting the knowledge from the dataset for the training data is further hard. Data pre-processing is performed to discover if there is unnecessary and terminated content present in the data. For pre-processing of the data there are several steps, such as cleaning, normalization, transformation, etc. The result of the data pre-processing is the vital dataset with minimal attributes.

The presence of redundant record is the main drawback with the healthcare dataset. The existence of redundant instances in the dataset can be a reason for unfair learning, biased towards frequent data and unfair towards infrequent data. For obtaining better accuracy,

the redundant records must be removed from the data. Thousands of records in the healthcare data are combined into one relational database table and the records which fail to match with the layout can be deleted from the record set. From the average value, if any instance or attribute comes out of this range then those instances are deleted from the dataset. Irrelevant and missing data are also deleted from the dataset. Throughout pre-processing process, the whole dataset is taken as an input and various data pre-processing approaches are applied for reducing the invalid instances from the dataset.

Pre-processing procedure is carried out as follows:

1. Thousands of records from the healthcare data are merged into one relational database table and the records which fail to match the layout were be deleted from the record set.
2. The records which go out of range of the average value were also deleted.
3. The record with missing value is replaced with average of the whole column.
4. Removal of instance that have more than one value, i.e., removal of multivalued attributes.

4.3 Machine Learning

Machine learning is the artificial intelligent model which provides various algorithms to make computer more intelligent. To make the computer to be an intelligent machine we have to make it learn from the past. Machine learning is a famous technique used in data mining, which discovers the non-trivial, unknown values and possibly useful information embedded in databases. Machine learning helps the computer system acquire knowledge. It provides various approaches for collecting, altering and modernizing information in the systems which are used in the time of decision making [36]. From past case examples and historical datasets, machine learning helps the machine to be intelligent through various learning algorithms. There are two types of machine learning, i.e., supervised learning and unsupervised learning.

4.3.1 Supervised Learning

The concept of supervised learning comes from the supervisor, acting as a teacher in the learning process [37]. The teacher teaches the topic with the good example so that student can remember, and then student derives the general rules from the specific

example. The training examples are in the form of pairs that consist of input value x and a desired output value y. The job of supervised learning algorithms is analysing the training data and producing a function. The training dataset consists of training samples, it concludes a function. In case of supervised learning, every sample consists of training value and testing value [38]. The supervised learning algorithm analyses, the training data and derives an activation function, which is used for mapping new sample. An optimal state will permit for the model to determine the class labels correctly. Supervised learning produces the function which maps the input to the desired output [39].

As stated earlier, in supervised learning, function maps the input to the desired output for defining about the new input to which class it belongs to.

Input data $$X = (x_1, x_2, \ldots\ldots, x_n) \quad (4.1)$$

Output data $$Y = (y_1, y_2, \ldots\ldots, y_n) \quad (4.2)$$

Hypothesis about the function $$h: X \rightarrow Y \quad (4.3)$$

h is the function of vector-valued input which is selected on the basis of training set of n input vector examples, i.e.,

$$X = (x_1, x_2, \ldots\ldots, x_n) \rightarrow h \rightarrow h(X) \quad (4.4)$$

$$\text{Training set} = \{X_1, X_2, \ldots\ldots, X_n\} \quad (4.5)$$

Therefore, the predicted value can be given as:

$$y = h(x) = \operatorname{argmax} f(x, y) \quad (4.6)$$

For solving any problem the supervised machine learning algorithm follows a number of steps, as given below:

1. The first and foremost step is the collection of the data required for solving a particular problem. It consists of identifying all the important features or attributes that are most relevant to the problem under study.
2. The second step is the pre-processing of data which is not suitable for training so that missing values or noise from the data can be removed. There are various methods for pre-processing of the data, but method used for the data depends upon the state. There are also some methods for detecting and handling noise.

3. The third step is feature subset selection. It consists of recognizing and eliminating the features that are redundant or that are not relevant for the problem. It improves the efficiency of the learning algorithms by decreasing the dimensionality of the data. In order to develop more accurate and efficient classifiers by removing redundancy, a process called feature construction is used. In this process new features are constructed from the existing basic features in situations where many features depend on one another.
4. The fourth step is evaluating the accuracy of the classifier. This step decides whether the classifier is fit to be used or some modifications are required. The evaluation of the classifier depends on the prediction accuracy. The classifier's accuracy can be estimated by splitting the data into two- training data and data for calculating accuracy (testing data).
5. If the error rate evaluation shows that the classifier is not efficient enough or is unacceptable, then the algorithm returns to previous stage and some factors are examined again. For example, features are checked again to eliminate irrelevant features, or the size of training set is checked again. Some other problems that might occur include too high dimensionality of the problem or class imbalanced data. However, if the evaluation shows satisfactory results, then the classifier is available for use.

Among many other learning examples, classification and regression are two important supervised learning problems. As discussed earlier, the training examples are in the form of pairs that consist of input x and output value y. The supervised learning analyses the training data and produces a function.

This function can take two forms, i.e., if the output value is continuous then it is regression function and if the output value is discrete then it is a classifier. The system is provided with labelled instances represented as (x, y) and the objective of supervised learning systems is to determine the label y for each new input x that it sees in future. When the value of y is real, then it is regression and when the value of y is discrete then it is classification.

4.3.1.1 Classification

In machine learning, we can define classification as the task of determining to which class among a new input belongs to. Training data will help in classifying the class because it contains the instances whose class is known. In classification, there are a number of classes and the goal is to develop a rule that classifies a new input into one of the existing classes. The algorithm that is used for classification is called a classifier. Classification algorithm implements the function that maps the input data to the desired class. There are certain issues which must be taken care of while developing a classifier, such as accuracy, speed, comprehensibility, and time to learn a classification rule. Classification can be either binary classification or multiclass classification.

- **Binary classification:** It consists of only two classes. For example, the classification of customers in the bank loan application. In this example, the input to the classifier is the information about the customer and the classifier function assigns the desired class, i.e., low-risk and high-risk customers. The customer information may include his salary, investments, age, and profession and so on. In this example, a classification rule learned is of if-then type, i.e., if the customer income is greater than a threshold and his savings are greater than a threshold then the customer can be classified into low-risk class else the customer will be classified into high-risk class. Such an example is called a discriminant function which separates the examples of different classes.
- **Multiclass classification:** Here, an object can be assigned to any one of a number of classes. There are a number of classification algorithms that have been developed. These include Logistic regression, k-Nearest Neighbour, NB, Random Forests, SVM, NN, Least Squares Support Vector Machines, DT, Bayesian Networks, etc.

4.3.1.2 Regression

Regression can be defined as a technique that is used for calculating the relationships between dependent variables and one or more independent variables [40]. It predicts continuous variable's value and is used for numeric prediction. In other words, we can say that the process of regression depicts the changes in the values of a dependent variable by varying the value of one of the independent variables while the other independent variables are kept fixed. In machine learning, regression can be defined as a technique that is used to fit an equation to a dataset. The types of regression techniques are:

- **Linear regression:** Here, the formula of straight line is used, i.e.,

$$y = mx + b \tag{4.7}$$

  and the suitable values for m and b are estimated in order to predict the value of y on the basis of a given value of x.
- **Multiple regression:** In this technique more than one input variable is used that fits more complex models, such as a quadratic equation.

Applications of regression are prediction and forecasting. There are a number of techniques for using regression. Least squares regression and linear regression are parametric methods. It means the function is described in form of a finite number of unknown parameters that are estimated from the data. Another form of regression is nonparametric regression in which the regression function is allowed to lie in a specified set of functions, which may have infinite dimensions. In order to explain the regression technique, we can take the example of a system that should be able to predict the car price. Inputs to the system are the car attributes such as capacity of engine, mileage, brand, and so on which show the worth of the car and the car price is the output. These kinds of problems where the output is in the form of number are known as regression problem.

4.3.2 Unsupervised Learning

In unsupervised learning approach, the idea is to find the hidden patterns from unlabelled data. For machine learning task, if training data has only set of inputs and not their classes then it is known as unsupervised learning, which will help in finding the relations among various input data. Clustering is a major form of unsupervised learning. In Data Mining, clustering is the procedure for grouping the data points in such a way that points inside a cluster have alike features, whereas, points in different clusters are different. In traditional clustering procedures, the usage of distances among points for clustering is not suitable for categorical and Boolean attributes. Further, clustering can be a partitioned or hierarchical algorithm. By the iterative relocation principle, it divides the document into a specific number of groups. Examples of partitioning algorithm are: k-medoids, k-means and bisecting k-means. Hierarchical clustering begins through every document in the document space as individual clusters and iteratively combines the best alike clusters. It creates the predetermined clusters from top to bottom.

Chapter 5
Proposed Work

Inspired by the growing rate of patients' death owing to heart disease each year, there is an increasing availability of patients' data which can help experts to extract important information by data mining techniques. This important information can help human experts to cure the heart disease [41], [42]. Moreover, it can facilitate the design of a hybrid model that can help the hospital management to encourage and give advice to the experts related to the diagnosis and proper treatment given to the patients having heart disease. This work describes some standard classifiers for disease risk estimation. These classifiers have further been analysed and compared with the proposed hybrid approach - Hybrid Classifier with Weighted Voting, referred to as HCWV, here. The HCWV includes nine classifiers which are employed to generate an ensemble. This provides better results than any of the classifiers, when applied in isolation. These algorithms which have been compared have been named earlier in the methodology (Chapter 4). The proposed framework of this study is given in Figure 5.1.

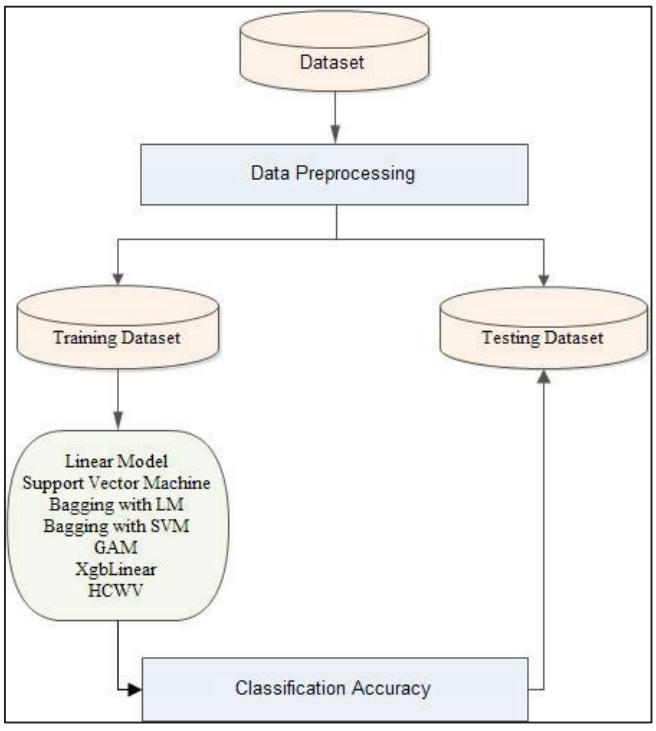


Figure 5.1 Flowchart of HCWV

5.1 Classification methods used

The following subsections summarize various methodologies used in the flow chart in Figure 5.1 which are used in the diagnosis of heart disease.

5.1.1 Generalized Linear Model

Generalized Linear Model (GLM) is the part of statistical theory. The linear models are the basis of a wide variety of statistical procedures. The model takes the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ \ldots\ldots\ldots\ldots\ldots + \beta_p x_p + \varepsilon \qquad (5.1)$$

where ε is normally distributed. The first part simplifies the y portion; the next the ε portion; and the third, the x portion of the linear model. Since the linear model cannot hold non-normal response, y, like proportions or counts. This inspires the expansion of GLM that can signify binary, categorical, and other forms of replies. In GLM, every result Y of the reliant on variables is expected to be made from a specific distribution in the exponential family like normal, binomial, Poisson and gamma distribution. The mean, μ, of the distribution be contingent on the independent variables, X, through:

$$E(Y) = \mu = g^{-1}(X\beta) \qquad (5.2)$$

Where,

E(Y): probable outcome that Y can take

(Xβ): linear predictor

g: link function

β: linear grouping of unidentified parameters.

The starting point for GLMs is the familiar GLM; the most widely taught and used method of data analysis in psychology and the behavioural science today. The GLM comprises both, analysis of covariance, analysis of variance and multiple regression. Analysis of variance and multiple regressions allow researchers to study the relationships among one or more independent variables and one continuous dependent variable. In multiple regression, the independent variables may be continuous or categorical; in analysis of variance, the independent variables are categorical, so that it can be measured as a special case of multiple regression. The form of relationship between each independent variable and the dependent variable can be linear and

curvilinear. This relationship can be general or conditional, potentially involving interactions between two or more independent variables.

5.1.2 Support Vector Machine

Support Vector Machine (SVM) is a statistical technique which was developed by Vapnick in 1982. It gives effective classification outcomes in several fields like face recognition, text categorization, bioinformatics and medical diagnosis [43].

In SVM, classification predictor is generated for each test set as input and corresponding output is produced which takes values of the two available classes thus creating binary classifier which is non-probabilistic. It gives the demonstration of the illustrations mapped in a manner that the illustrations of distinct classes are bifurcated by a gap which is clear and as separated as possible. New illustrations are then plotted into the similar space and expected to fit into a category on the basis that on which side of the predesigned gap they fall. This technique hence constructs a linear maximum margin hyper plane in a space with infinite dimensionality, which can further be used for regression, classification and other tasks. It is defined by w weight vector and b bias which is hyper plane distance from centre. Using kernel function, the separation of non-linear dataset is performed. The linear margin hyper-plane so found is maximum because a decent bifurcation is attained by the hyper-plane having the greatest distance to the closest training point belonging to several classes, as generally, the greater the margin, the lesser is the classifier error rate of the classifier. Let us take the scenario, in Figure 5.2 there are three hyper-planes A, B and C and all are separating the classes well. To recognize the correct hyper plane the right thumb rule is used: 'Select the hyper-plane which segregates the two classes better'. In the given Figure 5.2, hyper-plane C has brilliantly done this work.

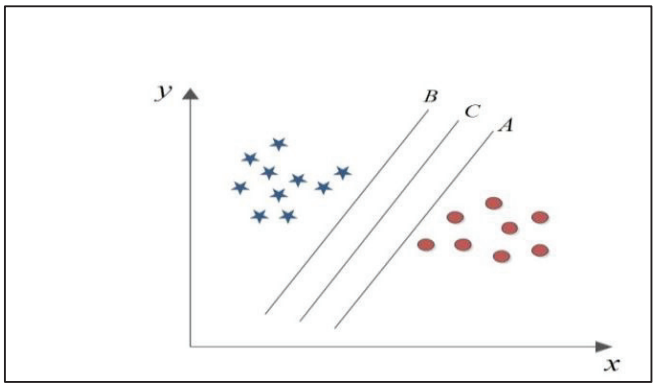


Figure 5.2 Hyperplane in SVM

The primary benefit of SVM is its maximum classification accuracy. It is utilized for pattern recognition and is fundamentally designed for the two-class classification. Actually, it works well with pure margin of parting and is operative in high dimensional spaces. It is not suitable for large datasets because required training time is high. It does not perform well on noisy datasets, i.e., where target class is overlapping.

5.1.3 Bagging algorithm

Bagging which stands for the process popularly known as Bootstrap Aggregation. For generating training datasets, bagging uses bootstrap and the learners are trained using an weak learning procedure, and weighted voting is taken during testing. This method works effectively if the base learner is weak. Bagging is a hybrid method which uses a training dataset to train the base classifiers on training set's random rearrangement. We randomly draw each training set of classifier using replacement technique. The results of each base classifier are assembled in a sample. In the bagging algorithm, the additional information is made available for the training data from the original dataset with replacement of the data to produce the multi-sets of similar sizes as in the original data. The framework of bagging is shown in Figure 5.3. These decrease the variance of predictive value and hence reduce dispersion. Through single classification rule, several pattern recognition problems cannot be answered. This occurs when there is complex data distribution, high dimensional data, and training data size is small. In such case, ensemble the classifiers for increasing the performance measure of single classification rule. In our work, bagging has been combined with the models, namely, GLM and SVM to increase the accuracy of the models already calculated under the bag free models category.

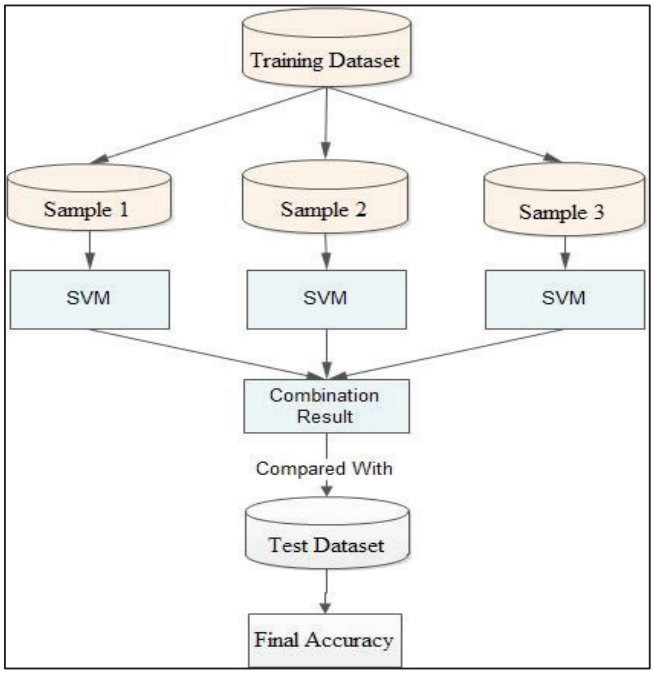


Figure 5.3 Bagging Framework

5.1.4 Boosting

Boosting is the method which combines the weak learners incrementally such that emphasis is set on the training data that were wrongly classified earlier and trains the new data. In this technique, the model is trained on the errors of the earlier learner. A learner is assumed to be weak if it derives models that achieve somewhat better than random guessing. In a weak learner, the error probability is ½. The basic notion behind boosting is that at every step, more stress is laid on the instances that were misclassified in the previous step. The intensity of this stress to be laid on an incorrectly classified instance is quantified by a weight linked to every instance in the next step. The main aim of boosting is to convert a weak learner into a strong learner. Figure 5.4 shows the procedure of boosting. In this work, for Boosting XGBLinear (eXtreme Gradient Boosting) and GAMBoost (Generalized Additive Models) are used. Xgboost() is eXtreme Gradient Boosting package. The package Xgboost() contains tree learning algorithm and linear model solver. It supports several functions like ranking, regression and classification. Xgboost() is expandable so that the users can effortlessly defined their objectives. It can do parallel computation automatically. GAM boost is Generalized Additive Model. The use of GAM boost is in statistics for the analysis of data which suffers from constraints to some exploratory variables and selection of smoothing parameter problem. The problems can be avoided by means of stepwise fitting of weak

learners and then boost them using GAM. The resultant of fitting process works for every experimental family including Poisson, binomial and normal response variables. The process merges the selected variables and resolves suitable quantity of smoothing.

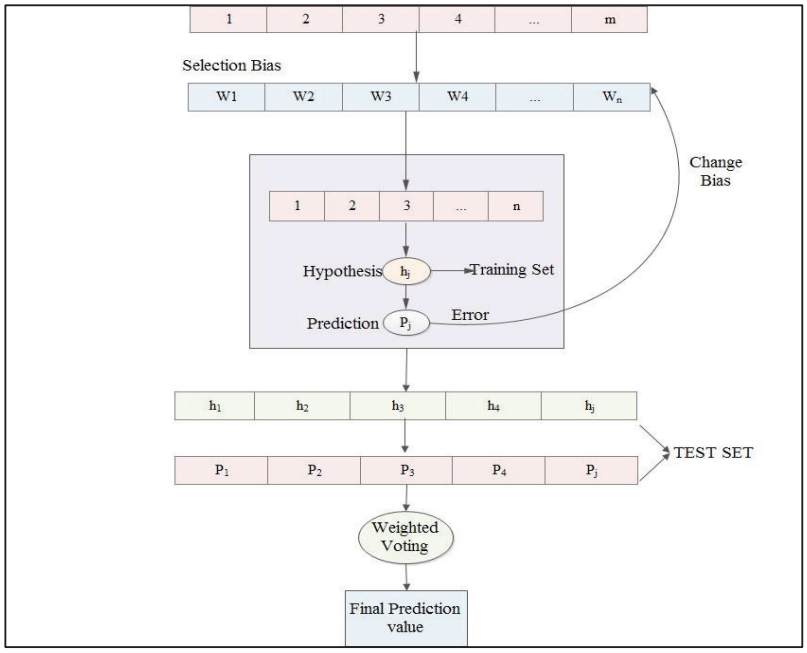


Figure 5.4 Boosting Framework

5.2 Hybrid Classifier with Weighted Voting

In Hybrid Classifier with Weighted Voting (HCWV), Machine Learning procedures are used to retrieve the hidden knowledge that can be used for good decision-making. Machine learning contains various techniques like decision tree, case based reasoning, artificial neural network, and rule based learning. Each technique has its own benefits and weaknesses. In the past many of hybrid machine learning systems were developed to bring the best from the two different machine learning methods. The proposed approach makes use 9 common classifiers, namely, SVM, Neural network, Decision tree, GLM, Lasso, Bayesian regularized Neural Network (brnn), Classification and Regression Tree (rpart2), Multivariate Adaptive Regression Spline (earth), Conditional Inference Trees (CTree) to be trained on the Training dataset and their accuracy is calculated by testing each model on the Test dataset. The models so generated are then arranged according to their accuracy in ascending order.

The models with the least accuracy are combined for the ensemble to give the Combination Result 1 which has better accuracy than the models so used on the basis of

Weighted Voting. Similarly, the Combination Result 2 and Combination Result 3 are found. These Combinations are again ensemble to predict the final output which is hence the output of HCWV. The output of HCWV is tested against the Test Dataset and has the accuracy greater than the accuracy of all the models used for its construction as shown in Figure 5.5. Through this, hybrid machine learning system is used to improve the classification accuracy of the proposed hybrid system that may be used for medical diagnosis.

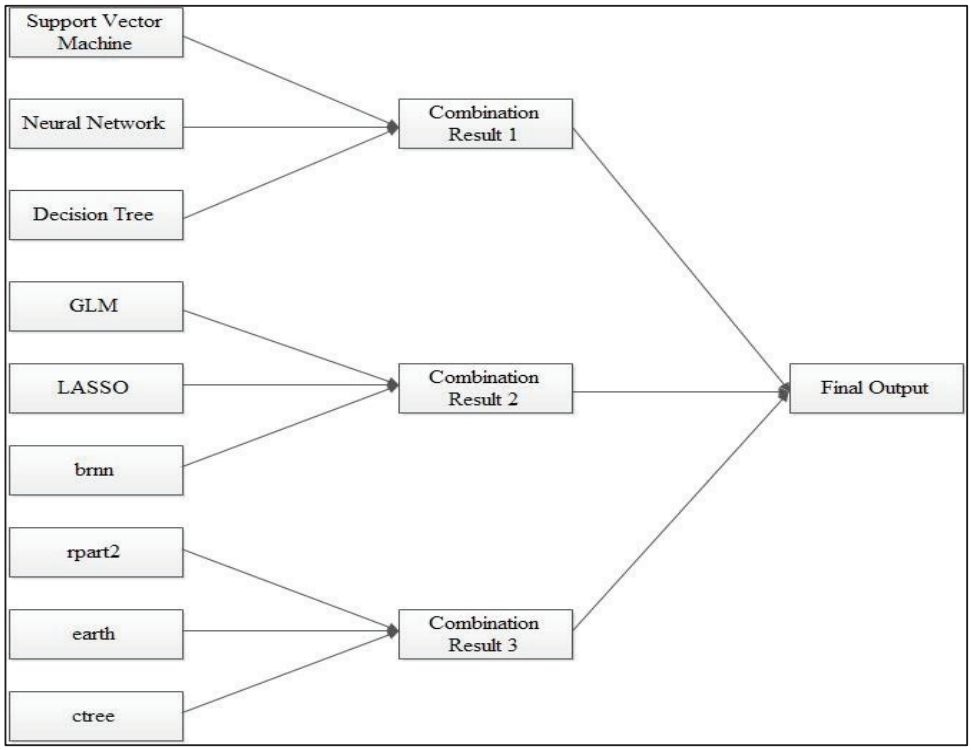


Figure 5.5 Proposed Framework for HCWV

The models used for ensemble are described below:

5.2.1 Level 1 classification

5.2.1.1 Support Vector Machine:

As explained earlier in Section 5.1.2.

5.2.1.2 Neural Network:

The Neural Network (NN) simulates a lot of heavily interconnected brain cells inside a computer in order to recognize patterns, and make decisions in a humanlike way. A

major benefit of NN is that you do not have to program it to learn explicitly, it learns all by itself, similar to our brain!

Neural network are systematized as layers, which are built from numeral interconnected nodes which hold the activation function. Through input layer patterns that are presented to the network, patterns communicate to a single or multiple hidden layers where the real execution is carried out. The hidden layers then connect to an 'output layer' where the response of the output is as shown in Figure 5.6.

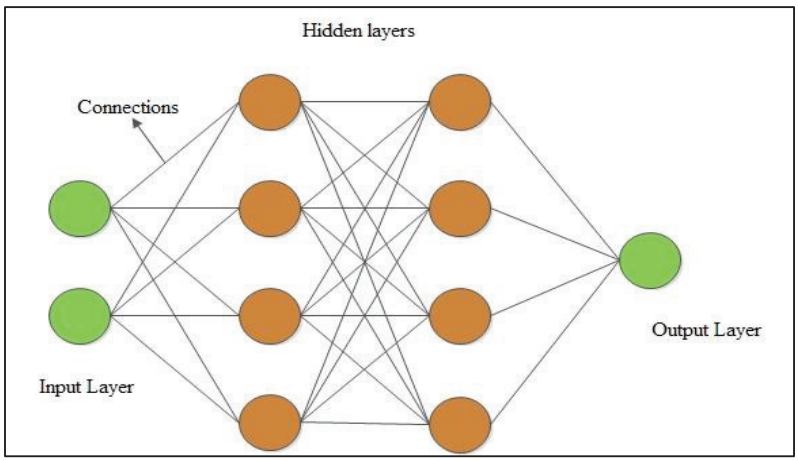


Figure 5.6 Neural Network

5.1.1.3 Decision Tree:

Initially decision tree was developed for statistics to determine the useful areas in the database. The whole procedure of hypothesis generation is automated by decision tree algorithm in a more integrated way as compared to other data mining approaches. The score of decision tree is so high on various significant features of data mining; they can also be helpful in solving the wide variety of problems for both prediction and exploration, without much pre-processing.

Decision tree is a flowchart like structure, where the testing attributes are represented by internal nodes and classes are represented by the leaf nodes. The highest node is root. The classification is given by the leaf node that relates to every case that reach the external node, or a probability distribution or classification sets over every probable classification. To classify unseen instances, according to the values of the attributes it is directed down the tree and checked in successive nodes, and when a leaf is obtained the instance is classified according to the class of the leaf.

In decision tree construction process, the important thing which is to be considered is the selection of attribute used for splitting the example set in to different classes. The attribute is selected at each non-leaf node built on the information gain and the attribute which produced its highest value is selected for splitting. The information gain shows the probable quantity of information which is necessary to specify about the new instance, to which class it belongs. For example, in the decision tree in Figure 5.7 the Location category is further sub-divided into three categories as Suburbs, Rural and City. The suburbs subcategory is divided into high and low on the basis of household income where the high income means no show is being attended and the low income means the show is being attended. The rural subcategory directly means the show being attended. The City subcategory is divided on the basis of registered voting where if the citizen is a registered voter he will not attend the show else if he is a registered voter, he will attend the show.

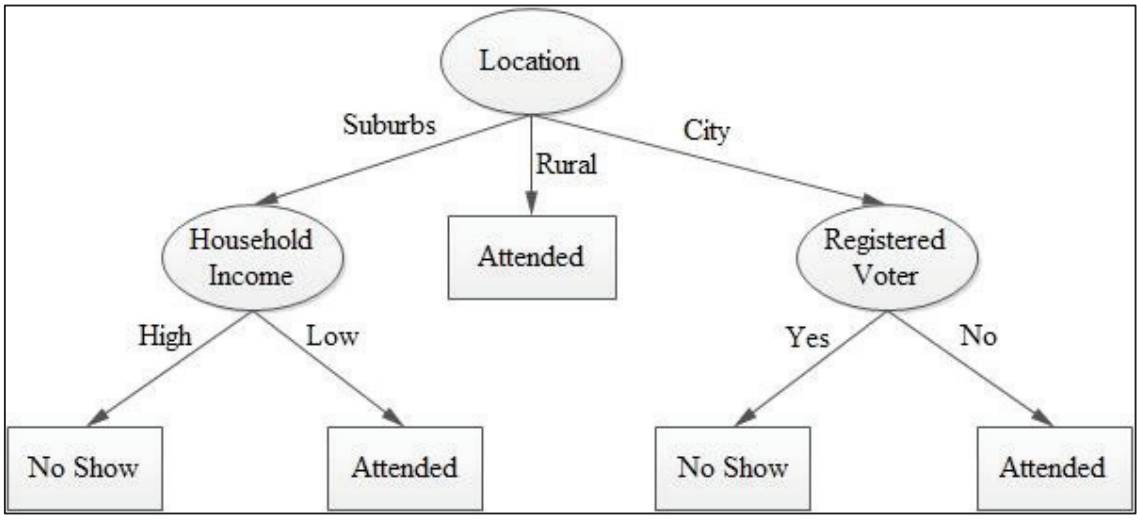


Figure 5.7 Description of Decision Tree

5.2.2 Level 2 classification

5.2.2.1 Generalized Linear Model

As explained earlier in the Section 5.1.1

5.2.2.2 Least Absolute Shrinkage and Selection Operator

Least Absolute Shrinkage and Selection Operator (LASSO) is a statistical analysis technique in which variable selection and regularization is performed to improve the

accuracy and interpretability of the statistical model. LASSO was initially formed for Least Square Model and this simple model discloses the significant quantity of performance of the estimator, with its connection to ridge regression and superlative subset variety and the links among coefficient estimates which is known as threshold. It also discloses that, if covariates are collinear then the estimation of coefficient is not essentially unique. However, after initial formulation of least squares, regularization of LASSO can be effortlessly extended to huge variations in the statistical models including proportional hazards models, GLMs, M-estimators and generalized estimating equations.

5.2.2.3 Bidirectional Recurrent Neural Networks

Bidirectional Recurrent Neural Network (BRNN) was developed to maximize the quantity of input knowledge existing on the network. Time Delay Neural Network (TDNN) and Multilayer Perceptron have limits on the input data suppleness, they need fixed input data. Similarly, Recurrent Neural Network (RNN) also has some constraints as the future input information cannot be obtained from the current state. The elementary idea behind of BRNNs is to link the two hidden layers with different directions to have similar output. Through this approach, the output layer can acquire information from the past and future states. When the context of input is needed, BRNN is used. For example, in handwriting recognition, the performance can be enhanced by knowledge of the letters located before and after the current letter.

5.2.3 Level 3 classification

5.2.3.1 Classification and Regression Trees

Classification and Regression Tress (rpart2) is classification type model. The required package for this model is rpart. The turning parameters are maxdepth (Max Tree Depth).

5.2.3.2 Multivariate Adaptive Regression Spline

Multivariate adaptive regression splines, implemented by the Earth class, are a flexible regression method that automatically searches for interactions and non-linear relationships. Earth models can be dimensional space. Each term in an Earth model is a product of so called ‘hinge functions’. A hinge function is a function that is equal to its argument where that argument is greater than zero and is zero everywhere else, as shown in Figure 5.8.

$$h(a-x) = [a-x] = \begin{cases} a-t, & a > x \\ 0, & a < x \end{cases} \quad (5.3)$$

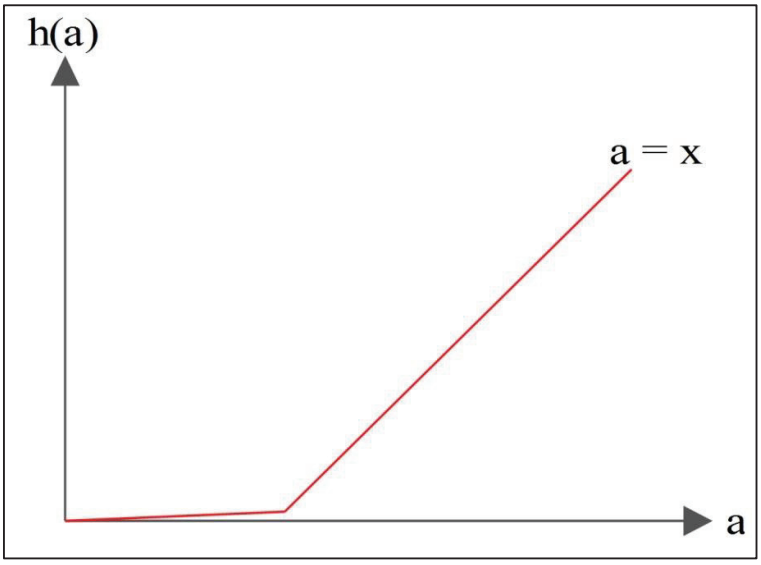


Figure 5.8 Basic Hinge Function

5.2.3.3 Conditional Inference Trees

Conditional Inference Trees (CTree) is the toolbox for partitioning the computational data recursively. In the package ctree(), conditional inference procedure implements the embedded tree-structured regression model into a definite concept of conditional inference procedures. This regression tree is of non-parametric class which can be appropriate for all types of regression problems which includes numeric, nominal, arbitrary and multivariate response variable types. cforest() delivers the execution of Breiman's random forests based on Conditional Inference Trees. Based on the parametric models like GLM, Linear models etc. the mob() function executes the algorithm for recursive portioning.

Chapter 6
Results

In order to check which model suits better for a problem, it is necessary to compare the performance of various models with each other. In data mining, it is a general condition and in this work Confusion matrix is used for this purpose. The confusion matrix here is used for several statistical events which are examined and whose implications are drawn. Here additional description is divided into two major portions; firstly the analysis part and theoretical background of the confusion matrix. The second part describes the uses to this problem. It is nearly a mutual condition that performance of several models has to be equated with each other to recognize the correctness of a model to a particular problem. This common condition exists in the data mining also. In this work accuracy, sensitivity and specificity is used for this purpose. In supervised learning, confusion matrix tool is used for analysing the performance. It is used to signify the testing outcomes in a predicted class. Each row signifies the actual class instances, each column signifies the predicted class instances.

Table 6.1 Confusion Matrix

Predicted class	Actual class		
		Positive	Negative
	Positive'	TP	FN
	Negative'	FP	TN

Where Positive and Negative values are Actual values and Positive' and Negative' values are Predicted values.

- TP (True Positive) defines to the number of samples or instances which actually belong to class 0 and also have been correctly classified to class 0.
- TN (True Negative) defines to the number of samples or instances which actually belong to class 1 and also have been correctly classified to class 1 itself.
- FN (False Negative) defines to the number of samples or instances which actually belong to class 0 but have been wrongly classified to class 1.

- FP (False Positive) defines to the number of samples or instances which actually belong to class 1 but have been wrongly classified to class 0.

a) **Accuracy:** One way of arbitrating the performance of a classifier is to relate the accuracy of every classification.

$$\text{Accuracy} = \frac{(\text{TP}) + (\text{TN})}{\text{Total number of Samples or instances}}$$

Where,

TP : Total number of correct positive classifiers

TN : Total number of correct negative classifiers

b) **Recall or sensitivity** : It is defined as the number of true positives (TP) over the number of true positives and the number of false negatives(FN).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

If a person has a disease, how often will the test be positive (true positive rate).

c) **Specificity:** It is defined as the ratio of number of true negative (TN) to the sum of number of true negatives and number of false positives(FP).

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

If a person does not have the disease how often will the test be negative (true negative rate).

Testing was conducted separately for each of the classification models and performance parameters were computed separately. The same training and test sets were used for all the models. This study has made the comparison between two normal classifiers which are Generalised Linear Model (GLM) and Support Vector Machine (SVM), two models which incorporated bagging using GLM and SVM, two models with the use of the boosting approach namely XGBLinear and Generalised Adaptive Model (GAM) and the proposed hybrid approach HPWV on the basis of accuracy for heart disease dataset as depicted in Table 6.2. Accuracy is the measure of how well the classifier is working in

predicting the target value of the instance in the test dataset as compared to its actual value. Higher the accuracy, better the models is and in this case the HCWV has shown the highest accuracy.

Table 6.2 Comparison of HCWV accuracy with other classifiers

Classifier	**Accuracy**
Support Vector Machine	61.45%
Generalised Linear Model	67.64%
XGBLinear	76.27%
Generalised Adaptive Model	79.36%
Bagging with LM	81.67%
Bagging with SVM	81.89%
HCWV	**82.54%**

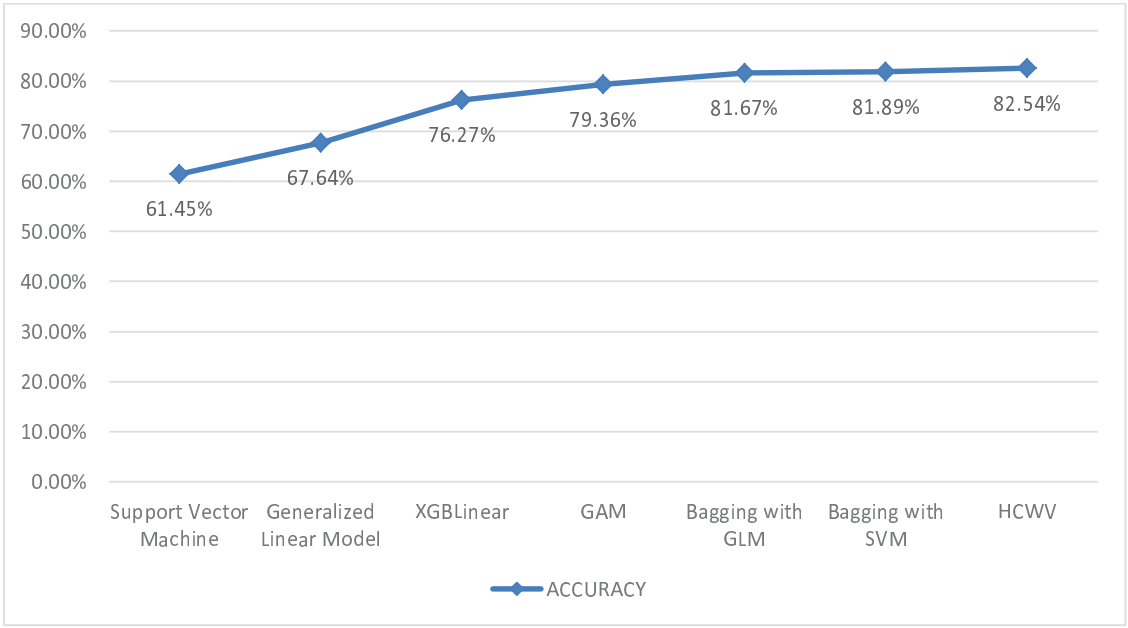


Figure 6.1 Accuracy of classifiers

Table 6.3 Comparison of HCWV Sensitivity with other classifiers

Classifier	**Sensitivity**
Support Vector Machine	0.89
Generalized Linear Model	0.82
XGBLinear	0.87
Generalised Adaptive Model	0.88
Bagging with GLM	0.83
Bagging with SVM	0.88
HCWV	**0.97**

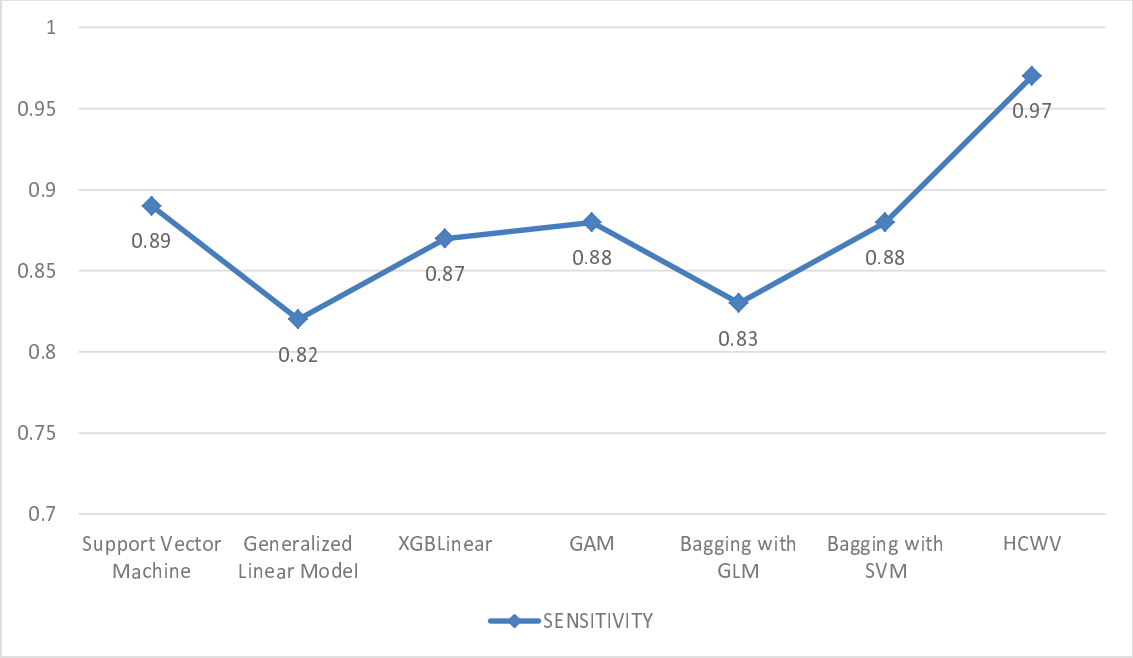


Figure 6.2 Sensitivity of classifiers

Table 6.4 Comparison of HCWV Specificity with other classifiers

Classifier	Specificity
Support Vector Machine	0.61
Generalized Linear Model	0.81
XGBLinear	0.61
Generalised Adaptive Model	0.65
Bagging with GLM	0.78
Bagging with SVM	0.80
HCWV	**0.82**

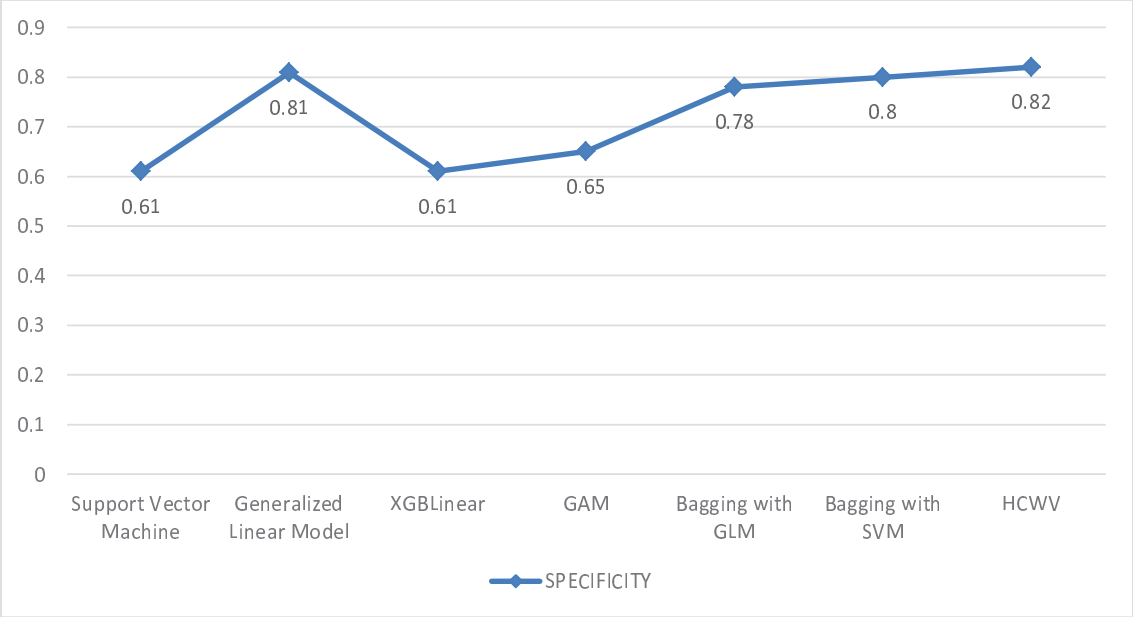


Figure 6.3 Specificity of classifiers

Chapter 7
Conclusion and Future Scope

This Chapter summarizes the research contributions of the thesis and gives future research directions.

7.1 Conclusion

Inspired by the overall expanding mortality of patient having heart disease every year, scientists are utilizing data mining strategies in the analysis of heart diseases. Even though data mining strategies to support healthcare experts in the analysis of heart disease is having certain achievement, yet the utilization of data mining systems is to recognize an appropriate treatment for patients with heart disease has received little consideration. Precise prediction safeguards correct healthcare services, likewise, using hybrid data mining methods demonstrates favourable outcomes in finding heart disease. To this end, in the current work we have successfully achieved higher accuracy by using hybrid data mining for heart disease analysis. The estimates made by the hybrid framework will be helpful in healthcare in making reliable diagnosis.

7.2 Future Scope

- In the future, automated heart disease prediction system can be implemented in the remote areas like in rural areas to replicate the human expert for the diagnosis. The prediction system is appropriate for supporting healthcare experts for in the diagnosis of heart disease.
- The attributes of the disease should be superior in obtaining the more accurate result.
- With the use of these attributes, the work can be extended to find other types of disease such as cancer, arthritis but early detection of other chronic disease should also be taken care of.
- In the automation of heart disease prediction, the work can be extended and improved by building a GUI, website or mobile application.
- The collection of actual data from the hospitals, healthcare organizations and healthcare agencies can be done in order to equate the ideal accuracy with other techniques in data mining.
- Fuzzy learning can be implemented to predict the strength of heart disease.

References

[1] "Cardiovascular diseases (CVDs)" World Health Organization. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs317/en/. [Accessed: 20-April-2017].

[2] Amin, S.U., Agarwal, K. and Beg, R. "Genetic neural network based data mining in prediction of heart disease using risk factors", In Information & Communication Technologies (ICT), IEEE, pp. 1227-1231, 2013.

[3] Baati, K., Hamdani, T.M. and Alimi, A.M. "A modified hybrid naïve possibilistic Classifier for heart disease detection from heterogeneous medical data", In Soft Computing and Pattern Recognition (SoCPaR), IEEE, pp. 353-358, 2014.

[4] Dewan, A. and Sharma, M. "Prediction of heart disease using a hybrid technique in data mining classification", In Computing for Sustainable Global Development (INDIACom), IEEE, pp. 704-706, 2015.

[5] Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J. "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann, 2016.

[6] Canlas, R.D. "Data mining in healthcare: Current applications and issues", School of Information Systems & Management, Carnegie Mellon University, Australia, 2009.

[7] Alizadehsani, R., Habibi, J. and et.al. "A data mining approach for diagnosis of coronary artery disease", Computer methods and programs in biomedicine, pp.52-61, 2013.

[8] Thuraisingham, B. "A primer for understanding and applying data mining", It Professional, pp.28-31, 2000.

[9] Giudici, P. "Applied Data Mining: Statistical Methods for Business and Industry", John Wiley & Sons, 2011.

[10] "What Is Atherosclerosis?", National Heart Lung and Blood Institute [Online] Available: https://www.nhlbi.nih.gov/health/health-topics/topics/atherosclerosis [Accessed: 26-Febrary-2017].

[11] "What Is Coronary Heart Disease?", National Heart Lung and Blood Institute [Online] Available: https://www.nhlbi.nih.gov/health/health-topics/topics/cad [Accessed: 2-March-2017].

[12] "What Is Carotid Artery Disease?", National Heart Lung and Blood Institute [Online] Available: https://www.nhlbi.nih.gov/health/health-topics/topics/catd [Accessed: 15-March-2017].

[13] "What Is Peripheral Artery Disease?", National Heart Lung and Blood Institute [Online] Available: https://www.nhlbi.nih.gov/health/health-topics/topics/pad. [Accessed: 26-April-2017].

[14] "About Chronic Kidney Disease", The National Kidney Foundation [Online] Available: https://www.kidney.org/atoz/content/about-chronic-kidney-disease. [Accessed: 9-May-2017].

[15] Srinivas, K., Rani, B.K. and Govrdhan, A. "Applications of data mining techniques in healthcare and prediction of heart attacks", International Journal on Computer Science and Engineering (IJCSE), pp.250-255, 2010.

[16] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. "From data mining to knowledge discovery in databases", AI magazine, vol. 17, no. 3, pp. 37, 1996.

[17] Koh, H.C. and Tan, G. “Data mining applications in healthcare”, Journal of healthcare information management, pp.65, 2011.

[18] Cheng, J. and Greiner, R. “Comparing Bayesian network classifiers”, In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pp. 101-108, 1999.

[19] Das, R., Turkoglu, I. and Sengur, A. “Effective diagnosis of heart disease through neural networks ensembles”, Expert systems with applications, pp.7675-7680, 2009.

[20] Srinivas, K., Rao, G.R. and Govardhan, A. “Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques”, In Computer Science and Education (ICCSE), IEEE, pp. 1344-1349, 2010.

[21] Peter, T.J. and Somasundaram, K. “An empirical study on prediction of heart disease using classification data mining techniques”, In Advances in Engineering, Science and Management (ICAESM), IEEE, pp. 514-518, 2012.

[22] Canlas, R.D. “Data mining in healthcare: Current applications and issues”, School of Information Systems & Management, Carnegie Mellon University, Australia, 2009.

[23] Gaber, M.M., Zaslavsky, A. and Krishnaswamy, S. “Mining data streams: a review”, ACM Sigmod Record, 34(2), pp.18-26, 2005.

[24] Alizadehsani, R., Habibi, J. and et.al. “A data mining approach for diagnosis of coronary artery disease”, Computer methods and programs in biomedicine, pp.52-61, 2013.

[25] Rajathi, S. and Radhamani, G. “Prediction and analysis of Rheumatic heart disease using kNN classification with ACO”, In Data Mining and Advanced Computing (SAPIENCE), IEEE, pp. 68-73, 2016

[26] Bashir, S., Qamar, U., Khan, F.H. and Naseem, L. “HMV: a medical decision support framework using multi-layer classifiers for disease prediction”, Elsesvier Journal of Computational Science, pp.10-25, 2016.

[27] El Bialy, R., Salama, M.A. and Karam, O. “An ensemble model for heart disease data sets: a generalized model”, In Proceedings of the 10th International Conference on Informatics and Systems, pp. 191-196, 2016.

[28] Onan, A., Korukoğlu, S. and Bulut, H. “BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting”, Australasian Physical & Engineering Sciences in Medicine, Springer, vol. 38, pp.305-323, 2015.

[29] Ismaeel, S., Miri, A. and Chourishi, D. “Using the extreme learning machine (elm) technique for heart disease diagnosis”, In Humanitarian Technology Conference (IHTC2015), IEEE, pp. 1-3, 2015.

[30] Bashir, S., Qamar, U. and Javed, M.Y. “An ensemble based decision support framework for intelligent heart disease diagnosis”, In Information Society (i-Society), IEEE,pp. 259-264, 2014.

[31] Bashir, S., Qamar, U., Khan, F. and Javed, M. “MV5: A Clinical Decision Support Framework for Heart Disease Prediction Using Majority Vote Based Classifier Ensemble”, Arabian Journal for Science & Engineering, Springer, 39(11), 2014.

[32] Amin, S.U., Agarwal, K. and Beg, R. “Genetic neural network based data mining in prediction of heart disease using risk factors”, In Information & Communication Technologies (ICT), IEEE, pp. 1227-1231, 2013.

[33] Chen, A.H., Huang, and et.al. “HDPS: Heart disease prediction system In Computing in Cardiology”, IEEE, pp. 557-560, 2011.

[34] Anbarasi, M., Anupriya, E. and et.al. “Enhanced prediction of heart disease with feature subset selection using genetic algorithm”, International Journal of Engineering Science and Technology, pp. 5370-5376, 2010.

[35] Sivagowry, S., M. Durairaj, and Persia, A. “An empirical study on applying data mining techniques for the analysis and prediction of heart disease”, In Information Communication and Embedded Systems (ICICES), IEEE International Conference, pp. 265-270, 2013.

[36] Novakovic, J. and Veljovic, A., “support vector classification: Selection of kernel and parameters in medical diagnosis”, In Intelligent Systems and Informatics (SISY), IEEE 9th International Conference, pp. 465-470, 2011.

[37] Abdul Razak, T., and Najeeb Ahmed, G. “Detecting Credit Card Fraud using Data Mining Techniques - Meta-Learning”, Indian Journal of Science and Technology, 2015.

[38] Rust, J. “Handbook of computational economics”, Numerical dynamic programming in economics, pp. 619-729, 1996.

[39] Kotsiantis, S.B., Zaharakis, I. and Pintelas, P. “Supervised machine learning: A review of classification techniques”, 2007.

[40] Sałabun, W. and Pietrzykowski, M. “Neural Networks in Economic Problems”, In Selected Issues in Experimental Economics, Springer International Publishing, pp. 245-266, 2016

[41] Helma, C., Gottmann, E., Kramer, S. “Knowledge discovery and data mining in toxicolog”, Statistical methods in medical research, vol. no 9, pp. 329-358, 2000.

[42] Bull,L ., Holmes, J., and Bernadó-Mansilla Ester. “Learning classifier systems in data mining. Berlin”, Springer-Verlag, 2008.

[43] Rajini, N.H., and Bhavani, R. “Computer aided detection of ischemic stroke using segmentation and texture features”, Measurement, Elsevier, pp.1865-1874, 2013.

Research Publications

[1] Meenal, Niyati Baliyan, and Vineeta Bassi, "Prediction of Heart Disease severity with Hybrid Data Mining", IEEE 2nd International Conference on Telecommunication and Networks, Amity University, Noida, 2017. [Accepted]

[2] Meenal, Niyati Baliyan, and Vineeta Bassi, "A Novel Hybrid Classifier for Heart Disease Prediction" [to be communicated]

Video link

https://youtu.be/6zUt_Ng0mzA

thesis_12July

ORIGINALITY REPORT

%11	%5	%6	%5
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	en.wikipedia.org Internet Source	%1
2	S. Rajathi, G. Radhamani. "Prediction and analysis of Rheumatic heart disease using kNN classification with ACO", 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), 2016 Publication	%1
3	Submitted to UT, Dallas Student Paper	%1
4	Submitted to Arab Academy for Science & Technology and Maritime Transport Student Paper	<%1
5	Submitted to Kaplan Professional School of Management Student Paper	<%1
6	Submitted to Higher Education Commission Pakistan Student Paper	<%1