IDENTIFYING SIGNIFICANT FEATURES AND DATA MINING TECHNIQUES IN PREDICTING CARDIOVASCULAR DISEASE

MOHAMMAD SHAFENOOR AMIN

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITI MALAYA
KUALA LUMPUR
2018

IDENTIFYING SIGNIFICANT FEATURES AND DATA MINING TECHNIQUES IN PREDICTING CARDIOVASCULAR DISEASE

MOHAMMAD SHAFENOOR AMIN

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF SOFTWARE ENGINEERING

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

UNIVERSITY OF MALAYA

KUALA LUMPUR

2018

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Mohammad Shafenoor Amin

Registration/Matric No: WGC150021

Name of Degree: Master of Software Engineering

Title of Project Dissertation: Identifying Significant Features and Data Mining Techniques in Predicting Cardiovascular Disease

Field of Study: Data Mining

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

ABSTRACT

Cardiovascular disease is one of the biggest cause for morbidity and mortality among the population of the world. Prediction of cardiovascular disease since regarded as one of the most important subject in the section of clinical data analysis. The amount of data in the healthcare industry is huge. This raw data is needed to be processed to make certain decision on various information. Data mining turns a large collection of data into knowledge. Therefore, the use of data mining in healthcare is obvious. A very high number of researches are conducted on this issue. Nonetheless, a very few has given attention towards the significant features that plays a vital role in predicting cardiovascular disease. Researchers have often focused towards the diagnosis by using different algorithms, sometimes even using the hybrid algorithm. Nonetheless, they have failed to generate an acceptable accuracy in prediction because of using wrong feature selection methods. It has been confirmed that correct features can be more effective when it comes to predicting cardiovascular disease at a very early stage. The problem of finding just the correct combination were addressed in some researches but still lacking effective attempts to improve the accuracy of prediction. A thorough analysis of the features needs to be conducted to select a combination of significant features that can increase the accuracy of the prediction. This research aims to identify significant features and data mining techniques that can improve the accuracy of predicting cardiovascular disease. This research predicts the cardiovascular disease using the identified significant features and data mining techniques. The significant features and techniques were evaluated and achieved accuracy of 87.41%.

ABSTRAK

Penyakit kardiovaskular adalah salah satu penyebab terbesar untuk morbiditi dan kematian di kalangan penduduk dunia. Ramalan penyakit kardiovaskular dianggap sebagai salah satu subjek yang paling penting dalam bahagian data klinikal analysis. Jumlah data dalam industri penjagaan kesihatan adalah besar. Data mentah ini perlu diproses untuk membuat keputusan tertentu bagi pelbagai maklumat. Perlombongan data menukar satu koleksi data yang besar ke dalam pengetahuan. Oleh itu, penggunaan perlombongan data dalam penjagaan kesihatan adalah jelas. Satu jumlah penyelidikan yang sangat tinggi telah dijalankan mengenai isu ini. Namun, sangat sedikit telah memberikan perhatian terhadap ciri-ciri penting yang memainkan peranan penting dalam meramalkan penyakit kardiovaskular. Penyelidik sering memberi tumpuan ke arah diagnosis dengan menggunakan algoritma yang berbeza, kadang-kadang menggunakan algoritma hybrid. Namun begitu, mereka telah gagal untuk menghasilkan ketepatan yang boleh diterima dalam ramalan kerana menggunakan kaedah pemilihan ciri yang salah. Ini telah disahkan bahawa ciri-ciri yang betul boleh menjadi lebih berkesan ketika datang ke meramalkan penyakit kardiovaskular pada peringkat awal. Masalah untuk mencari hanya kombinasi yang betul telah ditangani dalam beberapa peneyelidkan tetapi masih kekurangan percubaan yang berkesan untuk meningkatkan ketepatan ramalan.Satu analisis ciri-ciri yang teliti perlu dijalankan untuk memilih gabungan ciri-ciri yang penting yang boleh meningkatkan ketepatan ramalan. Penyelidkan ini bertujuan untuk mengenal pasti ciri-ciri yang penting dan teknik perlombongan data untuk meningkatkan ketepatan meramalkan penyakit kardiovaskular. Penyelidikan ini meramalkan penyakit kardiovaskular dengan menggunakan ciri-ciri dan teknik perlombongan data yang telah dikenal pasti. Ciri-ciri yang penting dan teknik perlombongan data telah dinilai dan mencapai ketepatan 87.41%.

ACKNOWLEDGEMENT

I express my deep sense of gratitude to my supervisors, Dr. Chiam Yin Kia and Dr. Kasturi Dewi A/P Varathan for their patience and guidance throughout the period of my research. They have been encouraging me with advice and constructive suggestions towards completing my research, giving me close supervision. Their dedication and keen interest and above all their overwhelming attitude to help their students had been mainly responsible for completing my work. Their timely advice, scholarly suggestions and scientific approach have helped me to accomplish this research.

I am also thankful for the panels who reviewed my research work and provided their suggestion for enhancement. Finally, I convey my sincere appreciation to my dear parents who give me fully support in all aspect of life.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Background

"Data mining turns a large collection of data into knowledge" (Han, Pei et al. 2011). The definition shows the importance of data mining in any field. An abundance of data may mean nothing if the data has not turned into meaningful information. Data mining plays a vital role in healthcare area. Researchers from all over the world have come up with different data mining techniques that could improve the disease prediction. Inappropriate use of data mining may not provide the desired outcome. To apply data mining techniques that can produce high accuracy in prediction, the understanding and processing of the data are equally important.

Cardiovascular disease (also known as heart disease) remains the biggest cause for mortality rate throughout the world for past decades. The World Health Organization (WHO) has estimated that 12 million deaths occur worldwide, every year, due to the cardiovascular diseases (Rajkumar & Reena, 2010). If we can predict the cardiovascular disease and provide warning beforehand, a handful of deaths can be prevented. The idea has spurred researchers to conduct research on the matter and come up with better solutions. On that note, data mining enters the scenario. Data mining is useful in an exploratory analysis because of nontrivial information in large volumes of data (Srinivas, Rao, & Govardhan, 2010). The healthcare industry has a huge amount of raw data that needs to be processed. Clinical choices are frequently made focused around instinct and experience of doctors instead of on the knowledge-rich information covered up in the database. This practice prompts undesirable biases, blunders and unnecessary medical expenses that influence the quality of services given to the patients (Anbarasi, Anupriya, & Iyengar, 2010).

The application of the data mining brings a new dimension on the cardiovascular disease prediction. Various data mining tools and techniques are used for identifying and extracting useful information from the clinical dataset with minimal user inputs and efforts (Srinivas, Rani, & Govrdhan, 2010). Over the decade, researchers explored various ways to implement data mining in healthcare to achieve an accurate prediction of cardiovascular diseases. It is crucial to identifying and extracting information using appropriate machine learning algorithm(s). In these studies, researchers usually emphasized on one aspect only and often overlooked other aspects of the data mining. Some tried improving one part of data mining (e.g. preprocessing, feature selection) while the others focus on analysing the performance and accuracy of machine learning algorithm(s) in heart disease analysis. At the end of the day, they all carry the burden of undeveloped aspects with their improvised solutions.

The redundant data in the healthcare sector reduces the accuracy of the algorithms used (Kavitha & Kannan, 2016). Extracting correct combination of feature hence regarded as one of the key factors that sway the outcomes of the algorithms. Furthermore, even with automated knowledge discovery, medical knowledge is important for the feature selection since computer automated process may overlook the importance of the feature from the clinical point of view (Nahar, Imam et al. 2013). Consequently, selecting features along with a suitable data mining technique should be taken into high consideration in order to ensure the prediction of heart disease is acceptable and accurate.

1.2 Problem Statement

The efficiency of data mining largely varies on the techniques used and the features selected. The medical datasets in the healthcare industry are redundant and inconsistent. It is harder to use data mining technique without prior and appropriate preparations. According to (Kavitha and Kannan, 2016), data redundancy and inconsistency in a raw dataset affect the predicted outcome of the algorithms. As a result,

to use the algorithms to its full potential, an effective preparation is needed to preprocess the dataset. Furthermore, unwanted features can reduce the performance of the data mining techniques as well (Paul, Shill et al., 2016). Thus, along with the data preparation, a proper feature selection method is needed to achieve high accuracy in heart disease prediction using significant features and data mining techniques.

Although it has been quite clear that feature selection is as important as the selection of a suitable technique, researchers are still struggling in combining a good technique with a proper set of features. According to (Shouman et al., 2011), there is an expectation to diagnose cardiovascular disease with a high accuracy but is not easy to achieve. According to the authors, a combination of significant features will definitely improve the accuracy. This show that an extensive experiment to identify significant features is necessary to achieve that goal.

The performance of data mining techniques used in predicting cardiovascular disease are greatly affected without a good combination of key features and also improper use of the machine learning algorithms (Khemphila and Boonjing, 2011). Thus, it is crucial to find the best combination of significant features that works incredibly well with the best performing algorithm. Authors highlighted that the accuracy of prediction depends on the best combination of features and also a proper data mining technique. Once again, the emphasis of this reearch focuses on finding the best technique with significant features that perform well in heart disease prediction. However, it is not easy to find the proper technique and select significant features. According to (Nahar et al., 2013), a proper evaluation and comparison to test the different combination of features, together with the data mining techniques are yet to be focused. Thus, a need for a thorough experimentation arises to provide proper identification of techniques and features.

Finally, the existing studies have shown that data mining techniques used in cardiovascular disease prediction are lacking and a proper examination is required to

identify significant features and techniques that will improve the performance. A proper evaluation of such a model is to be compared with the existing ones for truly being able to move forward by achieving higher accuracy.

1.3 Research Objective

This research aims to identify a combination of significant features with a suited technique that provides a higher accuracy in prediction of cardiovascular disease. To achieve this goal, we have to answer the following research questions;

RQ1. What are the significant features and data mining techniques that need to be taken into consideration in predicting cardiovascular disease?

RQ2. How to select significant features and data mining techniques to predict cardiovascular disease?

RQ3. How to evaluate the selected significant features and data mining techniques?

RQ4. How to develop a system to demonstrate the proposed model?

To answer these research questions, we have specified the objectives for this research. This research thrives to achieve the following objectives:

1. To identify significant features in predicting cardiovascular disease
2. To identify data mining techniques in predicting cardiovascular disease
3. To evaluate the performance of the prediction models using the selected significant features and data mining techniques
4. To develop a system to predict heart disease with the identified data mining techniques using the selected significant features

1.4 Research Scope

To achieve the aforementioned stated objective within the planned timeframe, the scope of this research is defined as follows.

Firstly, there are many data mining techniques that can be used to predict heart disease. In this research, only seven data mining techniques were used to perform the experiments. Six techniques are chosen based on popularity in data mining field. The last one, Vote is chosen due to explore its potential in heart disease prediction. However, this research will not focus on any tweaked algorithms.

Secondly, it is not easy to get the real-world patient datasets to predict heart disease, especially the datasets provided by local hospitals. In this research, the open datasets were used to perform the data mining. Both datasets were obtained from UCI machine learning repository in their respective studies.

Finally, there are many predictions can be done on heart disease datasets (e.g. predict the presence of a disease, predict the readmission). This research will solely focus on predicting the absence or presence of cardiovascular diseases in patients.

1.5 Significance of Research

The findings of this research would help in predicting cardiovascular disease and will contribute in heart disease prevention. This would make the diagnosis of the cardiovascular disease a lot easier than the current situation and that will lead to a lesser mortality rate of cardiovascular disease. From the perspective of future researchers, this study would shed new lights of cardiovascular disease prediction in terms of identifying significant features and data mining techniques. Moreover, this research would pave a way towards an accurate cardiovascular disease prediction which will be able to help the

patients by providing early discovery and swift treatment of the disease. Finally, the study identified significant features regarding the prediction of cardiovascular diseases which will raise public awareness of cardiovascular disease risks that can lead to the change of lifestyle and save a huge percentage of life.

1.6 Research Methodology

This section will provide a brief overview of the research methodology. The overall research methodology is discussed in chapter 3. The following methodology is applied for addressing the research problems.

The methodology consists of seven steps. The steps are conducting a literature review, identifying research gap, designing experiments with selective features and data mining techniques for heart disease prediction, evaluating the results and identifying significant features and data mining techniques for heart disease prediction, validating the findings with a different dataset, development of intelligent heart disease prediction system and conclude and summarize the research study. Figure 1.1 illustrates the research methodology of this research.

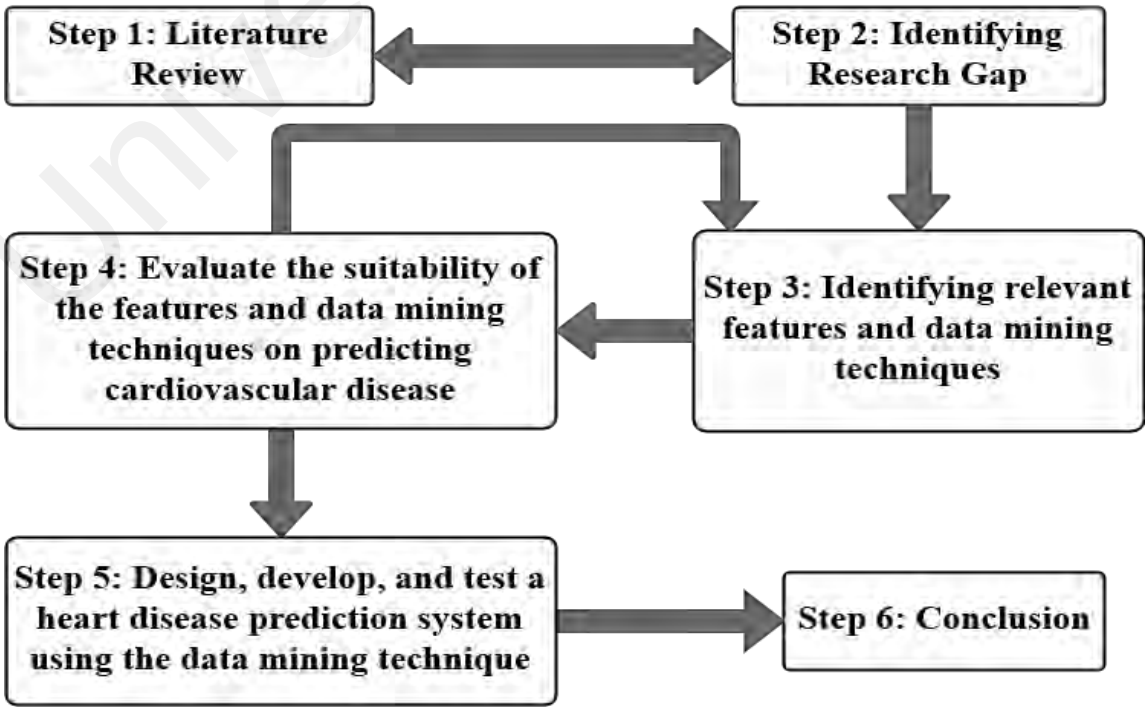

Figure 1.1: Research Methodology

1.6.1 Conducting Literature Review

A literature review was focused on the existing research studies in the cardiovascular disease prediction. By conducting literature review, the research gap was identified. Furthermore, the feature selection methods and data mining techniques used in the existing studies. The knowledge acquired by the literature review allowed us to explore the possibility of using any technique that has not been experimented with. Moreover, the knowledge obtained from the literature review justified the need of identifying significant features and data mining techniques for cardiovascular disease prediction. The detailed prospects and findings of literature can be found in chapter 2.

1.6.2 Identifying Research Gap

Literature review was conducted to identify research gap. This is a back and forth process that allow us to clearly understand the gaps in the existing studies. Based on the literature review, there is a gap concerning a balanced feature selection and data mining techniques. There is a substantial lack of feature selection methods in heart disease prediction.

1.6.3 Identifying Relevant Features and Data Mining Techniques

In this phase, experiments were designed and conducted to identify significant features and data mining techniques. The results from the experiments were analysed and discussed. Significant features that play vital role in predicting heart disease and data mining techniques that performs the very best were derived from the knowledge obtained from the results of the experiment. The entire process was conducted using a framework inspired from the CRISP-DM framework. The whole process is described in Chapter 3 more descriptively. This phase aims to address the research gaps identified in the existing studies. For this stage, nine significant features and top three data mining techniques were identified.

1.6.4 Evaluating the results to identify the significant features and data mining techniques

For this research, two-phase evaluation was conducted to reconfirm the findings of significant features and data mining techniques. In the first phase, experiments was conducted using another dataset, UCI Statlog Heart Disease dataset. The nine significant attributes identified earlier were used to validate the performance of the classification models created using the top three data mining techniques. In the second phase, the model created using significant attributes and data mining techniques was benchmarked against the existing studies. The details of the evaluation process are reported in Chapter 5. The performance of the proposed model (i.e. accuracy) was measured and benchmarked against the accuracy achieved by the models in the existing studies.

1.6.5 Design, Development, and Testing of Intelligent Heart Disease Prediction System

A system was developed in this stage with the identified significant features and data mining technique for the prediction of heart disease. The system was developed in Java with the help of WEKA API. The details of the development process are reported in Chapter 6.

1.6.6 Conclude and Summarize the Research Study

In this stage, the research work was concluded and the fulfilment of the objectives was justified. A conclusion was made based on the findings and evaluation results. Contributions and limitations of this research work were discussed and future research works were suggested.

1.7 Thesis Outline

This thesis contains seven chapters including the introduction. Each chapter describes a specific part of the entire thesis.

Chapter 1 gives an introduction to this research. It consists of background, problems statement, research objective, scope, significant of the research and methodology used to conduct this research. It provides an overview of the research work.

Chapter 2 presents the literature review. It provides review and analysis on the current research and the related studies. This chapter also discusses the research gaps and theoretical platform of the current research.

Chapter 3 shows the research methodology. It describes the steps used to conduct this research.

Chapter 4 explains the identification of significant features and data mining techniques. This chapter describes the methodology used to perform the experiment and reports the outcome of the experiment.

Chapter 5 presents the evaluation of the identified significant features and data mining techniques and discusses the evaluation results.

Chapter 6 describes the design and development of tool based on the proposed model.

Chapter 7 concludes this research and presents a summary of contribution, limitations and potential future research.

Chapter 2: Literature Review

2.1 Introduction

This chapter contains the overview of the literatures that are relevant to our study. Section 2.2 of the chapter describes background of cardiovascular disease and data mining. Section 2.3 discusses the analysis on the literatures we reviewed. It gives a comparative analysis on all the literatures we have selected as our study. The limitation of the existing studies in the field of data mining in cardiovascular disease prediction is discussed in section 2.4. Finally, the entire chapter is summarized and presented in section 2.5.

2.2 Background

First, this section presents the background of heart disease and an overview of data mining. Lastly, the previous reviews related to this work are discussed.

2.2.1 Cardiovascular Disease

Cardiovascular disease or Heart disease is a class of diseases that involve disorders of the heart or blood vessels. It includes coronary artery disease (CAD) such as angina and myocardial infarction (commonly known as a heart attack), stroke, heart failure, hypertensive heart disease, rheumatic heart disease, cardiomyopathy, heart arrhythmia, congenital heart disease, valvular heart disease, carditis, aortic aneurysms, peripheral artery disease, and venous thrombosis. In 2008, heart attacks were responsible for 7.3 million deaths and strokes were responsible for 6.2 million deaths (Mendis, Puska & Norrving, 2011). That is almost over 75% of the total deaths (17.3 million) related to heart disease. There are three main types of heart diseases: Atherosclerotic, Rheumatic and Congenital. Atherosclerotic is a complex pathological process in the walls of blood vessels that develop over many years. It causes the disease most commonly known as

heart attack and stroke. Rheumatic heart disease is caused by damage to the heart muscle and heart valves from rheumatic fever, following a streptococcal pharyngitis/tonsillitis. Congenital heart disease is used to describe any malformation of heart at the time of birth (Mendis, Puska, & Norrving, 2011; World Health Organization, 2017).

A heart attack is preventable if we can predict the risk of a patient getting heart disease and recommend appropriate treatment as early as possible. Researchers (Mendis et al., 2011) have identified the behavioural, metabolic and other risk factors that may cause heart diseases. The risk factors are as follows:

Behavioural risk factors:

- Tobacco use
- Physical inactivity
- Unhealthy diet (rich in salt, fat and calories)
- Harmful use of alcohol

Metabolic risk factors:

- Raised blood pressure (hypertension)
- Raised blood sugar (diabetes)
- Raised blood lipids (e.g. cholesterol)
- Overweight and obesity

Other risk factors:

- Poverty and low educational status
- Advancing age
- Gender

- Inherited (genetic) disposition
- Psychological factors (e.g. stress, depression)
- Other risk factors (e.g. excess homocysteine)

Collection of electronic medical records for cardiac patient encourage the prediction of heart disease using data mining techniques. The identified risk factors help to predict the heart diseases and to recognize the patterns and trends of heart diseases.

2.2.2 Data Mining

Data mining techniques or Knowledge Discovery of the Databases (KDD) is a branch of artificial intelligence. The rapid growth of datasets has led to the discovery of knowledge intelligently using different approaches (Liao, Chu & Hsiao, 2012). Machine learning is one of the commonly used approaches for data mining to turn a large amount of data into knowledge and information. There are two main methods for machine learning: supervised and un-supervised learning. The idea is to learn a way through large amounts of data with the help of data mining and apply the machine learning algorithms in the hopes of discovering relations in an overloaded dataset (Han, Pei, & Kamber, 2011).

2.2.2.1 Machine Learning Algorithms

In the supervised method, researchers specify the desired variable for the machine to learn from the data, on the other hand, the unsupervised method does not specify any label or desire variable. The application of data mining can help in various ways to the different sectors of science. The medical science is benefitted by the application of the data mining in the knowledge discovery. For each method, there are techniques and machine learning algorithms that create models to analyse the hidden information of a dataset. There are four basic types of data mining techniques:

Classification, Clustering, Association, and Regression (Han et al., 2011). Examples of supervised learning are classification and regression. For unsupervised learning, examples are clustering and association.

Classification technique is the form of analysing data that extracts models by giving data classes or labels. These types of models predict categorical (discrete, unordered) class labels. It consists of two steps: Learning and Testing. In the first step, a classifier is built using a predetermined set of data classes or concepts. The next step is to check the classifier model on a given data to determine its acceptability. Figure 2.1 shows the two steps method for classification where data are labelled and represented accordingly.

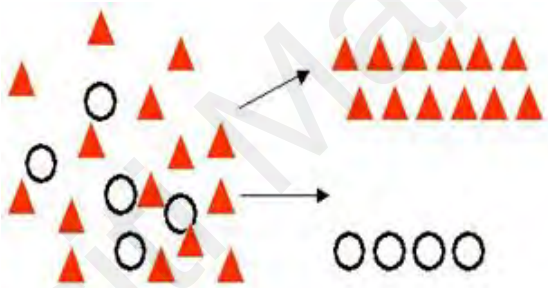

Figure 2.1: Classification Techniques (adopted from Oracle (n.d.))

Clustering technique divides the data into small groups in such a way that each member of a group is very similar to other members of the same group. On the other hand, members of different groups will be very different from one another. Unlike classification, in clustering the labels are unknown and the purpose is to discover those (Han et al., 2011). Figure 2.2 shows the clustering as it is being described. The process of clustering is done by an algorithm. There are many algorithms with different ways to

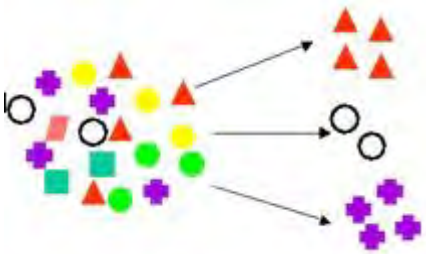

Figure 2.2: Clustering Techniques (adopted from Oracle (n.d.))

cluster data but the end result of all is to discover the clusters within the scattered data (Han et al., 2011).

Association technique, often known as frequent pattern mining, is the process of finding out recurring relations in a dataset. In other words, association rules down the behaviour of a dataset (Han et al., 2011). Let I = [i1, i2,…., in] be a set of n binary attributes called items and D = [t1, t2,…..,tm] be the set of transaction called database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. Then the rule is defined as P=>Q, where P, Q are a disjoint subset of item I (Agrawal, Imieliński, & Swami, 1993). Figure 2.3 shows the process of association mining where associated objects are being put together to show their association.

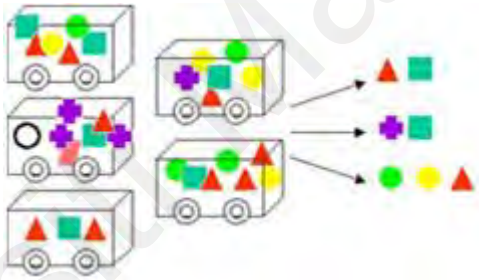

Figure 2.3: Association Techniques (adopted from Oracle (n.d.))

Regression technique is a statistical method that is mainly used for numeric prediction. Along with classification, this does the work of prediction in the field of data mining (Han et al., 2011). It focuses on the relationship between a dependent variable with one or more independent variables. Independent variables are called predictors. Figure 2.4 shows the concept of a regression technique.

Figure 2.4: Regression Techniques (adopted from Oracle (n.d.))

There are plenty of algorithms in each of the techniques developed over the years that can be used to predict and diagnose the heart disease datasets. The techniques and algorithms used in data mining have their strengths and limitations in analysing and predicting heart disease for valuable knowledge. According to (Abdar, Zomorodi-Moghadam, Das and Ting, 2017), it is crucial to ensure the accuracy in the analysis and application of medical data to avoid the mistakes in the results that may trigger threats to human life.

2.2.2.1.1 Naïve Bayes (NB)

Naive Bayes classifier is based on Bayes theorem. The algorithm assumes that an attribute value on a given class is independent of the values of other attributes. The Bayes theorem is as follows (Han et al., 2011):

Let X={x1, x2, ..., xn} be a set of n attributes. In Bayesian, X is considered as evidence and H be some hypothesis means, the data of X belongs to specific class C. We must determine P (H|X), the probability that the hypothesis H holds given evidence i.e. data sample X. According to Bayes theorem the P (H|X) is expressed as;

$$P(H|X) = P(X|H)\,P(H) / P(X)$$

2.2.2.1.2 KNN

KNN classifiers are based on learning by comparing a given test tuple with training tuples that are similar (Han et al., 2011). The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. Thus, all training tuples are stored in a n-dimensional pattern space. For an unknown tuple, it searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k "nearest neighbours" of the unknown tuple. "Closeness" is often derived by Euclidean distance.

2.2.2.1.3 Decision Tree (DT)

Decision tree builds classification models in the form of a tree structure. Entire dataset is broken down into subsets while an incremental development of tree is built in this technique (Han et al., 2011). The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches while leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node.

2.2.2.1.4 Neural Network (NN)

Neural network is a set of connected input/output units in which each connection has a weight associated with it (Han et al., 2011). During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples. It consists of an input layer, one or more hidden layers, and an output layer. Each layer is made up of units. The inputs to the network correspond to the attributes measured for each training tuple. The inputs are fed simultaneously into the units making up the input layer. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer of "neuronlike" units, known as a hidden layer. The outputs of the hidden layer units can be input to another hidden layer, and so on. The number of hidden layers is arbitrary, although in practice, usually only one is used. The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction for given tuples (Han et al., 2011).

2.2.2.1.5 Support Vector Machine (SVM)

A Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors (cases) that

define the hyperplane are the support vectors. SVM algorithm can be described into three basic steps (Han et al., 2011):

Step-1: Define an optimal hyperplane: maximize margin

Step-2: Extend the above definition for non-linearly separable problems: have a penalty term for misclassifications.

Step-3: Map data to high dimensional space where it is easier to classify with linear decision surfaces: reformulate problem so that data is mapped implicitly to this space.

2.2.2.1.6 Logistic Regression (LR)

Logistic regression predicts the probability of an outcome that can only have two values (Han et al., 2011). The prediction is based on the use of one or several predictors (numerical and categorical). It produces a logistic curve which is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.

2.2.2.1.7 Vote

Ensemble methods are techniques that create multiple models and then combine them to produce improved results (Wu et al., 2008). Vote is an ensemble method. The first step is to create multiple classification models using some training dataset. Each base model can be created using different splits of the same training dataset and same algorithm, or using the same dataset with different algorithms, or any other method. Every model makes a prediction (votes) for each test instance and the final output prediction is the one that receives more than half of the votes.

2.2.2.2 Preprocessing

Preprocessing is the most important step when applying data mining model to a dataset. “Data preprocessing allows transforming the original data into a suitable shape to be used by a particular mining algorithm” (Romero & Ventura, 2007). It is quite common that any or most of the datasets are often filled with missing value, erroneous values, unexpected values, duplicate entries and missing attributes. Attempting to use these datasets cannot only cause the algorithm to produce ineffective results but also can sometimes produce faulty results. Therefore, it is important that this step is not being ignored. The preprocessing normally consists of four steps; Data Cleaning, Data Reduction, Data Transformation and Data Integration. The steps are about getting a clearer dataset suitable for applying mining algorithm. In cleaning part, filling the missing value and resolving inconsistencies are taken care of. Then in the reduction, duplicate entries are erased without inflicting any harm on the pattern of the dataset. Then comes the transformation of garbage data or unreadable data in recognized readable formats. The generalization and scaling are done in this part. The part about integration is very important if different sources have been used for the generation of the dataset. Any anomalies occur in the coding schemes or any discrepancies while combining dataset from different source are dealt in this part (Acuña, 2011). A well preprocessed dataset is bound to produce an effective result on the algorithm.

2.2.2.3 Feature Selection Methods

Feature Selection is the process for eliminating attributes with little or no predictive information. The significance of this step is very highlighted throughout data mining as a good selection increases the efficiency of the techniques applied. Furthermore, finding a correct subset of predictive feature can be problematic and often leads to discovery of new ways of implementing a technique (Kim, Street et al. 2003). In (Nahar, Imam et al. 2013, Rajeswari, Vaithiyanathan et al. 2013), it was proven beyond

doubt that a good feature selection can cause greater improvements in techniques applied. There are many different types of feature selection techniques used and almost all of them has its own perks and downs. Most common practice for many studies is to use the full features. Studies that bothers to use feature selection before applying techniques are often found to have been using Motivated Feature Selection (MFS), Correlation-based Feature Selection (CFS), Wrapper-based Feature Selection (WFS), Rough Set Theory etc.

Motivated Feature Selection (MFS) is the feature selection method based on the opinion of the medical experts. Basically, some expert chooses a combination of attributes best suited for the purpose and use it with the techniques.

Correlation-based Feature Selection (CFS) evaluates a subset of features by considering the predictive capabilities and the level of redundancy of each feature. A good feature subset is a subset of features that contain features that are highly correlated with the class and each feature is not correlated (Setiawan, Prabowo et al., 2014).

Wrapper-based Feature Selection (WFS) method utilizes the induction algorithm. It basically evaluates a subset for a given technique and then chooses the highest accurate one. Algorithms like Particle Swarm Optimization, Genetic Optimization etc. comes into play while using this method.

Rough Set Theory (RST) is a mathematical method for finding the patterns hidden inside the dataset. Originated by Pawlak this method can also be utilized for feature selection (Setiawan, Prabowo et al., 2014).

2.2.2.4 Model Evaluation

In data mining, model evaluation is the way to validate the model performance and identify the best model that represents the heart disease data. To evaluate the models, Hold-out and Cross Validation are the two main methods that can be used. Hold-out is mostly used for the large dataset and randomly divided the data into three subsets:

Training set, Validation set and Test Set. Some algorithms may not need a Validation set. On the other hand, k-fold cross-validation is suitable for model evaluation when the size of available data is limited. The data are divided into k subsets of equal size, the subsets are further partitioned into Training set, and Test set. For both methods, the Test set is used to evaluate the performance of a model in data mining.

2.3 Review on the application of data mining techniques in heart disease prediction

This section describes the research methods used to conduct this review.

2.3.1 Review Methods

This section describes the research methods used to conduct this review.

2.3.1.1 Objectives

Many studies have been conducted over the years to explore the usage of data mining techniques in predicting heart disease. Researchers thrive to get better performance and accuracy by applying different types of data mining technique. Each researcher has their own way of applying the techniques. The objectives of this review are:

- to gather and present the available research studies or significant past literature that have conducted diagnosis and prediction of heart diseases using data mining techniques, and
- to review the current state of data mining techniques for use in improving the accuracy of the diagnosis and prediction of heart diseases.

2.3.1.2 Search Strategies

This research aims to review and analyse the studies that have applied data mining in heart disease prediction, starting from 2010 to 2016. To meet the research

objectives, we developed a search strategy. We used the Google Scholar (www.scholar.google.com) to retrieve all the relevant papers. The database was chosen because it provides efficient search facility to retrieve a large number of research papers from different journals and conferences in a single search. The query string used for the searching is ("data mining" heart disease). We have performed three searches in this study. Firstly, we searched the papers and custom the year range to retrieve publications from 2010 to 2016. We included the papers from the first 20 pages of the search results. Upon realizing that this list does not contain any papers from 2015 and 2016, we performed the second search and set the custom range from 2015 to 2016. We collected papers from the first 5 pages of the search results and found that there were no papers from 2016. Eventually, we conducted the third search and set the custom range from 2016 to 2016.

2.3.1.3 Inclusion and Exclusion Criteria

Once we retrieve all the papers, we started the study selection process by reading the title, abstract and full text. We selected and shortlisted the papers based on the inclusion and exclusion criteria specified in Table 2.1. A list of 76 primary study papers was selected for our final study. We excluded all the non-English papers. Irrelevant studies are excluded in this review.

Table 2.1: Selection Criteria

Inclusion Criteria	**Exclusion Criteria**
Studies are written in English Studies that describe data mining technique applied to heart disease dataset Studies published between 2010 to 2016 Peered-reviewed academic articles (Full journal or conference papers)	Studies that are books Studies that use text mining Studies that use ECG signals Studies that use Image processing Studies that are review or comparison only Studies that do not have sufficient information on applying data mining technique

	Studies that do not apply any data mining technique

2.3.1.4 Study Selection

Study selection was performed in three phases. Firstly, the authors conducted the three searches to retrieve all the potential publications. Next, the first author selected an initial list of papers based on the inclusion and exclusion criteria. Once initial list of papers is selected, the other two researchers read the selected papers to cross-check the initial list. A check for duplication, missing studies and the unfit studies was conducted in a few research meetings to resolve the conflicts and finalise the selection. After thorough analysis, the final list of 76 papers that fulfill the inclusion criteria was shortlisted. In this study, we only included journal and conference papers.

The papers were selected following the steps as illustrated in Figure 2.5. The search was made in Google Scholar database with three different range; 2010 to 2016, 2015 to 2016 and 2016 to 2016. All three searches were made using the same search string ("data mining" heart disease). The search results yield 17,400, 13,100 and 6,600 papers respectively. For the study selection, only the first 20, 5 and 5 pages were filtered respectively to collect the initial list of papers. Each page consists of 10 results. The list was narrowed down to 200, 50 and 50 search results. Next, all the researchers using the inclusion and exclusion criteria to finalise and select the final list of papers screened the studies. After applying the filtering criteria, the number of studies selected for this study was 61, 6 and 9 respectively. 76 papers were selected for this review (List of selected studies can be found in Appendix A).

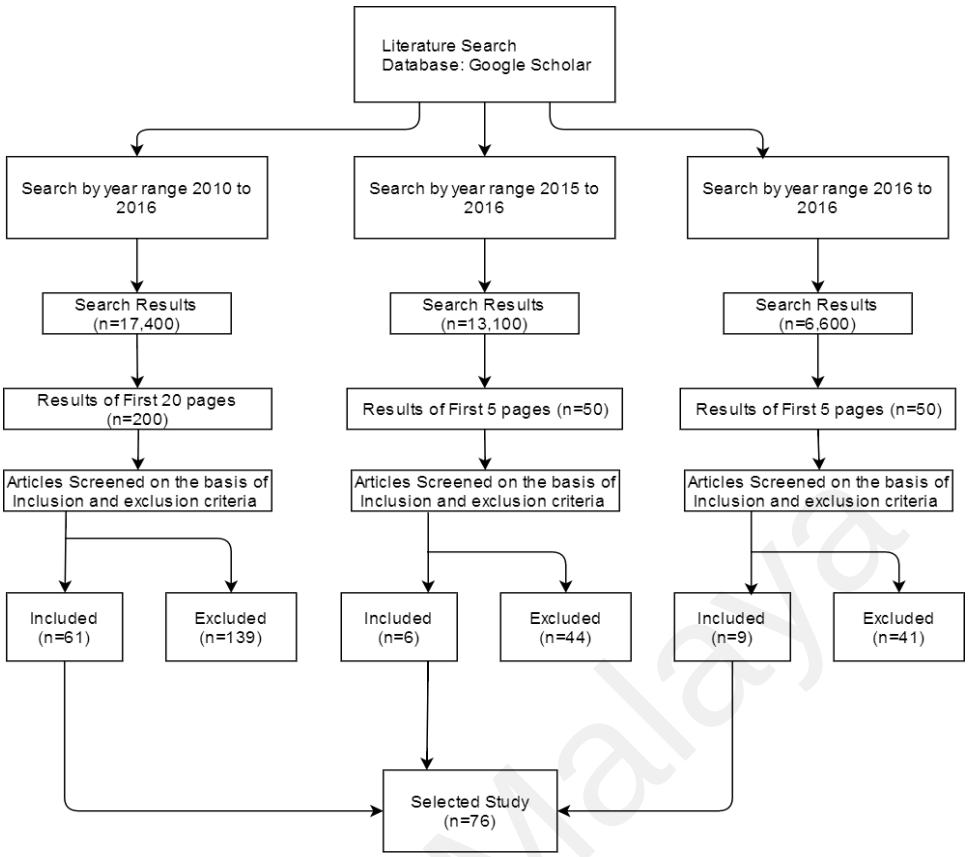

Figure 2.5: Study Selection Process

2.3.1.5 Data Extraction

Data extraction was performed by one of the researchers of this paper. The other two researchers played the role to validate the extracted data of each paper. The researchers discussed and resolved all the issues regarding the data extraction during research meetings. This extracted data was later analysed by the authors to answer the research question of this paper. The following data were extracted from the selected studies by reading the full paper:

- Article bibliographic information (author and publication year)
- Article type (journal, conference)
- Dataset used in the article to predict heart disease or another disease
- Features used in the heart disease prediction

- Method use to separate dataset into a training set and testing set
- Method used for feature selection
- Techniques used in data mining
- Tool used to perform data mining
- Method and criteria used in evaluating the data mining model

2.3.2 Comparative analysis of the data mining techniques in heart disease prediction

This section analyses and compares the dataset, preprocessing method, feature selection method, data mining techniques, algorithms, model evaluation techniques and tools used in the selected studies for heart disease prediction.

2.3.2.1 Input Dataset

A clinical dataset is an input data that can be analysed and turned the raw data into meaningful knowledge that can help to predict heart disease. A dataset with too many inconsistencies can reduce the performance of any machine-learning algorithm in data mining. Furthermore, a dataset with too many missing values can question the credibility and accuracy of the prediction. All the papers were studied to extract the input dataset used in their research. Most of the studies, except three studies (i.e. S4, S16, S50), have mentioned the source of the input datasets. There are three main types of data sources: publicly available dataset, a collection of data through surveys, and electronic medical records (EMR) collected by hospitals, medical and research centres or health maintenance organization (HMO) data warehouse.

Figure 2.6 shows the analysis of the type of dataset source used in the selected studies. Based on the analysis, it is clear that most of the researchers (i.e. 55 studies) using the publicly available heart disease dataset provided by UCI Machine Learning Repository (Blake & Merz, 1998). Only 16 studies collected the patient EMR through

various means on their own and/or with the help from hospitals, medical and research centres or HMO data warehouse in their respective countries. The EMR collected was converted into clinical datasets. Bandyopadhyay et al. (2015) conducted their research has used different sources of datasets provided by HMO Research Network Virtual Data Warehouse (HMORN VDW) from a healthcare system in Midwestern United States. Table 2.2 shows the list of studies and the source of the dataset used in the research. The table corresponds to the IDs of the papers that have used the dataset source. Among these papers, 10 studies (i.e. S15, S18, S28, S31, S48, S51, S53, S62, S64, S76) have used more than one type of clinical datasets along with the heart disease dataset in their experiments.

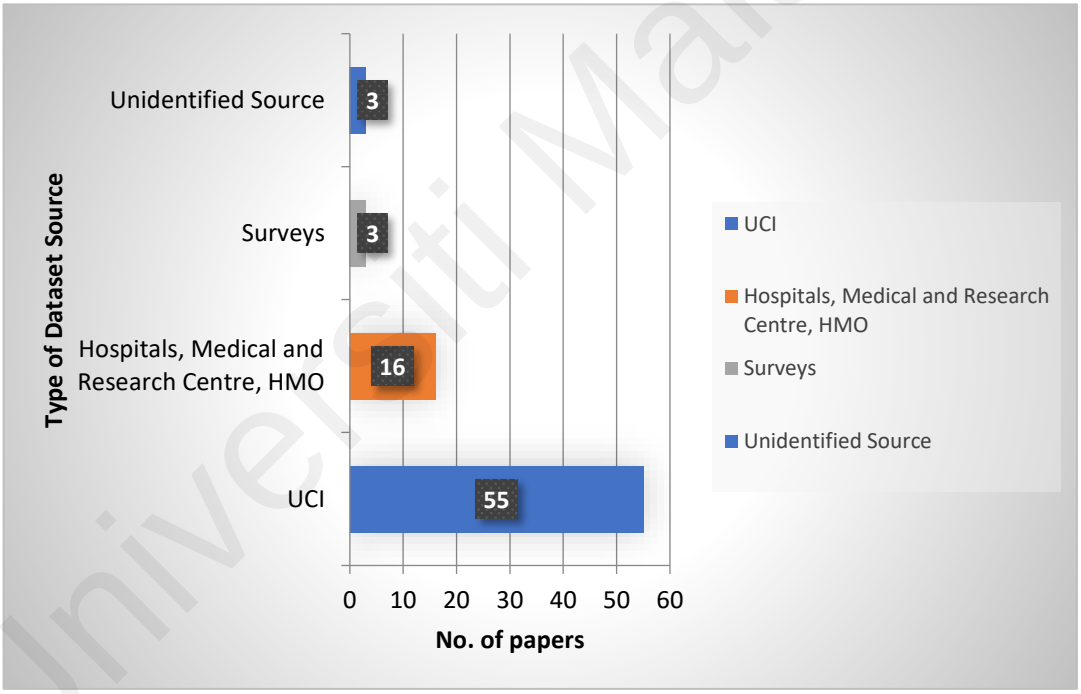

Figure 2.6: Datasets source used in the selected studies

Table 2.2: List of studies and dataset used

Dataset Source	Study ID
Publicly available dataset (UCI repository)	55 studies: S1, S3, S7, S8, S9, S11, S12, S13, S14, S15, S17, S18, S19, S21, S22, S23, S24, S25, S26, S27, S28, S29, S30, S32, S34, S35, S36, S37, S38, S41, S42, S43, S44, S45, S48, S51, S52, S53, S54, S55, S56, S58, S59, S60, S61, S64, S65, S68, S70, S71, S72, S73, S74, S75, S76
EMR collected at hospitals, medical and research centres or HMO	16 studies: S5 (Cyprus - Paphos General Hospital) S6 (China - Longhua Hospital) S10 (India - Diabetic Research Institute in Chennai) S20 (China, Dongzhimen Hospital) S31 (USA)

	S33 (Russia, Central Clinical Hospital No. 2 of Russian Railways JSC) S39 (India - Madras Medical College, Chennai) S46 (Iran - Academic and Educational Hospital of Rajaei Cardiovascular Medical & Research Center in Tehran) S48 (India - Various Corporate Hospital in Andhra Pradesh state) S49 (India - The Postgraduate Institute of Medical Education and Research (PGI) is a premier medical and research institution in Chandigarh) S51 (India - Various Corporate Hospital in Andhra Pradesh state) S51 (South Africa - Collected from medical practitioners in South Africa) S62 (India - Diagnostic Centre) S66 (Iran – collected from a hospital in Iran) S67 (USA - HMO Research Network Virtual Data Warehouse (HMORN VDW) from a healthcare system from Midwestern United States) S69 (India - Indira Gandhi Medical College (IGMC), Shimla, India)
Surveys	3 studies: S2 (India, survey conducted in Andra Andhra Pradesh state) S40 (The USA, survey collected by American Heart Association) S63 (Korea - KNHANES-VI, a survey study conducted by the Korea Centers for Disease Control and Prevention)
Source of dataset unidentified	3 studies: S4, S16, S50

Figure 2.7 illustrates the heat map generated from the countries of the dataset source. The heat map shows the graphical representation of the location of the heart disease datasets. The dark green indicates over 30 studies are using dataset collected from that country (i.e. USA). There are 58 papers using heart disease datasets from the USA. Among them, 55 studies using UCI dataset from Cleveland, USA, 1 study using the survey data collected by American Heart Association and 2 studies using data collected from hospitals and HMO data warehouse. Most of the researchers prefer to use the publicly available UCI dataset from the USA, even though the researchers are from a different country. On the other hand, light green indicates less than 10 studies using the dataset from that country. Besides the USA, there are 8 studies collected data from India, 2 from China, 1 from Korea, Cyprus, Russia and South Africa, respectively.

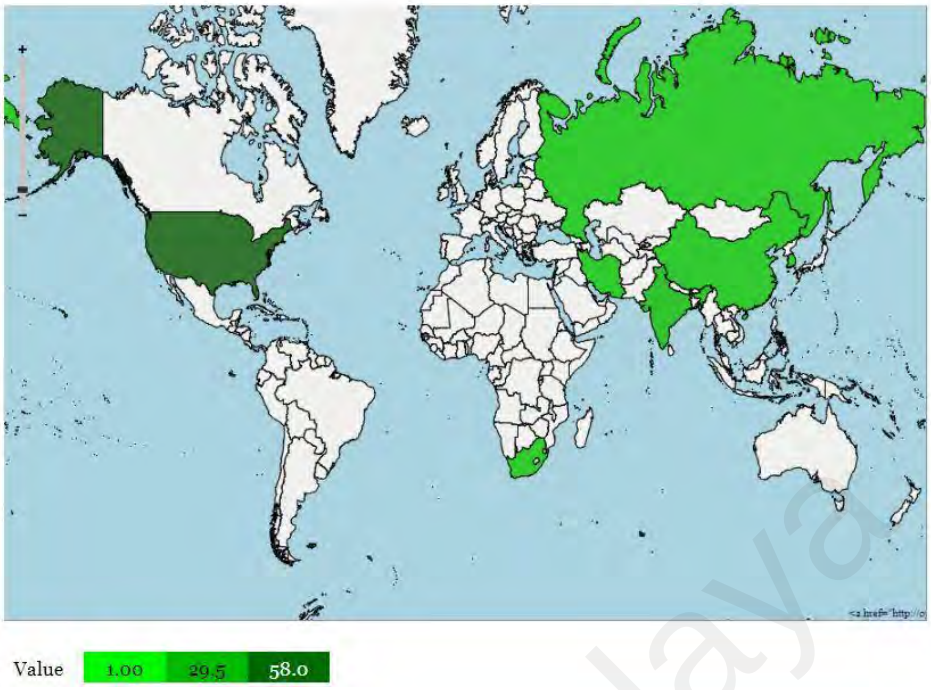

Figure 2.7: Heat map generated based on the country of dataset source

2.3.2.2 Feature Selection and Methods

In machine learning, feature selection otherwise known as attribute selection is the process for eliminating attributes with little or no predictive information. It is significant in improving the performance of the technique applied. Furthermore, finding a correct subset of predictive feature can be problematic and often leads to the discovery of new ways of implementing a technique (Kim, Street, & Menczer, 2003). In this study, we identified feature selection algorithm(s) used to choose the most significant features to predict heart disease. In this review, we have identified five studies (Jabbar et al., 2011; Peter & Somasundaram, 2012; Rajeswari et al., 2012; Nahar et al., 2013b; Kavitha & Kannan, 2016) that mainly focus on feature selection in order to improve the performance of a technique applied in data mining.

Table 2.3 shows some of the algorithms used in the feature selection methods found in the recent studies. It shows the diversity of algorithms available for feature

selection. The algorithms are often combined with the data mining techniques discussed in Subsection 2.2.2.1, except for using weights in the features to indicate the importance of the feature. Furthermore, there have been initiatives to select features manually by experimenting on different combination of features. Such a technique is brute force. According to (Rajeswari et al., 2013), brute force method was applied to identify a significant attribute which seems to increase the accuracy of the techniques applied.

Table 2.3: Feature selection methods used in existing studies

Source	**Feature Selection Methods Used**
Jabbar et al. (2011)	CBARBSN
Shouman, Turner, & Stocker (2011)	Information Gain, Gini Index, Gain Ratio
Khemphila & Boonjing (2011)	Information Gain
Peter and Somasundaram (2012)	CFS with Bayes Theorem
Nahar et al. (2013)	MFS
Rajeswari et al. (2013)	Genetic Algorithm
Subanya & Rajalaxmi (2014)	Artificial Bee Colony Algorithm
El-Bialy, Salamay, Karam, & Khalifa (2015)	Selected Attribute
Paul et al. (2016)	Genetic Algorithm
Verma, Srivastava, & Negi (2016)	PSO
Kavitha & Kannan (2016)	PCA
Anbarasi, Anupriya, & Iyengar (2010)	Genetic Algorithm
Peker (2016)	Attribute Weights

Cluster Based Association Rule Mining Based on Sequence Number (CBARBSN) is a new association rule mining approach proposed by (Jabbar et al., 2011) which suggest the selection of features based on sequence numbering and clustering to predict heart disease. The efficiency of CBARSN and other algorithms was measured in terms of the support and execution time to mine the association rules. Frequent item sets were generated on each cluster with an iteration that identifies frequent itemsets for heart disease. Age, BP, Max Heart Rate, Old peak and Thal were found as frequent features in heart disease dataset. Applying the CBARBSN algorithm has outperformed other algorithms in mining association rules.

Peter and Somasundaram, (2012) proposed a new hybrid method that combines computerised feature selection (CFS) and Bayes Theorem. As compared to CFS that has chosen 3 attributes, the hybrid algorithm selected only 3 significant attributes from 14 attributes. They used the 3 attributes to identify the best classification algorithms by calculating the accuracy. The performance was notable for Naive Bayes, Multilayer Perception, J48 and KNN, which achieved accuracy above 80%. However, On the other hand, in overall compared with other feature selection methods, the proposed hybrid method achieved higher accuracy (84.07%).

Neural Network is one of the popular techniques applied in feature selection for mining heart disease datasets. Thus, a study (Rajeswari et al., 2012) found a way that utilised feed forward neural network to reduce the number of attributes without affecting the accuracy of the prediction. The researchers went through an iteration based on feature reduction and calculated accuracy for each turn. The highest accuracy was achieved using 12 attributes that yielded the accuracy of 89.4% in training and 82.2% in testing.

A medical feature selection method called MFS was proposed by (Nahar et al., 2013b). The idea was to factor in the medical knowledge while selecting significant features that are often overlooked in the feature selection process. The researchers merged MFS and CFS to select medically significant features and the experiments achieved some interesting results in heart disease prediction. The combination of MFS and CFS yields good performance for Naïve Bayes, IBK and Support Vector Machine (SMO) algorithms. Though there were instances where the proposed MFS and CFS method was performed poorly, the researchers concluded the method can further be developed with an extension to achieve a consistent result.

Genetic Algorithm (GA) (Anbarasi et al., 2010; Kavitha et al., 2010; Bhatla & Jyoti, 2012a; Atkov et al., 2012; Ephzibah & Sundarapandian, 2012; Amin et al., 2013;

Jabbar et al. 2013c; Manikantan & Latha, 2013; M. A. Jabbar et al., 2013a; Methaila et al., 2014) is a popular machine learning algorithm for feature selection, especially when it works together with neural network. The combination of these two algorithms outperforms any traditional neural networks and also reduce the execution time required for training the model. The traditional backpropagation algorithm that is used for the neural network has two major drawbacks. Firstly, it starts with the blindly weighting process and secondly, it takes a longer time for the model learning process. After combining with GA, neural network overcomes the two drawbacks and shows better performance.

Principal Component Analysis (PCA) (Kavitha & Kannan, 2016; Kaur & Singh, 2016; Dey et al., 2016) is a feature extraction method that seems to gain popularity in recent years. In one of the studies (Kavitha & Kannan, 2016), PCA was combined with information gain ratio for feature selection. The researcher claimed that the framework developed based on PCA has increased efficiency, accuracy and speed as compared with other scoring function. However, there are no experiment details provided in the paper to support this argument. Another two studies (Kaur & Singh, 2016; Dey et al., 2016) conducted experiments to evaluate the PCA based feature selection method. Kaur and Singh (2016) combined PCA with SVM algorithms while Dey et al. (2016) applied SVM, Naive Bayes and Decision tree together with PCA. These two studies prove that the reduction of attributes using PCA based feature extraction has increased the accuracy of prediction, especially for SVM.

Maximal Frequent Itemset Algorithm (MAFIA) (Khaing, 2011; Manikantan & Latha, 2013; Banu & Gomathy, 2014) and Particle Swarm Optimization (PSO) (Kalaiselvi & Nasira, 2015; Verma et al., 2016) are the two rising algorithms for feature selection. MAFIA is very efficient for the larger database (Banu & Gomathy, 2014). MAFIA is the identification of frequent item sets from a database. K-mean based MAFIA

implemented with ID3 and C4.5 algorithms has achieved high precision, recall and accuracy in prediction (i.e. precision of 0.82, recall of 0.89 and accuracy of 89.5%) (Banu & Gomathy, 2014). On the other hand, PSO is a type of population-based optimization algorithm. Combining PSO with k-means clustering algorithm (Verma et al., 2016), the model developed using multinomial logistic regression (MLR) has yielded an accuracy of 91.36% in heart disease prediction. In another experiment performed by Kaliselvi and Nasira (2015), PSO was used with neuro-fuzzy and the prediction results gave an accuracy of 98%.

Table 2.4 show the features selected by some of the existing studies for cardiovascular diseases using UCI Cleveland dataset. The table highlights that most of them uses the full features while others may choose but doesn't do the job efficiently. From the table we can see that most of the papers have used 13 attributes and lowest was 6 attributes. Attributes named age, cp and thalach is the highest used among the recent studies whereas others are not so left behind with the exception from sex.

Table 2.4: Features used in existing studies

Source	**Age**	**Sex**	**cp**	**trestbps**	**chol**	**fbs**	**restecg**	**thalach**	**exang**	**oldpeak**	**slope**	**ca**	**thal**	**Total**
(Anooj 2012)	√			√	√			√		√			√	6
(Shouman, Turner et al. 2011)	√	√	√	√	√	√	√	√	√	√	√	√	√	13
(Nahar, Imam et al. 2013)	√		√	√	√	√	√	√	√					8
(Medhekar, Bote et al. 2013)	√	√	√	√	√	√	√	√	√	√	√	√	√	13
(Ziasabounchi and Askerzade 2014)	√		√	√	√	√	√	√						7
(Rajeswari, Vaithiyanathan et al. 2013)			√		√			√	√	√	√	√	√	8
(Subanya and Rajalaxmi 2014)	√				√	√	√	√			√	√		7
(Saxena and Sharma 2015)	√	√	√	√	√	√	√	√	√	√	√	√	√	13
(El-Bialy, Salamay et al. 2015)			√					√				√	√	4
(Ismaeel, Miri et al. 2015)	√	√	√	√	√	√	√	√	√	√	√	√	√	13

(Khemphila and Boonjing 2011)	√		√					√	√	√	√	√	√	8
(Chaurasia and Pal 2013)	√	√	√	√	√	√	√	√	√	√	√	√	√	13
(Fathurachman, Kalsum et al. 2014)	√	√	√	√	√	√	√	√	√	√	√	√	√	13
(Turabieh 2016)	√	√	√	√	√	√	√	√	√	√	√	√	√	13
(Paul, Shill et al. 2016)	√	√	√	√			√	√	√	√	√	√	√	12
Total	13	8	13	11	12	10	11	15	11	11	11	12	12	

2.3.2.3 Data Mining Techniques

The core of the research is the data mining techniques used to predict heart disease. Moreover, using an appropriate technique can often mean reduce mistakes in pointing false positives. The first and foremost concern is the performance of the techniques. Furthermore, there are various data mining techniques can be used to predict heart disease. Over the years, many machine-learning algorithms were applied to achieve

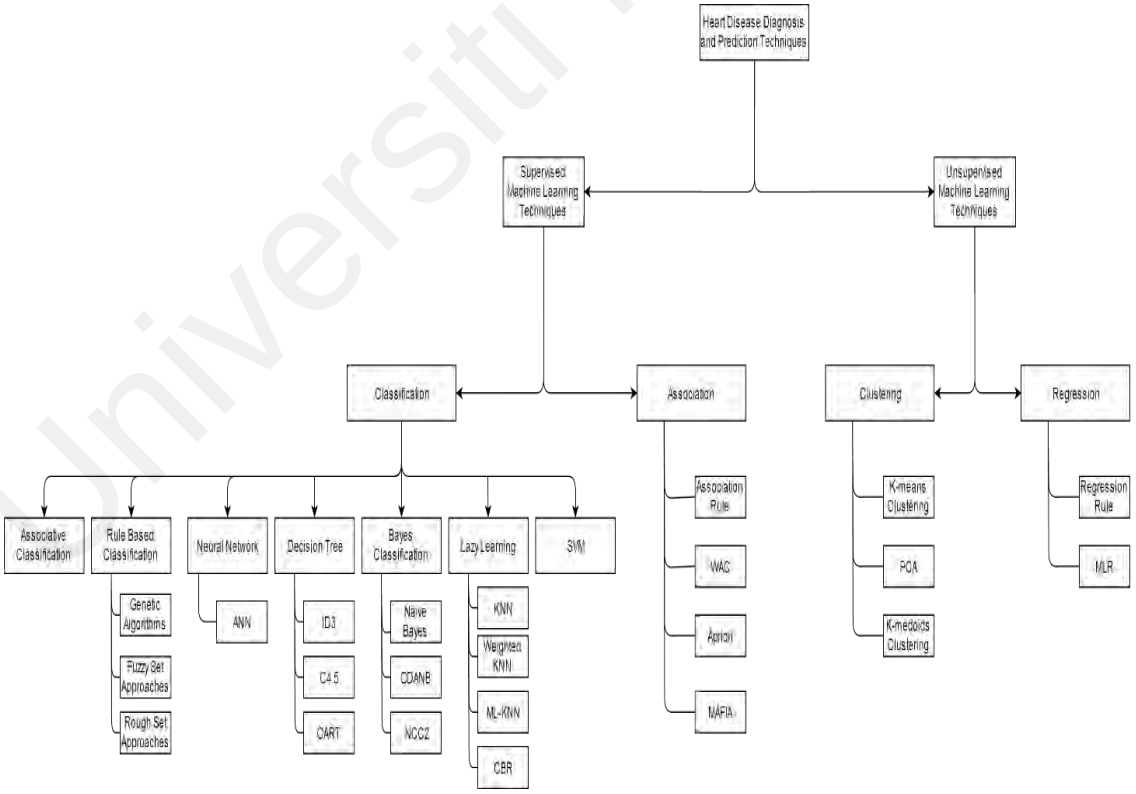

Figure 2.8: Classification of data mining techniques used in heart disease prediction

better performance in prediction. Some studies also explored the performance of hybrid techniques, which is a combination of two or three techniques. In this review, all the data

mining techniques that have been applied for heart disease diagnosis and prediction in the 76 selected studies are identified. Figure 2.8 classifies all the data mining into four basic types of data mining techniques: Classification, Clustering, Association, and Regression (Han et al., 2011). First, all the techniques are grouped into two main categories: Supervised and Unsupervised. The supervised techniques are divided into Classification Techniques and Association Techniques. On the other hand, unsupervised techniques are divided into Clustering Techniques and Regression Techniques. Although the regression techniques are often referred as a statistical method, they have been implemented in data mining for outlier detection on numerous occasions. The leaf nodes of the tree are some algorithms that researchers mentioned in their studies with a correlation of relevant techniques. Based on Figure 2.9, classification techniques are the most commonly used data mining techniques to create models for heart disease prediction.

Figure 2.9 shows the top-ten data mining technique and the number of times the techniques are used in the selected studies. This analysis includes the all the data mining techniques used by the researchers. There are studies that use more than one techniques

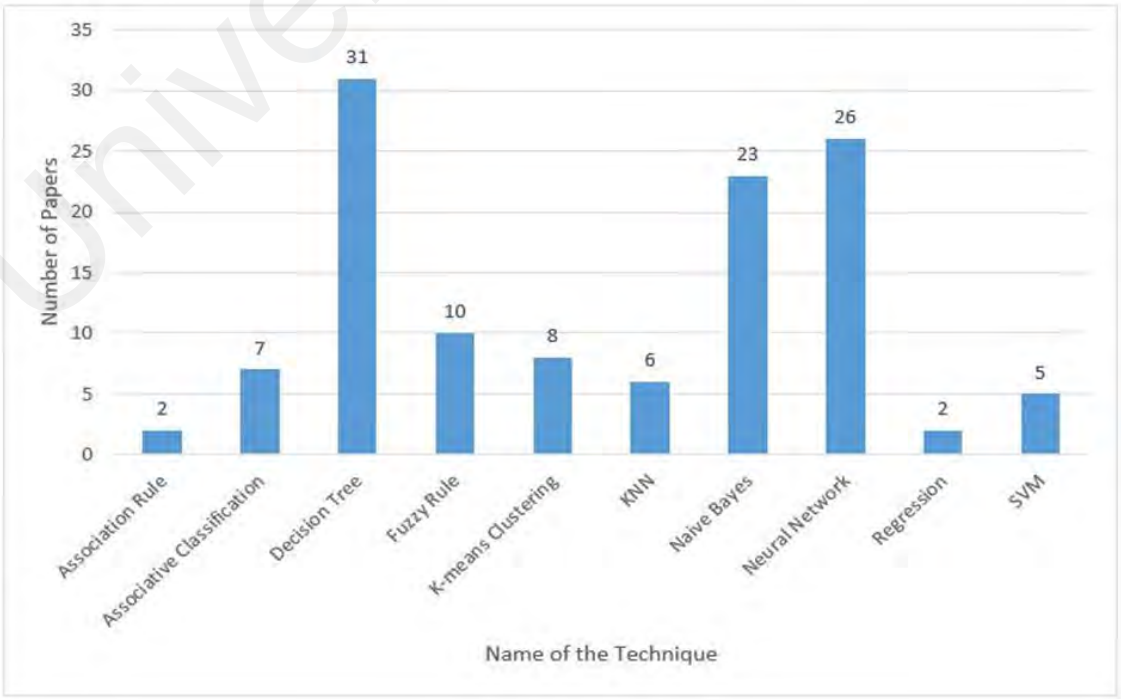

Figure 2.9: Data mining techniques used in the selected studies

separately or use hybrid techniques. In our analysis, we have counted each usage of the technique separately. According to Figure 2.9, the top three data mining techniques are Decision Tree, Neural Network and Naïve Bayes. The decision tree has always been number one while the neural network was catching up to it on two occasions. Naïve Bayes have had a significant drop in the most recent year. The trend suggests that Decision Tree is not only giving better results but also easy to tweak or modify. In recent years, researchers attempted to extend the existing algorithms to achieve higher accuracy. It is understandable that any technique that does not allow modification or cannot simply outperform the current algorithm even with tweaking is declining from their popularity. Decision Tree has maintained its status throughout the years and also increase its popularity in recent years. This suggests that decision tree has been used in various tweaking and modifications for prediction of heart disease.

Choosing a type of data mining technique (i.e. classification, clustering, association, regression) for the research is not the ending of the story. Once the technique is selected, researchers try to improvise the machine learning algorithms or apply the data mining model in a different way to achieve a better prediction result than the existing one. Moreover, the researchers usually compared the performance of the basic algorithms with the improved version of the chosen technique. Besides the top three algorithms, Associative Classification, K-means Clustering, KNN, Fuzzy Rule and SVM are also popular choices among researchers. The most recent trend suggests that researchers are trying to hybridize two types of data mining models, for example clustering and classification, to get better result (Anbarasi et al., 2010; Soni et al., 2011a; Bhatla & Jyoti, 2012b; Shouman et al., 2012b; Devi & Saravanan, 2012; Banu & Gomathy, 2014). Three studies (Shouman et al., 2012b; Devi & Saravanan, 2012; Banu & Gomathy, 2014) showed that combining k-means clustering and decision tree can achieve a better prediction result. In the study conducted by Banu and Gomathy (2014), the researchers

achieved an accuracy of 89% when they built a model using K-mean based MAFIA with decision tree (ID3 and C4). However, over the years, the top three techniques remained the most popular one to be combined with other techniques. This is because the prediction models constructed based on these three popular techniques provide more stable solutions and better results.

2.3.2.4 Model Evaluation

In this review, cross validation is identified as the most popular model validation technique. Cross validation is about dividing the entire set into different subsets. Furthermore, with each turn, one subset acts as a testing set and all others become the training set. The most commonly used technique is 10-fold cross validation, which was used in 10 studies (i.e. S4, S15, S27, S30, S36, S41, S45, S49, S66, S68). On the other hand, there are eight studies (i.e. S3, S7, S15, S39, S62, S63, S68, S74) partitioned the datasets into 70% for training and 30% for testing. On the other hand, some studies divided the datasets with ratios such as 80:20, 50:50, and 90:10. Besides k-fold cross validation, some studies used Random Sub Sampling and Hold-out method. There are

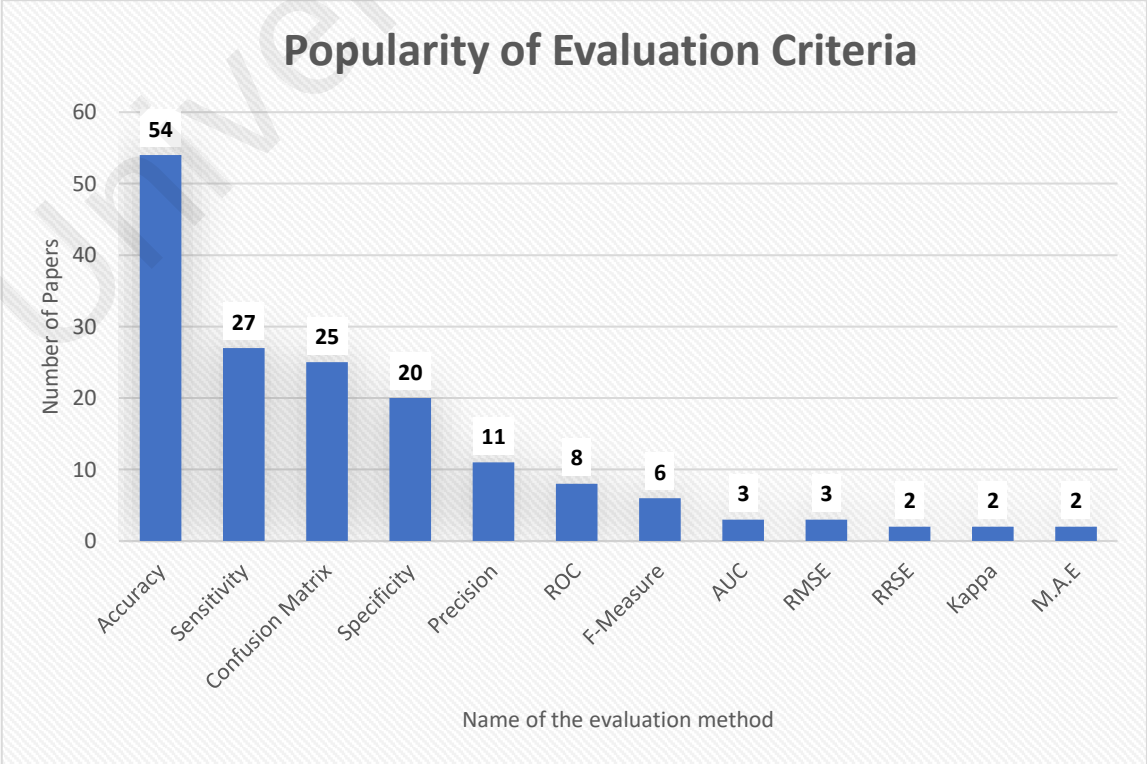

Figure 2.10:Popular evaluation criteria and their selection among the researchers

many evaluation criteria can be selected by the researchers to evaluate the model performance. They usually choose more than one criteria to compare the models created by different data mining techniques. Based on the analysis, 30 evaluation criteria were identified from all the studies. Figure 2.10 shows the evaluation criteria that have been used in at least two studies. The chart represents the number of studies using a criterion. Accuracy is the most popular evaluation criterion and it has been used in 54 studies. Sensitivity, Specificity and Precision are used in 27, 20 and 11 studies respectively to evaluate prediction model. Accuracy, Specificity, Sensitivity and Precision can be acquired from the confusion matrix and 25 studies provided confusion matrix to show the evaluation results. Additionally, there are some studies (i.e. S21 and S22) measured the effectiveness based on the runtime of the algorithms (Jabbar et al., 2011; Bhatla & Jyoti, 2012a).

2.3.2.5 Output

The existing studies created models to predict an outcome. The output often reflects the performance of the proposed models. Outputs may not be the only result of diagnosis or prediction of heart disease; it can often be the selected features or risks prediction in a patient or patterns of heart disease in an area. In this review, we found that the most common output used in the selected studies is predicting whether the patient has a heart disease or not. The second most popular predicted output is the predicting the risk level of heart disease. Besides, some studies also generated fuzzy or associative rules to identify significant factors.

Nevertheless, there are papers that worked with different outcome rather than predicting heart disease. In terms of prediction of heart disease, there are two types of the result being generated; Nominal and Confirmation. Five papers were found to have worked with Feature Selection (M. Jabbar et al., 2011; Peter & Somasundaram, 2012;

Rajeswari et al., 2012; Nahar et al., 2013b; Kavitha & Kannan, 2016) and applied the algorithm which reduces the number of attributes in the dataset but still providing faster execution and improving the accuracy. One of the papers has worked with the length of stay in the hospital as their outcome (Hachesu et al., 2013). One paper was predicting the risk of heart disease in potential patients over the period of 5 years (Bandyopadhyay et al., 2015). At the end of the day, most outcomes are emphasized based on prevention of heart disease.

2.3.2.6 Tool(s) Used

There are various factors affect the selection of data analytics tools to conduct the experiments and product the prediction results. The factors include cost, researcher's familiarity with the tool, machine learning algorithms, features, guidelines and support provided, platform supported, limitations and usability of the tool. In this review, we identified all the tools used by the researchers to predict heart disease in their studies. However, some studies did not specify the tool(s) used in their experiments. Table 2.5 presents all the tools used that we found. The researchers either used non-programming tool(s) and/or developed their own tool(s). Figure 2.11 shows the popularity percentage of the tools used by the researchers.

Among the tools, Waikato Environment for Knowledge Analysis (WEKA) is the most popular non-programming data analytics tool. It is used in 22 studies. This is because Weka is a free, open source and platform independent tool, which provides many machine learning algorithms and features to support the heart disease prediction using different data mining techniques. This java written software is well supported by an established research group in The University of Waikato, New Zealand. The Classifier class can be extended and customised to meet the research needs.

MATLAB has the second highest usage (31%) among researchers. Besides these two tools, RapidMiner, TANAGRA, SPSS, Microsoft SQL Server and KEEL are the other four non-programming data analytics tools used in some studies to create data prediction models. Microsoft SQL Server provides a query language, Data Mining Extensions (DMX) to create and train new prediction models. Other than non-programming tools, researchers also developed their own heart disease prediction tools using programming languages such as C, C#, Java and Phython.

Table 2.5: Data analytics tools used in the selected studies

Tool	Study ID
Weka	**22 studies:** S1, S2, S3, S10, S17, S21, S22, S25, S26, S31, S37, S41, S44, S45, S49, S56, S57, S58, S60, S67, S69, S75
Matlab	**12 studies:** S6, S29, S32, S38, S40, S44, S58, S59, S62, S63, S64, S74
Tanagra	**2 studies:** S4, S68
SPSS	**2 studies:** S46, S63
Microsoft SQL Server (DMX)	**2 studies:** S46, S24
RapidMiner	**1 study:** S47
KEEL	**1 study:** S75
Self-developed new tool	**7 studies:** S5, S11, S12, S19, S24, S72, S76

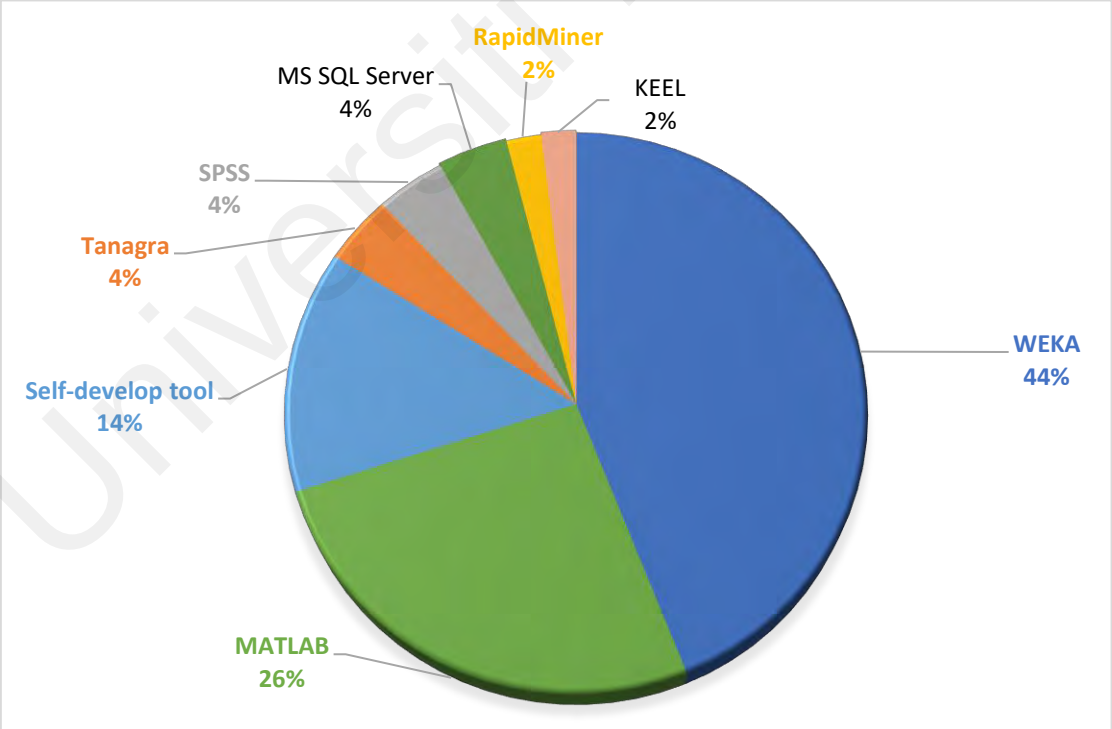

Figure 2.11: Popularity of tools used in the selected studies

2.3.3 Discussion

The selection of the data source depends on the availability and completeness of the datasets. To save time and effort in data collection and data management, most

researchers decided to use a publicly available dataset in their studies. However, the UCI Heart disease dataset was donated in year 1988 and there is no new data been added to the registry. It is considered an outdated clinical dataset that consists of 14 attributes only. Additionally, the researchers should extend their studies to collect data from their own country, rather than using the USA dataset. For example, if the researchers are doing their research in India, then it is quite impractical to collect data from the USA rather than India. To get a more realistic data mining model and accurate predictions, the researchers should collect localised and recent heart disease data from their own country to identify the significant risk factors and improve the healthcare in their country.

It has been proven that feature selection can help to improve the accuracy of the prediction models. Applying feature selection algorithm(s) in an incorrect manner can reduce the performance of data mining technique. In recent years, researchers explored the possibility of combining different types of feature selection methods (e.g. CFS and MFS) to identify medically relevant factors. Nevertheless, the feature selection process is still lacking its importance considering the potential it holds. More research should be done on feature selection using both CFS and MFS to identify significant features that can improve the accuracy of prediction when we have limited heart disease data.

We noticed that hybrid techniques are being explored but there is a need to study more on the combination of different types of data mining techniques that can improve the performance of prediction models. Some machine learning algorithms can be extended to improve the accuracy. 90% of the studies only provide prediction output as whether there is a possibility to have heart disease or not. To have a better risk prediction, more research should be done to provide more analysis that can assist medical experts and clinicians to make informed decision. Some studies on fuzzy or associative rule mining discovered many rules to detect whether the patients is heathy or has risk to have heart disease without the help of medical experts or cardiologists. The accuracy of the

rules is greatly relying on the positive and negative rules generated from the training dataset. However, in most cases, many irrelevant and impractical rules will be generated. The rule-based approach still need to be explored further to get medically relevant rules that can classify the patients into different risk levels.

Most of the data mining tools only provide simple prediction output to show the accuracy of the algorithms or the risk level. Research can be conducted to design and develop specific heart disease prediction tools together with an interactive analytics dashboard that present the results to non-statistician decision makers and users from medical domain. This can help to monitor cardiac patient population by detecting and preventing heart disease.

2.4 Limitation of the Existing Studies

Cardiovascular diseases are the most efficient killer in the current world. Thus, researchers have been putting their efforts in to achieving greater success in controlling its kill counter. Every year we are taking another step towards improving what was the margin last year. Nonetheless the improvements are still something researchers are thriving for.

Researcher are underlying the fact that only techniques are not enough to achieve greater efficiency and so they are putting more focus on other steps like feature selection to see how that yields. A good choice of data mining technique along with a suitable group of features can often get a good result. In (Rajeswari, Vaithiyanathan et al. 2013), feature selection was highly emphasized due to its impact on the algorithms used. A correct combination that suits the technique can give much higher accuracy. In (Arabasadi, Alizadehsani et al. 2017), the researchers found almost 10% increase in

accuracy after applying a feature selection technique. In this context feature selection is becoming a concern for data mining in cardiovascular disease diagnosis.

One of the reason for feature selection being held as important is that the healthcare industry contains noisy data. Noisy data can hinder the result of an algorithm and throw it off its course (Paul, Shill et al. 2016). Nevertheless, just munching off features carelessly will achieve nothing as well. So, researchers in (Nahar, Imam et al. 2013) emphasizes that choosing features must be done with an expert's view. Thus, the researchers also suggested that using a computerized feature selection may lead to features that are irrelevant in the first place. As per the scenario suggests the researchers are looking for balanced features that will have positive impact on the techniques applied. In (Rajeswari, Vaithiyanathan et al. 2013), the researcher suggests that it might be worth a while to go through all the combinations for finding the correct combinations of features that works well with the technique.

In (Khemphila and Boonjing 2011), researcher acknowledge the fact that focusing only on feature selection might not yield enough improvement needed. As they found that even focusing on the feature selection methods techniques still perform somewhere around 80% accuracy. The reason might be the use of the techniques associated with feature groups. The selection of correct technique is equivocally important in this matter. In (Subanya and Rajalaxmi 2014), researchers have suggested a good technique with a correct subset of feature only gives a higher result than other scenarios. The focus they put in is not only in the feature selection but also in the choice of a good technique. It is obvious from the current research that now it has be a package deal if any improvements are to be made in the field where the package being the combination of technique and feature selection.

In (Purusothaman and Krishnakumari 2015), researcher demands from the future researches to come up with an efficient technique that is high in accuracy and

speedy. Using feature selection can speed t up the technique as well as give higher accuracy. However, using ensembles in the techniques can be beneficiary for the technique as well. In (Chaurasia and Pal 2014), an ensemble method was used to detect heart disease. The results were pretty amazing as the ensemble beats all the techniques in accuracy. Ensemble techniques are being used more and more in all other sectors. But in cardiovascular disease prediction not many has seen the light. Vote is such an ensemble. In (Paris, Affendey et al. 2010), vote was used and compared with other techniques. It has outperformed all those techniques in a good manner. But the current researches in cardiovascular diseases haven't recognize its potential yet.

At the end, it is easy to comprehend that most researchers goal was to improve the efficiency of the techniques used by whatever means necessary. As if even 1% accurate result can save up to thousands of lives.

2.5 Summary

This review study has discussed 76 studies that describe the use of data mining techniques to support the heart disease prediction. This review helps us understand the state of the art in applying data mining techniques to improve the accuracy of predicting heart disease. We also identify gaps and future research directions. This review experience helps us to have a better understanding of issues related to the heart disease prediction using data mining. Performing prediction using data mining techniques require appropriate selection of dataset, preprocessing technique, feature selection method, machine learning algorithms, performance evaluation and tools. This will ensure high accuracy of prediction that will benefit the medical experts or clinicians to give advice or decisions based on the results of the prediction.

The use of data mining techniques and tools to support the heart disease prediction is an active area of research. However, although many studies have used many data mining techniques to support the heart disease predictions, there are still limited information provided as the output of the prediction. Moreover, although the preprocessing and feature selection are of great importance in the data mining process, most researchers have neglected these two aspects. Furthermore, the identified tools support the prediction only provide analysis more on performance accuracy, rather than present uncertainties and statistical results to non-statistician decision makers with more understandable information. In this review, we have suggested some future research directions to encourage the researchers to explore different aspects in heart disease prediction. It is crucial to use local heart disease dataset to predict the trend of the cardiovascular diseases based on the cardiac patient population characteristics in that country.

Chapter 3: Research Methodology

3.1 Introduction

This chapter aims to describe the research methodology used in this study. The methodology used to conduct this research is adapted from the CRISP-DM methodology. The chapter describes the entire methodology and how we adapted CRISP-DM to this study. Section 3.2 gives an overview of the CRISP-DM. Section 3.3 describes the adapted methodology for this research and how have we changed to adapt our study. Moreover, this section also describes the phases of the methodology we used. Finally, section 3.4 summarizes the chapter.

3.2 Overview of Crisp DM

The Cross Industry Standard Process for Data Mining (CRISP-DM) was a concept developed many years ago. It is still the most popular methodology applied in data mining projects. In this methodology, the entire data mining process is divided into six phases that go back and forth to satisfy the needs of data mining projects. The six phases of the methodology are: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. It ensures the knowledge discovery process is conducted correctly for a project and reduces the mistakes in each task. It is a very robust and stable methodology (Chapman, Clinton et al. 1999). Overall this methodology serves as a guideline for many data science projects.

3.3 Our Methodology

The methodology has emerged based on CRISP-DM (Azevedo et al., 2008). Figure 3.1 shows the methodology used in this research. This research methodology adapted five phases out of the six phases from CRISP-DM. The five phases are: Data Understanding, Data Preparation, Modeling, Evaluation and Development. In this

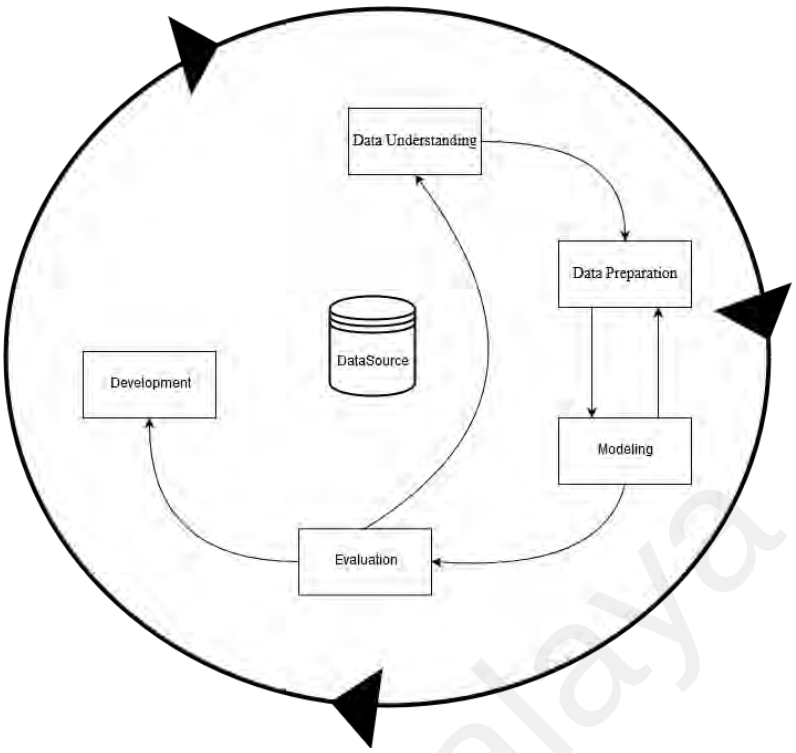

Figure 3.1: Research Methodology adapted from CRISP-DM

research, Business Understanding phase was not included because this is a research project and literature review was conducted to identify research problems, define objectives and scope from research perspectives, not from business perspectives. There is back and forth flow in between Data Preparation and Modeling phases. After performing the evaluation, limitations and constraints of the data mining model are identified to provide feedback to Data Understanding phase. This allows us to develop a prediction model without the fear of getting it wrong. Moreover, if anything goes wrong we can go back to our data understanding phase and reiterate the phases.

3.3.1 Data Understanding

This is the phase where the datasets were analyzed. The heart disease datasets were collected from the data source, UCI Machine Learning Repository and went through the four databases (i.e. Cleveland, Hungary, Switzerland, and the VA Long Beach) to understand the data provided by each database. Finally, Cleveland dataset was selected because it is the commonly used database by machine learning researchers and has the most complete records. Although the database has 76 raw attributes, only 14 subsets of

attributes with more complete values are actually used in the past experiments. In this phase, all the 14 attributes and their presence and impacts on the dataset were analyzed to identify problems in data quality in the 303 records. For instance, the number of records that are male and female in each category, how many of them have high blood pressure, how many of them are diabetic. All the information was gathered from analyzing the dataset by focusing one single attribute called 'sex'. Thus, the phase continued with other attributes as well. After analyzing the entire dataset properly, this phase was ended and ready to move on to the next phase, Data Preparation.

3.3.2 Data Preparation

This phase prepares the data for the experiments. Any missing values are dealt with in this phase. The preparation of the data contains many things like scaling attributes or converting multiple class into binary but none of this is done in an established pattern. These changes can be done more than once to ensure a clean and sound dataset. For this research, 6 records with missing values were removed from the dataset and 297 records are used in the experiments. Additionally, the conversion of the predicted attribute for presence of heart disease in a patient from the multiple values (0 for absence and 1, 2, 3, 4 for presence) to the binary values (0 for absence; 1 for presence of heart disease). After preprocessing the data and setting the type of the attribute, the Cleveland dataset was imported to RapidMiner tool and organize the data for our experiment. After this phase, our data was ready for the experiments.

3.3.3 Modeling

In this phase, seven classification techniques (k-NN, Decision Tree, Naïve Bayes, Logistic Regression, Vote, Support Vector Machine and Neural Network) were applied to create prediction models for this experiment using the prepared dataset. The phase starts with selecting a data mining technique for the model and then perform the

prediction. The results and performance (i.e. accuracy, f-measure and precision) of the data mining technique are recorded. After recording the results, another technique was selected to build the prediction model. The same process was repeated for all techniques. Every classification technique was used with the different combination of features. Furthermore, the results for each combination of features was recorded. Next, all the results were analyzed to identify significant features and data mining techniques in predicting heart disease. Nine significant features and top three classification techniques were selected to create the best performing model. In next phase, evaluation was conducted to validate the top three classification techniques with the selected significant features.

3.3.4 Evaluation

A final barrier before proceeding with the development of the heart disease prediction system. This phase focuses on the evaluation of significant features and classification techniques identified from the previous phase using Statlog heart disease dataset from UCI Machine Learning Repository. This dataset contains 270 records without missing value. All the records were used for the experiments. In the evaluation phase, a performance analysis of the classification techniques along with the combination of different features was performed. All the results were recorded for all combination of techniques. After completing the experiments, a proper analysis was performed to compare the results obtained from the previous phase and the results obtained during evaluation phase. The results were benchmarked with existing studies as well. Finally, the evaluation results show that the features and top three data mining techniques identified during the Modelling phase are significant and appropriate to predict heart disease with acceptable accuracy (more than 87%).

3.3.5 Development

Nine significant attributes and Vote classification technique identified in this research to create a best performing prediction model. During development phase, the model was used to design and develop heart disease prediction system. The system aims to predict the presence of heart disease for a list of patients or a specific patient using the significant attributes. Development phase starts from eliciting and analyzing the functional requirement for this system, followed by the system and interface design. Next, the system was developed using Java programming language. The model identified was applied using WEKA-API on Java SE. Testing was conducted to evaluate the system functionalities.

3.4 Summary

This chapter gives an overview of the research methodology used in this research. Although the main concept was inspired from the CRISP-DM methodology, the methodology has been adapted to suit this research. Some phases were changed and the first phase, Business Understanding was omitted from the original CRISP-DM. The five phases were described to explain the process of conducting the entire experiments for this research project: Data Understanding, Data Preparation, Modeling, Evaluation and Development. Overall these processes were iterated throughout the experiments to ensure the achievement of the research objectives. A more detailed explanation of the experiments is reported in Chapter 4 and Chapter 5.

Chapter 4: Identification of Significant Features and Data Mining Techniques

4.1 Introduction

This chapter discusses the identification of significant features and data mining techniques to predict heart disease in this research. An experiment was conducted to identify the features and data mining techniques. The following subsections describe different phases of the experiments. Section 4.2 describes the data source to collect heart disease dataset used in this research to identify significant features and data mining techniques. Section 4.3 explains the preparation to setup the experiment which includes Data Preprocessing, Feature Engineering, Classification Modelling using Data Mining Technique and Performance Measure. The process of feature engineering is described to illustrate the selection of significant features in heart disease prediction. Section 4.4 presents the results of the experiment, which is the performance evaluation of the model created using seven data mining techniques. Section 4.5 discusses the analysis conducted to identify significant features and data mining techniques to create the best performing model.

4.2 Dataset

The heart disease data were collected from UCI machine learning repository (Lichman, 2013) There are four databases (i.e. Cleveland, Hungary, Switzerland, and the VA Long Beach)), the Cleveland database was selected for this research because it is the commonly used database by machine learning researchers and has the most complete records.. The dataset contains 303 records. Although the Cleveland dataset has 76 attributes, the dataset provided in the repository only provides information for a subset of 14 attributes. The data source of the Cleveland dataset is Cleveland Clinic Foundation. Table 4.1 describes the 14 attributes and their attribute types. Based on the table, we can

see that there are 13 attributes that feature in heart disease prediction and one attribute serves as the output or prediction. predicted attribute for presence of heart disease in a patient

The Cleveland dataset contains an attribute named ‘num’ to show the diagnosis of heart disease in patients on different scales from 0 to 4. In this scenario 0 being the absence of the heart disease and all the values from 1 to 4 represent patients with heart disease where the scaling refers to the severity of the disease (4 being the highest). Figure 4.1 shows the distribution of ‘num’ attribute among the 303 records.

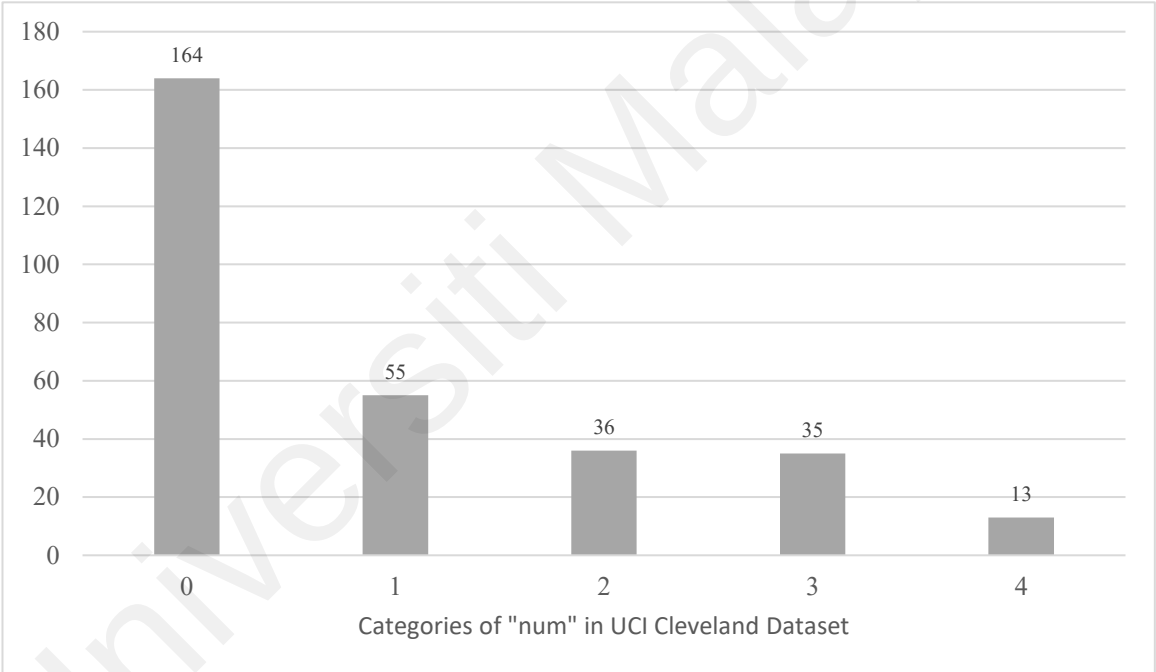

Figure 4.1: Distribution of "num" in UCI Cleveland Dataset

Table 4.1: Description of attributes from the UCI Cleveland Dataset

Attribute	**Description**	**Type**
Age	Age of the patient in years	Numeric
Sex	Gender of the patient (1 for male and 0 for female)	Nominal
Cp	Chest pain type described with 4 values; Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic	Nominal

Trestbps	Resting blood pressure (in mm/Hg on admission to the hospital)	Numeric
Chol	Serum cholesterol in mg/dl	Numeric
Fbs	Fasting blood sugar > 120 mg/dl; 1 if true and 0 if false	Nominal
Restecg	Resting electrocardiographic results in 3 values; Value 0: normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria	Nominal
Thalach	Maximum heart rate achieved	Numeric
Exang	Exercise induced angina (1 for yes and 0 for no)	Nominal
Oldpeak	ST depression induced by exercise relative to rest	Numeric
Slope	The slope of the peak exercise ST segment Value 1: upsloping Value 2: flat Value 3: downsloping	Nominal
Ca	Number of major vessels (0-3) coloured by fluoroscopy	Numeric
Thal	The heart status described with 3 values Value 3: normal Value 6: fixed defect Value 7: reversable defect	Nominal
Num	It represents the diagnosis of heart disease with 5 values. 0 meaning absence and 1 to 4 means presence of heart disease.	Nominal

4.3 Experimental Setup

In this research, RapidMiner Studio was used to conduct the experiment. RapidMiner Studio is used because it provides a robust and easy to use visual design environment for building predictive analytic workflows. The visual representation of the workflow is one of the efficient features for the beginners. Moreover, its support towards open source innovation, availability and effective functionality. Figure 4.2 shows the flowchart for the experiment. In the experiment, the UCI Cleveland heart disease dataset was imported into RapidMiner. The data mining process starts from -pre-processing

phase, followed by the feature engineering to select different combination of attributes and classification modelling to create models for prediction using data mining techniques. The feature selection and modelling were repeated for all the combination of attributes . The loop iterates as a subset containing minimum 3 attributes are chosen from the 13 attributes and applied the model to it. The performance of each model created based on the selected attributes and data mining technique during each iteration is recorded and the output of the results is shown after the entire process is completed. Section 4.3.1 to Section 4.3.3 describes more detailed on the data preprocessing, feature selection and classification modelling.

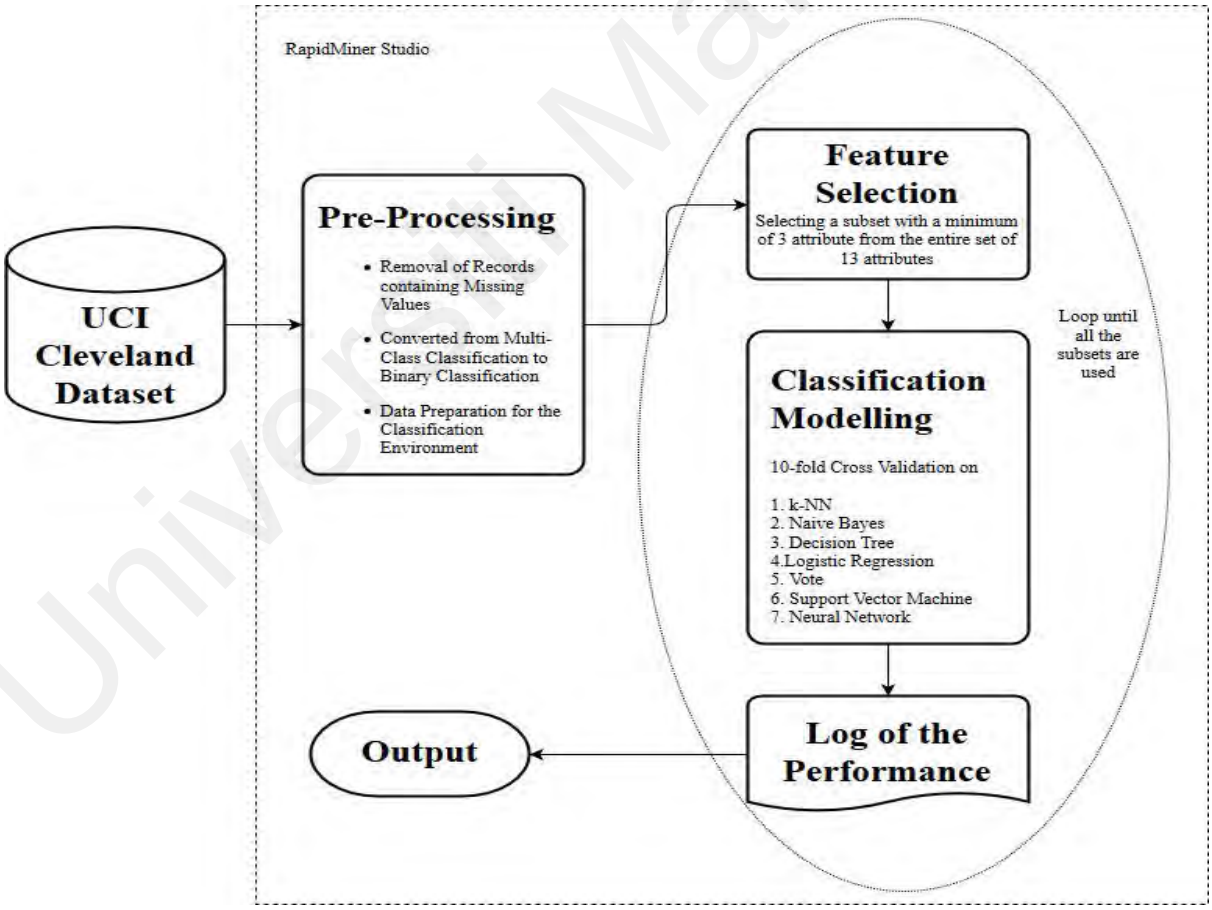

Figure 4.2: Flowchart of Experiment

4.3.1 Data Preprocessing

The data were preprocessed after collection. There were 6 records that have missing values in Cleveland dataset. All the records with missing values were removed

from the dataset thus reducing the number of records from 303 to 297. Next, the values of predicted attribute for presence of heart disease in the dataset was transformed from multiclass values (0 for absence and 1, 2, 3, 4 for presence) to the binary values (0 for absence; 1 for presence of heart disease). The data preprocessing task was performed by converting all the diagnosis values from 2 to 4 into 1. The resulting dataset thus contains only 0 and 1 as the diagnosis value where 0 being the absence and 1 is the presence of heart disease. After the reduction and transformation, the distribution of 297 records for 'num' attributes become 160 records for '0' and 137 records for '1'.

4.3.2 Feature Selection

Among the 13 features used in heart disease prediction, only 'age' and 'sex' is personal information of the patient. The remaining 11 features are all test based scores collected from various medical examinations. In this experiment, a combination of features was selected to be used with 7 classification techniques; k-NN, Decision Tree, Naïve Bayes, Logistic Regression, Vote, Support Vector Machine and Neural Network to create the classification model. For this purpose, brute force technique was applied to limit its lower bound (minimum 3 features). The procedure was to test each possible combination of features with all the techniques. In the experiment, firstly, all possible combination of 3 features from the 13 attributes were chosen and each combination was tested by applying the 7 data mining techniques. Next, the experiment was repeated to select all possible combination of 4 features from 13 attributes.

The total number of combination achievable from a set of 13 attributes excluding the empty set is represented by $2^n - 1$. In this research, a single subset of the combination of features cannot have less than 3 attributes. Thus, all the subsets of combination achieved by having 2 attributes and 1 attribute are omitted. The equation used to calculate the total number of combinations is derived as follows.

Total number of combination

$$=2^{n}-\left(\frac{n!}{1!(n-1!)}\right)-\left(\frac{n!}{2!(n-2)!}\right)-1$$

$$=2^{n}-n-\frac{n\,(n-1)}{2}-1$$

$$=2^{n}-(\frac{2n+n^{2}-n}{2}+1)$$

$$=2^{n}-(\frac{n^{2}+n}{2}+1)$$

where n represents the total number of features used to generate the subsets of combination, which is 13 for our experiment. Thus, a total of 8100 combinations of the feature was selected and tested in this experiment.

4.3.3 Classification Modelling using Data Mining Technique

After selecting the features, the models were created with the 7 most popular classification techniques in data mining: k-NN, Decision Tree, Naive Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Neural Network and Vote (a hybrid technique with Naïve Bayes and Logistic Regression). 10-folds cross validation technique was used validate the performance of the models. In this technique, the entire dataset is divided into 10 subsets and then processed 10-times. 9 subsets are used as a testing set and remaining 1 subset is used as training set. Finally, the results are shown by averaging each 10 iterations. The subsets are divided using stratified sampling meaning that each subset will have the same class ratio of the main dataset.

4.3.4 Performance Measure

The performance of the classification models was measured with three performance measures: Accuracy, f-measure and precision. Accuracy is the percentage of correctly predicted instances among all instances. F-measure is the weighted mean of

the precision and recall. Precision is the percentage of correct predictions for the positive class.

To identify the significant features, these three performance measures was used whereas to identify data mining technique for creating best performing models, the accuracy and precision measures was used. For identification of significant features, the three performance measures give a better understanding of the overall behaviour of the different combination of features. On the other hand, analysis of data mining techniques focusses on the best performing models that can produce high accuracy in heart disease prediction because accuracy and precision are the most intuitive evaluation metrics on performance. For each classifier, performances have been measured separately and all the results are recorded properly for further analysis.

4.4 Results

The performance of 7 data mining techniques on 8100 combinations of features were experimented one by one. The experiment was conducted using RapidMiner tool to measure the accuracy, precision and f-measure of each model, all the experiment results were gathered for further analysis. Table 4.2, Table 4.3 and Table 4.4 describe the highest accuracy, highest precision and highest f-measure achieved by each data mining technique and also the combination of features used in the model.

Table 4.2: Highest accuracy achieved by data mining techniques

Technique	**Accuracy**	**Combination**
Support Vector Machine (SVM)	86.87%	age, sex, cp, chol, fbs, exang, oldpeak, slope, ca
Vote	86.20%	sex, cp, fbs, thalach, exang, slope, ca, thal
Naïve Bayes	85.86%	sex, cp, thalach, exang, oldpeak, ca
Logistic Regression	85.86%	age, sex, cp, chol, restecg, oldpeak, slope, ca, thal
Neural Network	84.85%	sex, cp, trestbps, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal

k-NN	82.49%	sex, cp, fbs, restecg, oldpeak, ca, thal
Decision Tree	82.49%	sex, cp, fbs, restecg, oldpeak, ca, thal

Table 4.3: Highest precision achieved by data mining techniques

Technique	**Precision**	**Combination**
k-NN	95.00%	sex, restecg, exang
Decision Tree	95.00%	sex, restecg, exang
Vote	90.27%	cp, trestbps, fbs, thalach, exang, oldpeak, slope, ca, thal
Naïve Bayes	87.92%	sex, cp, fbs, slope, ca, thal
Support Vector Machine (SVM)	86.86%	sex, cp, chol, slope, ca
Neural Network	86.43%	sex, ca, thal
Logistic Regression	86.42%	sex, cp, trestbps, thalach, exang, oldpeak, slope, ca, thal

Table 4.4: Highest f-measure achieved by data mining techniques

Technique	**F-measure**	**Combination**
Support Vector Machine (SVM)	88.22%	age, sex, cp, chol, fbs, exang, oldpeak, slope, ca
Naïve Bayes	87.35%	sex, cp, thalach, exang, oldpeak, ca
Logistic Regression	87.27%	age, sex, cp, chol, restecg, oldpeak, slope, ca, thal
Neural Network	85.98%	sex, cp, trestbps, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal
Vote	84.41%	sex, cp, fbs, thalach, exang, slope, ca, thal
k-NN	84.05%	sex, cp, fbs, restecg, oldpeak, ca, thal
Decision Tree	84.05%	sex, cp, fbs, restecg, oldpeak, ca, thal

The three tables describe the performance of 7 classification techniques in three different categories. Based on the analysis showed in the tables, we can see that highest accuracy (86.87%) is achieved by SVM with 9 attributes. On the other hand, the highest precision (95.00%) was achieved by both Decision Tree and k-NN using the same combination of 3 features (i.e. sex, restecg, exang) whereas the highest f-measure was

given by SVM with 9 attributes. Table 4.2 shows the top three best performing techniques with accuracy more than 90% are SVM, Naïve Bayes and Vote. The results also indicate that both Decision Tree and k-NN has the lowest accuracy (82.49%) as compared to other techniques. However, these two techniques gave highest performance in precision. Table 4.5 shows the average accuracy achieved by each technique on all the 8100 combinations of features. Table 4.6 and Table 4.7 shows the average precision and f-measure of each technique.

Table 4.5: Average accuracy achieved by data mining techniques

Technique	**Average Accuracy Achieved**
Vote	78.20%
Naïve Bayes	78.20%
Support Vector Machine (SVM)	78.15%
Logistic Regression	78.03%
Neural Network	75.18%
k-NN	63.50%
Decision Tree	63.50%

Table 4.6: Average precision achieved by data mining techniques

Technique	**Average Precision Achieved**
Vote	79.41%
Naïve Bayes	78.76%
Support Vector Machine (SVM)	78.15%
Logistic Regression	76.27%
Neural Network	76.20%
k-NN	66.43%
Decision Tree	66.43%

Table 4.7: Average f-measure achieved by data mining techniques

Technique	**Average F-measure Achieved**
Logistic Regression	80.98%
Support Vector Machine (SVM)	80.25%
Naïve Bayes	80.17%
Vote	78.10%
Neural Network	77.33%
k-NN	65.87%
Decision Tree	65.87%

Based on Tables 4.5, Vote, Naïve Bayes and SVM are the top three techniques for all the 8100 combinations by getting an average of 78.20% and 78.15% in accuracy. On the other hand, the average values of precision shown in Table 4.6 indicate that Vote,

Naïve Bayes and SVM are the top three techniques. According to Table 4.7, the top three techniques that have achieved highest average F-score are LR, SVM and Naïve Bayes.

4.5 Analysis of Features and Data Mining Technique

This section describes the selection of significant features and data mining techniques based on the results obtained from the experiments. By analyzing these results, the significant features and data mining technique that have significant impacts in creating best performing models are identified to predict heart disease. Section 4.5.1 and Section 4.5.2 discusses the significant features, and best performing data mining techniques selected in this research.

4.5.1 Feature Selection

The result achieved from experiments were analyzed to identify the significant attributes. In order to identify the significant attributes, an analysis was conducted to find out how many times an attribute was selected in the model that has performed the highest accuracy, precision and F-measure. Table 4.8 shows the analysis of attributes that have resulted in best performances on all the data mining techniques. Among all the 8100 combinations, there is 1 combination of features for every technique that has resulted the highest performance for that specific technique. Thus, 7 techniques have different combinations that resulted in highest performance. In this table, an attribute that has occurred in those highest performing combination has been counted and compared with other attributes. The first row of the Table 4.8 depicts how many times each of those attributes found among the combinations that resulted the highest accuracy among the 7 techniques. Similarly, the second and third rows depict the occurrence of the attributes which gives the highest performing F-measure and precision. Lastly, a summation of all occurrences of each attribute are calculated.

Among all the 13 attributes, 'sex' is the attributes that has the highest number of total occurrence that appear 21 times in all the combinations. This indicates that this attribute is the most significant attribute that has impact on predictions that give high accuracy, F-measure and precision. In this research, the attributes that have appeared at least 10 times in resulting the highest performance are identified as significant features in heart disease prediction. Based on the analysis in Table 4.8, 9 attributes are identified as significant features in prediction: "sex", "cp", "fbs", "restecg", "exang", "oldpeak", "slope", "ca" and "thal". Furthermore, based on the experiment results, these nine attributes have been used in the highest accuracy models created using 4 or more data mining techniques.

Table 4.8: Comparison between Attributes resulting highest performance

	Age	Sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
Occurrence in Highest Accuracy	2	7	7	1	2	5	4	3	4	6	4	7	5
Occurrence in Highest F-measure	2	7	7	1	2	5	4	3	4	6	4	7	5
Occurrence in Highest Precision	0	6	4	2	1	2	2	2	4	2	4	5	4
Total Number of Occurrence	4	20	18	4	5	12	10	8	12	14	12	19	14

4.5.2 Data Mining Technique Selection

For the completion of our proposed model, we require data mining technique to go with the selected significant attributes. In this research, top three data mining techniques are identified according to highest average accuracy and precision obtained from the experiments. According to Table 4.5, the top three best performing data mining techniques in terms of highest average accuracy and precision are selected. These three techniques are Vote, Naïve Bayes and SVM. In order to finalize the choice of the top

three techniques, Table 4.2. and Table 4.3 were cross checked and the results show that these three techniques has appeared as one of the top three or top four techniques in highest accuracy and precision. Thus, Vote, Naïve Bayes and SVM are selected as the best three data mining techniques in this research that can be applied to create high performance models.

4.6 Summary

This chapter has presented the methods used to conduct the experiments and discussed the results achieved in the experiments. The chapter provides a thorough description of the dataset used in the experiment. It also explained the preprocessing and feature selection done with the datasets. Moreover, the entire process of analyzing the experiment results and identifying the significant attributes and data mining technique was described in detail. All the experiments were conducted using RapidMiner studio. Nine significant attributes: “sex”, “cp”, “fbs”, “restecg”, “exang”, “oldpeak”, “slope”, “ca” and “thal” and top three best performing techniques: Vote, Naïve Bayes and SVM, were identified based on the analysis of the experiment results.

Chapter 5: Evaluation

5.1 Introduction

In the previous chapter, 9 significant attributes and top three data mining techniques are identified. This chapter describes the process of two-phase evaluation and discusses the outcomes of the evaluation. The objective of this evaluation is to validate the identified significant attributes and data mining techniques. The evaluation was performed in two phases. The first phase conducted an experiment using another dataset, UCI Statlog Heart disease dataset to confirm the findings. To analyze how the best data mining technique achieve its high-performance levels, the second phase benchmarks the highest accuracy achieved by the best technique identified from the first phase against the highest accuracy achieved in the existing studies.

5.2 Dataset

The Statlog heart disease dataset was selected to run the experiments for phase one evaluation. Same as Cleveland dataset, it was collected from UCI machine learning repository (Lichman, 2013). The structure of the Statlog heart disease dataset is similar to the Cleveland heart disease dataset. Table 5.1 shows the comparison between Statlog and Cleveland datasets. Both datasets has 13 attributes that features the heart disease and 1 predicted attribute to show the presence of heart disease. All the names of the attributes are same. The only difference in the attributes between the two datasets is the value that they used to represent the class attribute, “num”. The output for the Statlog dataset contains two values: 1 and 2. “1” is the absence of heart disease and “2” is the presence of heart disease in patient. On the other hand, The Cleveland dataset has five different levels of “num” scaling from 0 to 4. That is the only difference between the attributes from the Cleveland dataset and Statlog dataset.

In the previous experiment, Cleveland dataset was converted from the multiclass values for 'num' into binary values (0, 1). This overcome the difference of 'num' values between the two datasets. After preprocessing and converting the Cleveland data into binary data, , the distribution of 297 records for 'num' attributes becomes 160 records for '0' and 137 records for '1'. On the other hand, as seen in Table 5.1, the Statlog dataset contains a total of 270 records. The dataset does not contain any missing values. The distribution of "1" and "2" as the value of "num" is 150 and 120 respectively. Since there is no missing values, the dataset did not require much preprocessing before using the data in this evaluation. Furthermore, the ratio of the number of records for absence and presence of heart disease is almost the same (i.e. Cleveland's ratio =160: 137; Statlog's ratio = 150: 120).

Statlog dataset has a very clean representation of the records which has increased its popularity among the researchers as well. Many recent studies (Subbulakshmi et al., 2015, Nahato et al., 2015, Bashir et al., 2015, Srinivas et al., 2015) have used Statlog dataset in their experiments.

Table 5.1: Comparison between Cleveland and Statlog dataset without preprocessing

Comparison Category	**Cleveland Dataset**					**Statlog Dataset**	
No. of Attributes	13					13	
Attributes	age, sex, cp, trestbps, chol, fbs, restecg, exang, oldpeak, slope, ca, thal					age, sex, cp, trestbps, chol, fbs, restecg, exang, oldpeak, slope, ca, thal	
Class Attribute	num					num	
Different values for "num"	0,1,2,3,4					1,2	
Distribution of "num"	0	1	2	3	4	1	2
	164	55	36	35	13	150	120
Records with Missing Values	6					0	
Total number of instances	303					270	

Due to all the similarity and the quality of the data, the Statlog dataset was identified as the best dataset to be used in our evaluation to validate the proposed significant attributes and data mining techniques.

5.3 Experimental Setup

This subsection describes the experiment conducted to evaluate the significant features and top three classification techniques proposed in this research. The nine significant attributes are sex, cp, fbs, restecg, exang, oldpeak, slope, ca and thal. Furthermore, the top three data mining technique are Vote, Naïve Bayes and Support Vector Machine (SVM). Figure 5.1 shows the entire process of data mining used to conduct the experiment for the evaluation. This experiment was performed with the help of RapidMiner Studio as well. The Statlog heart disese dataset was downloaded from the UCI Machine Learning repository and was loaded on RapidMiner.

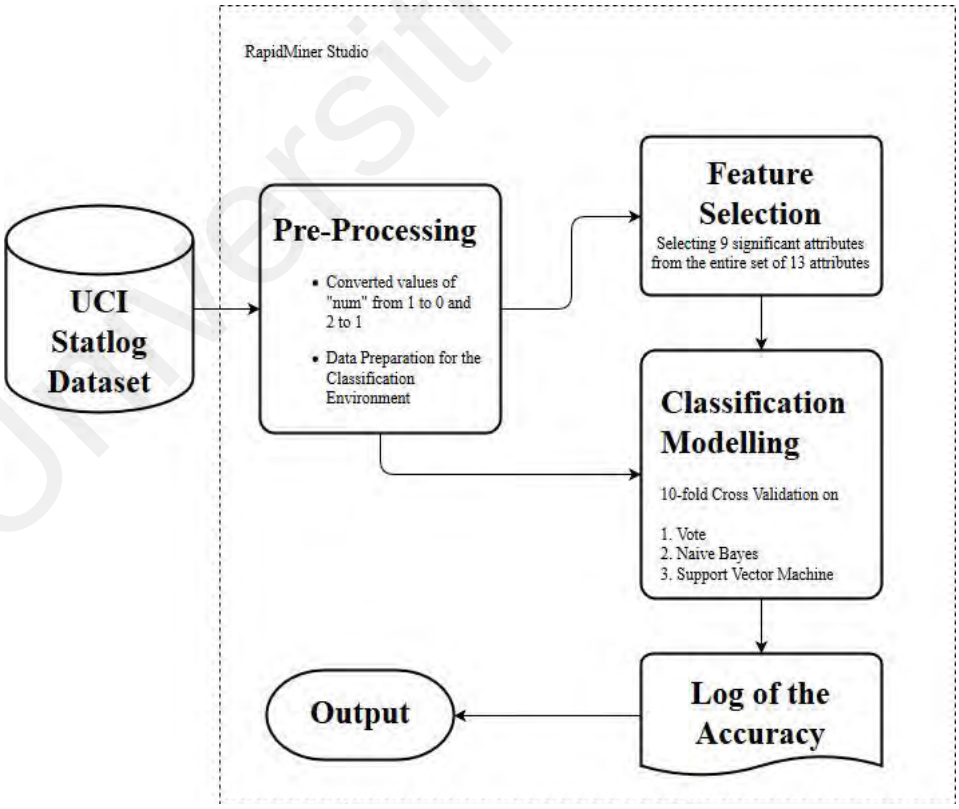

Figure 5.1: Flowchart of the Experiment for Evaluation

5.3.1 Data Preprocessing

First, the Statlog data were preprocessed before using the dataset for evaluation. To make the Statlog dataset similar to the Cleveland dataset, class value of "num" was converted from "1" to "0" and from "2" to "1". The resulting dataset thus contained 0 and 1 as the predicted output values where 0 is the absence and 1 is the presence of heart disease. After the transformation, the distribution of 270 records becomes 150 instances for '0' and 120 instances for '1'. The data was then ready to be used for the classification environment.

5.3.2 Feature Selection

The experiment used the dataset together with the top three techniques in two ways. Firstly, the dataset was passed to modelling without any feature reduction. In this stage, the dataset with 13 attributes were tested on the top three techniques. In the second stage, the nine significant features identified in Chapter 4 (i.e. sex, cp, fbs, restecg, exang, oldpeak, slope, ca and thal) were selected from the preprocessed dataset. These nine attributes were used in the first phase of evaluation to validate the performance of the classification models created using the top three data mining techniques.

5.3.3 Classification Modelling

The classification models were developed using the 3 top data mining techniques (i.e. Vote, Naive Bayes and Support Vector Machine) for all features and also the nine significant features identified in Chapter 4; In this experiment,10-folds cross validation technique was used to measure the performance of the models. This technique, partitioned the entire dataset into 10 equal size subsets (9 subsets for training and 1 subset for testing) and then processed for 10-times. Each time a subset is used as a testing set and remaining other subsets are used as training sets. Finally, the results are shown by averaging the

results obtained from the 10 iterations. The subsets are divided using stratified sampling meaning that each subset will have the same class ratio of the main dataset.

5.3.4 Performance Measure

The evaluation of the model was performed with the help of confusion matrix. The confusion matrix is a kind of a table that is often used to describe the performance of a classification model. It is one of the commonly used methods for performance evaluation. Table 5.2 shows the confusion matrix that is used in our evaluation. There are four outcomes based on this table: True Positive, True Negative, False Positive and False Negative. Each of them describes the relation between the predicted value and the actual value. True Positive refers to correctly predicted value, where the prediction value is 'yes' and the actual value is 'yes' too. Likewise, False Positive refers to the value predicted incorrectly, where the prediction value is 'yes' but the actual value is 'no'. It goes the same way for both True Negative and False Negative. Based on the confusion matrix, different formulas can be used to measure different evaluation criteria. In this research, this formula was used to measure the accuracy of the classification models: Accuracy = (TP+TN)/N (Powers & Martin, 2011). Accuracy is selected because it is one of the most popular and intuitive criteria used in many existing studies to evaluate the performance of classification models.

Table 5.2: Confusion Matrix

N=Total Number of Instances	**Predicted Yes**	**Predicted No**
Actual Yes	True Positive (TP)	False Negative (FN)
Actual No	False Positive (FP)	True Negative (TN)

5.3.5 Results

The confusion matrix for the data mining techniques (Vote, Naïve Bayes and SVM) are shown through Table 5.3-5.8. According to the tables, from a total of 270 instances of we applied the formula described in previous section to measure the accuracy of the data mining techniques.

Accuracy of Vote with 13-attribute dataset= (103+130)/270=236/270=0.8630

Accuracy of NB with 13-attribute dataset = (97+130)/270=236/270= 0.8407

Accuracy of SVM with 13-attribute dataset = (89+133)/270=236/270=0.8222

Accuracy of Vote with 9-attribute dataset = (100+136)/270=236/270=0.8741

Accuracy of NB with 9-attribute dataset = (97+132)/270=236/270=0.8481

Accuracy of SVM with 9-attribute dataset = (94+136)/270=236/270=0.8519

Table 5.3: Confusion Matrix of Vote with 13-attribute dataset

N=270	**Predicted "1"**	**Predicted "0"**
Actual "1"	103	17
Actual "0"	20	130

Table 5.4: Confusion Matrix of Naïve Bayes with 13-attribute dataset

N=270	**Predicted "1"**	**Predicted "0"**
Actual "1"	97	23
Actual "0"	20	130

Table 5.5: Confusion Matrix of SVM with 13-attribute dataset

N=270	**Predicted "1"**	**Predicted "0"**
Actual "1"	89	31
Actual "0"	17	133

Table 5.6: Confusion Matrix of Vote with 9-attribute dataset

N=270	Predicted "1"	Predicted "0"
Actual "1"	100	20
Actual "0"	14	136

Table 5.7: Confusion Matrix of Naïve Bayes with 9-attribute dataset

N=270	Predicted "1"	Predicted "0"
Actual "1"	97	23
Actual "0"	18	132

Table 5.8: Confusion Matrix of SVM with 9-attribute dataset

N=270	Predicted "1"	Predicted "0"
Actual "1"	94	26
Actual "0"	14	136

Table 5.9 describes the accuracy obtained from the experiment. This table shows the accuracy for the entire dataset with 13 attributes and the accuracy for the dataset that contains 9 significant attributes. Based on Table 5.3, it shows that the accuracy for the dataset with 9 attributes has performed better than the dataset with the 13 attributes. The highest accuracy in the 13-attribute dataset was achieved by vote (86.30%). Additionally, the highest accuracy for the 9-attribute dataset was also achieved by the vote (87.41%).

Table 5.9: Accuracy obtained from the experiment for evaluation

	Vote	Naïve Bayes	Support Vector Machine
Accuracy obtained with 13 attributes	86.30%	84.07%	82.22%
Accuracy obtained with identified 9 significant attributes	**87.41%**	84.81%	85.19%

5.3.6 Discussion

The results presented in Table 5.3 indicate that the identified significant features have improved the accuracy of all the three top data mining techniques. This confirms the findings in Chapter 4 on the significant attributes in heart disease prediction. Based on

Table 5.3, the Vote is the best data mining technique that has achieved the highest accuracy, 87.41% using the 9 significant attributes proposed in this research. Since Vote outperforms the other two techniques in the second experiment and show consistent accuracy for both experiments, the Vote is identified as the best performing technique among the top three techniques. Vote technique is used to create a classification model that will be designed and implemented as an intelligent heart disease prediction system (see Chapter 6 for more details). The highest accuracy achieved by Vote technique is used in the second phase of evaluation to benchmark the accuracy with the results achieved by existing studies. According to the evaluation results in phase one, a classification model

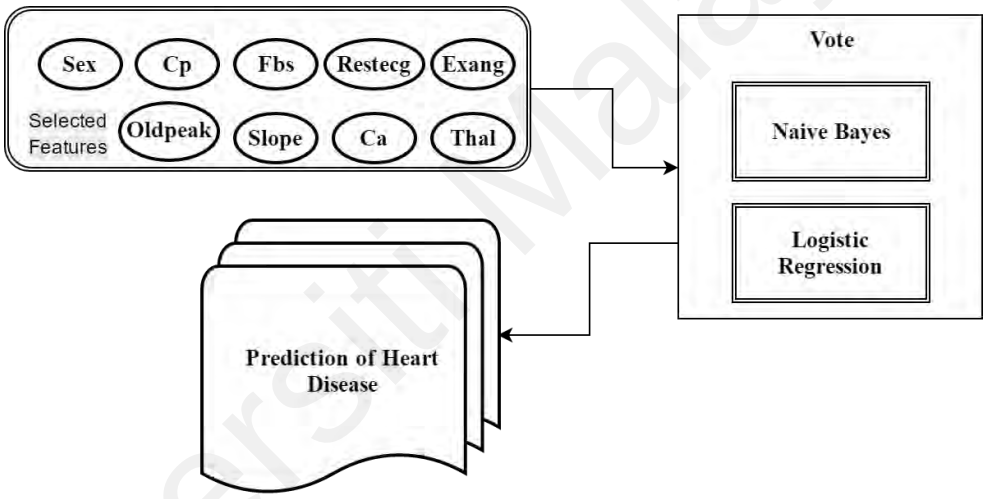

Figure 5.2: The Proposed Model

was proposed in this research using the 9 significant attributes (sex, cp, fbs, restecg, exang, oldpeak, slope, ca and thal) and and the Vote technique. Figure 5.2 shows the proposed model according to the identified features and data mining technique.

5.4 Benchmarking of the Proposed Model

In phase two, evaluation was conducted on the proposed model using the benchmarking technique. Benchmarking is comparing one's performance against the performance achieved in existing work. This helps to evaluate whether the proposed model has achieved acceptable accuracy as compared to the accuracy achieved by other

studies. In this section, the proposed model was benchmarked against the other six studies that have been conducted in the recent years using UCI machine learning repository.

In this benchmarking, only the accuracy of the proposed model with Vote is included since it has outperformed the other two data mining techniques we used in the previous experiment. So, the benchmarking will only focus on that specific proposed model with the vote technique. Table 5.3 shows the benchmarking of the performance of the proposed model against the performance of the models in the six existing studies. Based on Table 5.3, we can see that the proposed model has performed better as compared to the existing studies. The most recent researches were conducted in 2016 (Paul et al., 2016; Verma et al., 2016), the accuracy is 80% and 80.68%. The difference is very astounding when it comes to the proposed model. Based on the comparison, it is apparent that this research has generated higher accuracy using the hybrid technique, Vote, as compared to other classification techniques used in existing studies.

In the research conducted by (Paul et al., 2016), Neural Network was applied with fuzzy to ensure greater accuracy. However, the accuracy they achieved was 80% which was good comparing to other studies using Neural Network. In (Verma, Srivastava and Negi, 2016), a hybrid feature selection method was applied to identify risk factors which is used for the improvement on the techniques. They combined the correlation based feature subset (CFS) selection with particle swam optimization (PSO) search method and K-means clustering algorithms. The accuracy has increased 11.4%. The highest accuracy was obtained using this feature selection method and the decision tree algorithms 80.68%. In (Ismaeel, Miri, Sadeghian, & Chourishi, 2015), the authors tried a new technique named extreme learning machine to predict heart disease and their proposed technique has obtained an accuracy of 86.50%.

In the study conducted by (El-Bialy et al., 2015), the authors work on improving the feature selection by applying machine learning analysis on different heart disease

datasets to reduce the inconsistency, missing data problems during the data collection. They used the selected features with decision tree technique on the dataset which resulted in the accuracy of 78.54%. On the other hand, (Subanya & Rajalaxmi, 2014) used artificial bee colony algorithm to identify features in heart disease dataset. Later they used the identified features with SVM and has achieved 86.76% accuracy. (Nahar et al., 2013) introduced a feature selection method based on MFS (Medical based Feature Selection). They used the feature selected by medical experts in their models and obtained an accuracy of 69.11% using Naïve Bayes algorithm Although the accuracy is not high it gave insights of applying a new feature selection method to select medical relevant attributes.

In (Chaurasia and Pal, 2013), the authors have developed prediction models using Classification and Regression Tree (CART). The highest accuracy achieved by their model is 83.49%. (Khemphila and Boonjing, 2011) proposed a classification model using neural network that trains the classification using genetic algorithm. Their model is faster in comparison to other neural network models since it does not use the usual backpropagation algorithm. The model only used 8 out of 13 attributes and has achieved the accuracy of 80.99%. In the last study (Shouman, Turner, & Stocker, 2011), the authors applied different types of decision trees algorithms with different combination of features and weighting in heart disease diagnosis to get better accuracy. Their experiment yielded best result in decision tree with gain ratio for feature selection. The accuracy obtained by the technique is 84.10%.

Bechmarking against the six recent studies proves that this research has successfully identified significant features and data mining technique that outperform other existing studies. Additionally, the classification model proposed in this research has proved to have acceptable accuracy and perform better than other studies. The nine

significant attributes and the Vote can be further explored in the future work to improve prediction in heart disease using data mining.

Table 5.10: Benchmark of the Proposed Model

Source	Technique Used	Accuracy Achieved
The proposed model	Vote with Naïve Bayes and Logistic Regression	**87.41%**
Paul, Shill, Rabin, & Akhand, (2016)	Neural Network with Fuzzy	80%
Verma, Srivastava, & Negi, (2016)	Decision Tree	80.68%
Ismaeel, Miri, Sadeghian, & Chourishi, (2015)	Extreme Learning Machine	86.50%
El-Bialy, Salamay, Karam, & Khalifa, (2015)	Decision Tree	78.54%
Subanya & Rajalaxmi, (2014)	SVM	86.76%
Chaurasia & Pal, (2014)	Bagging with J48	85.03%
Nahar, Imam, Tickle, & Chen, (2013)	Naïve Bayes	69.11%
Chaurasia & Pal, (2013)	CART	83.49%
Khemphila & Boonjing, (2011)	Neural Network with Genetic Algorithm	80.99%
Shouman, Turner, & Stocker, (2011)	Decision Tree with Gain Ratio	84.10%

5.5 Summary

This chapter describes the entire evaluation process starting from dataset to finishing with benchmarking. The Statlog dataset was used for the evaluation purpose. The proposed model was trained with the identified significant attributes. The model was evaluated and has achieved better results than other existing studies. In the evaluation of the results obtained from the experiment, it was clear that this research has identified significant attributes and data mining techniques that increase the accuracy of heart disease prediction. According to the benchmarking, it has confirmed the high accuracy compared to other existing studies. The identified data mining technique from the proposed model has performed well paving a way for future research on it. Based on the results, the most appropriate data mining techniques were selected to develop a heart disease prediction system.

Chapter 6: System Analysis, Design and Implementation

6.1 Introduction

This chapter aims to describe the process of design and develop an intelligent heart disease prediction system based on the proposed classification model using nine significant features and Vote technique. Section 6.2 presents the tool and functional requirements. A use case diagram was used to illustrate the system functionalities. Section 6.3 shows the system architecture and system design using layered architecture, activity diagram, class diagram. . Screenshots of the user interface design are presented too. Section 6.4 describes the testing of the system.

6.2 System Requirements

This section provides an overview of the basic system requirements. The proposed model is designed and developed into a heart disease prediction system in a Java Standard Edition (SE) environment.

6.2.1. Tool Requirement

Table 6.1 describes the tools used in the development of the system. The system was developed as a desktop based application. The development of the system was carried out using Netbeans IDE. Moreover, the system uses WEKA API for the application of machine learning algorithms. For the training purpose, UCI Cleveland Heart Disease dataset was used to train the data.

Table 6.1: Tools used in System Development

Category	Software used
Operating System	Windows 10
Programming Language	Java
Integrated Development Environment (IDE)	Netbeans
Library for Data Mining	Weka api
Dataset for Training	UCI Cleveland

6.2.2. Functional Requirement

The functional requirements of the system is described in this section. Functional requirement defines what the system should do. For example, a functional requirement for a glass would typically be holding water without leaking. Similarly, all systems have functional requirements which define its basic behavior. The functional requirements of this system are defined as follows;

F1. System shall provide the prediction of heart disease for a single patient

F2. System shall provide the prediction of heart disease for an entire dataset

F3. System shall provide the performance evaluation results of a given dataset

6.2.3. Use Case Diagram

Use case diagram depicts what system can or should do from the user perspective. Figure 6.1 shows the use case diagram for this system. There are three use cases: Predict Single Instance, Predict Entire Dataset and Evaluation.

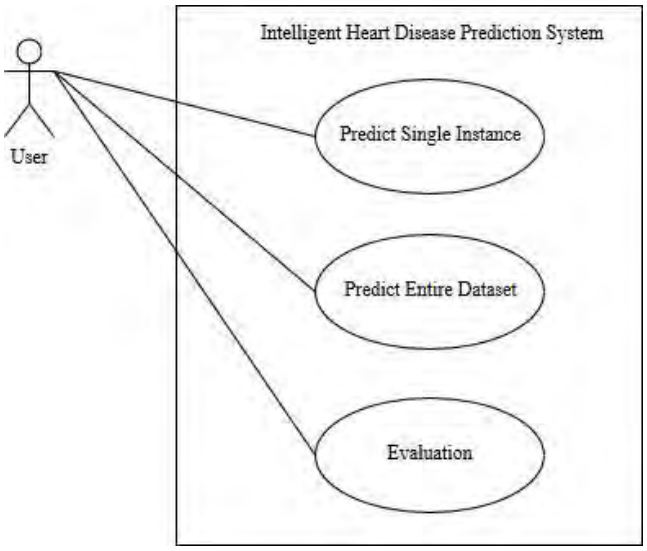

Figure 6.1: Use Case Diagram

6.2.4 Use Case Description

The three use cases of the system are described in this section

Table 6.2: Specification for Use Case "Predict Single Instance"

Use Case ID	U01	**Use Case Name**	Predict Single Instance
Actors	User		
Description	The user uses this to predict heart disease for a single patient by entering details of the patient condition as per the categories are shown in the system.		
Pre-Condition	The model is trained and loaded		
Post-Condition	Log the prediction along with the entered values describing patient condition		
Flow of Events		**Actor Input**	**System Response**
	1	Select “Predict Single Instance”	Present user with an interface for entering attribute regarding patient condition
	2	Enters attributes	Validate the entered attributes
	3	Press “Predict” button	Predicts the heart disease based on the entered attributes and shows the result

Table 6.3: Specification for Use Case "Predict Entire Dataset"

Use Case ID	U02	**Use Case Name**	Predict Entire Dataset
Actors	User		
Description	The user uses this to predict an entire dataset containing patient records.		
Pre-Condition	The model is trained and loaded.		
Post-Condition	Save the predictions in a new dataset and save it on the desktop.		
Flow of Events		**Actor Input**	**System Response**
	1	Select “Predict Entire Dataset”	Present user with an interface for entering the location of the dataset
	2	Press “Predict” button	Predicts the heart disease for each record in the dataset and shows the result

Table 6.4: Specification for Use Case "Evaluation"

Use Case ID	U03	**Use Case Name**	Evaluation
Actors	User		
Description	This is used to evaluate the model with the dataset provided by the user		
Pre-Condition	The model is trained and loaded.		
Post-Condition			
Flow of Events		**Actor Input**	**System Response**
	1	Select “Evaluation”	Present user with an interface for entering the location of the dataset
	2	Press “Evaluate” button	Evaluate the model based on the dataset provided by the user as the test set and display the results using confusion matrix.

6.3 Design and Architecture

This section describes the system architecture, system design and user interface design of this system.

6.3.1 System Architecture

This system was developed using a layered architectural pattern. . Figure 6.2 describes the system architecture. The architecture consists of three layers: UI layer, Logic Layer and Data Layer. The UI layer represents a layer that interacts with the user to get inputs and show outputs. Logic layer works with different control logics that control the system while the data layer consists of weka API and dataset.

6.3.2 Class Diagram

According to Unified Modeling Language (UML), a class diagram describes the structure of a system by giving a visual representation of system's classes, their attributes, operations and the relationships. Figure 6.3 describes the class diagram to show the structure of the system. Four classes named Model, Hospital, Patient and PatientRecords are defined to represent the main objects of this system. Each class includes its attributes and operations.

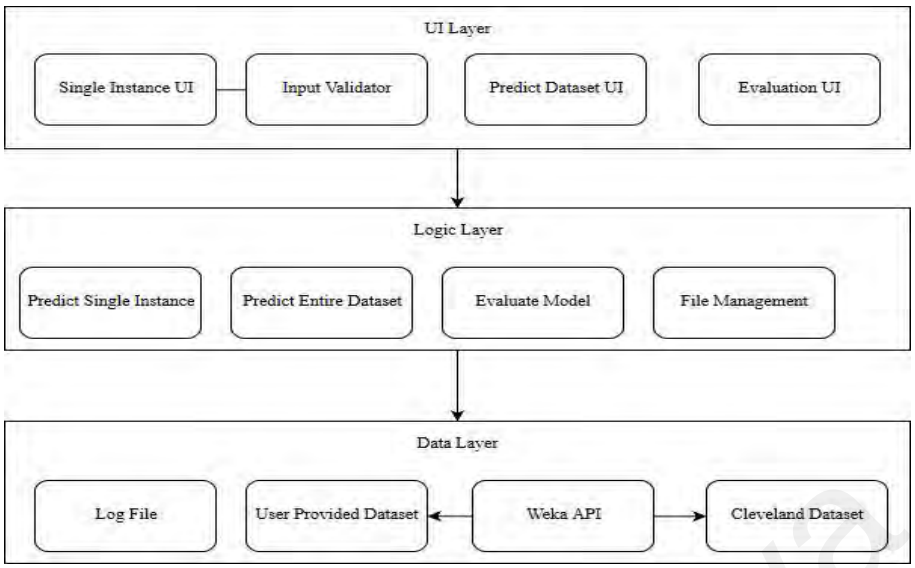

Figure 6.2: System Architecture (Layered Pattern)

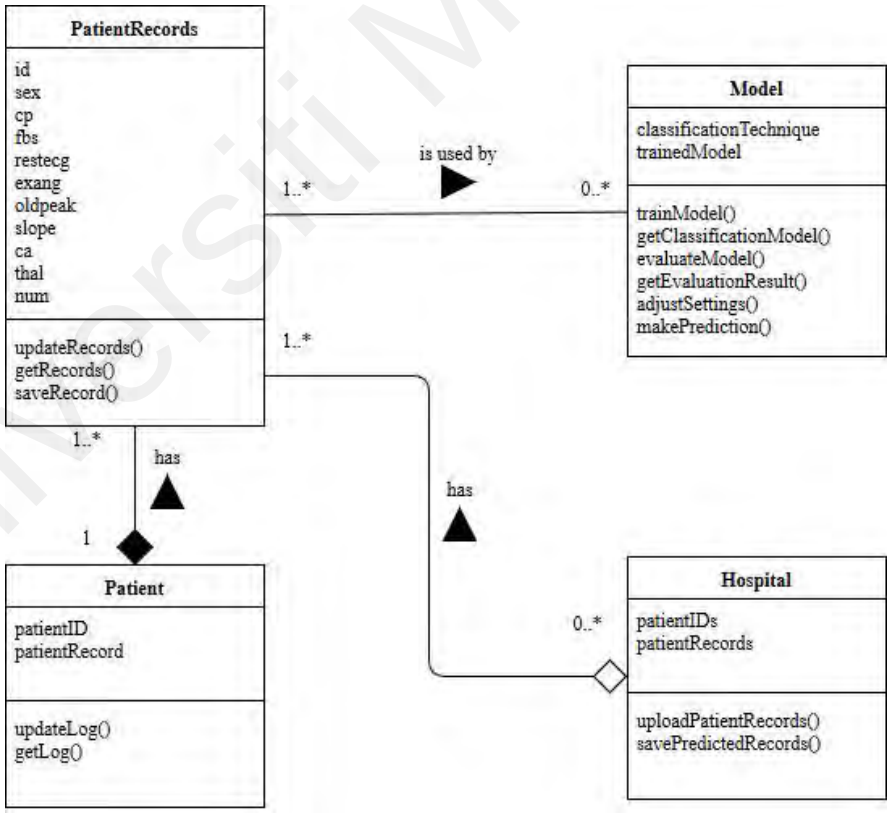

Figure 6.3: Class Diagram of the system

6.3.3 User Interface Design

In this research, an Intelligent heart disease prediction system was developed to improve the diagnosis of heart disease in patient. It employs the proposed model and uses it to predict heart disease with nine significant features and Vote classification technique. The system has three main functions: Predict an entire Dataset, Predict Single Instance, and Evaluation. The system is a desktop-baseded software developed in Java programming language.

As seen from Figure 6.5, there are three options provided in this system. The second tab allows users to predict heart disease for one patient only. The users need to input the patient records for the nine significant features. The users can predict the heart disease status by clicking on the "Predict" button after providing the information. The system displays the results of prediction with the help of a pop-up dialogue box as shown in figure 6.4.

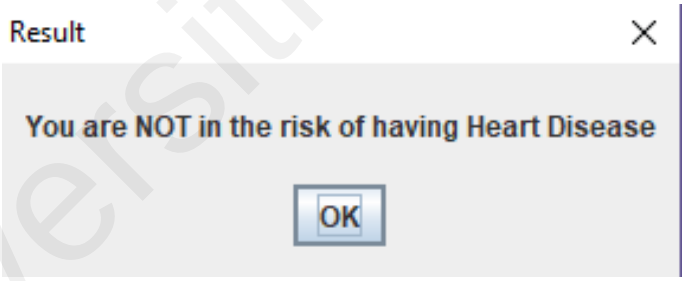

Figure 6.4: Pop-up Dialogue box of the system

Figure 6.5: Make Prediction Interface of the system

Figure 6.6 shows the the first tab of the system, “Predict an entire dataset”. This tab allows users predict heart disease for a dataset that has more than one patient records. The interface allows the users to upload a dataset by clicking the “Choose File” button. Once the users click the ‘Predict’ button, the system will display the predict results for every patient in the dataset based on the information provided. This system automatically saves a copy of the prediction results in the desktop for future uses.

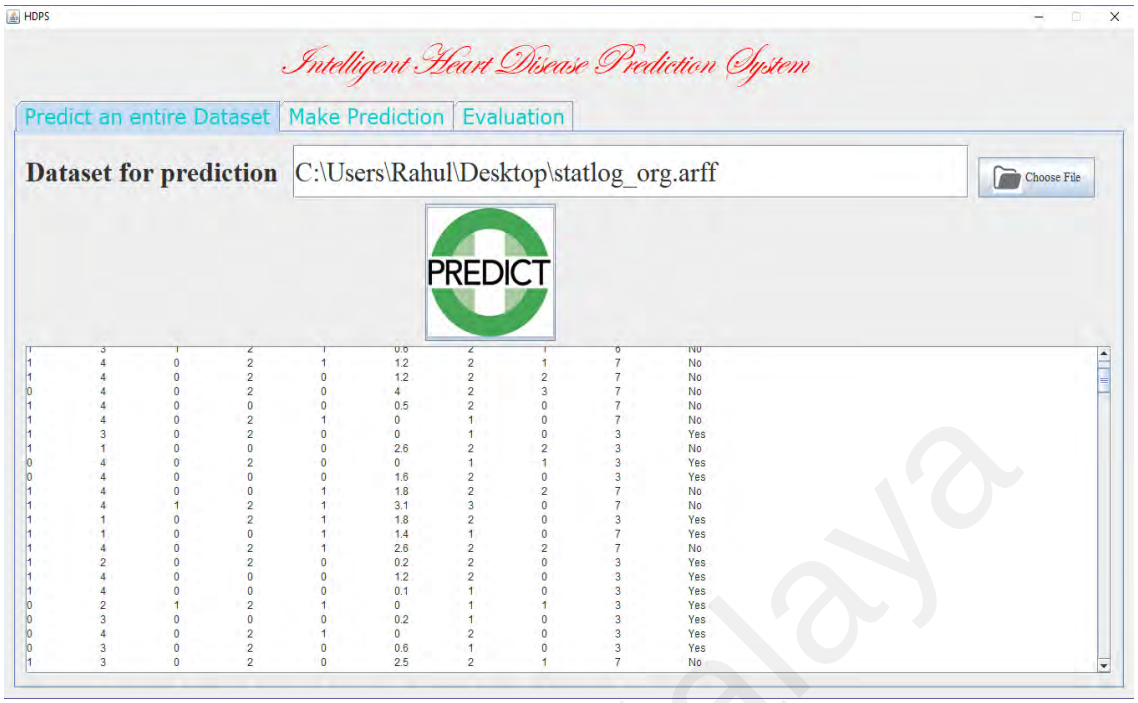

Figure 6.6: Predict an entire Dataset Interface of the system

Figure 6.7 shows the third tab of the system. In this tab, users can evaluate the performance of the model by uploading an arff file that contains a dataset with nine significant features. This function allows one to upload the dataset by clicking the "Choose File" button. After uploading the file, the users click the "Evaluate" button to produce the performance evaluation results of the model for the uploaded dataset. The system displays the results together with a confusion matrix.

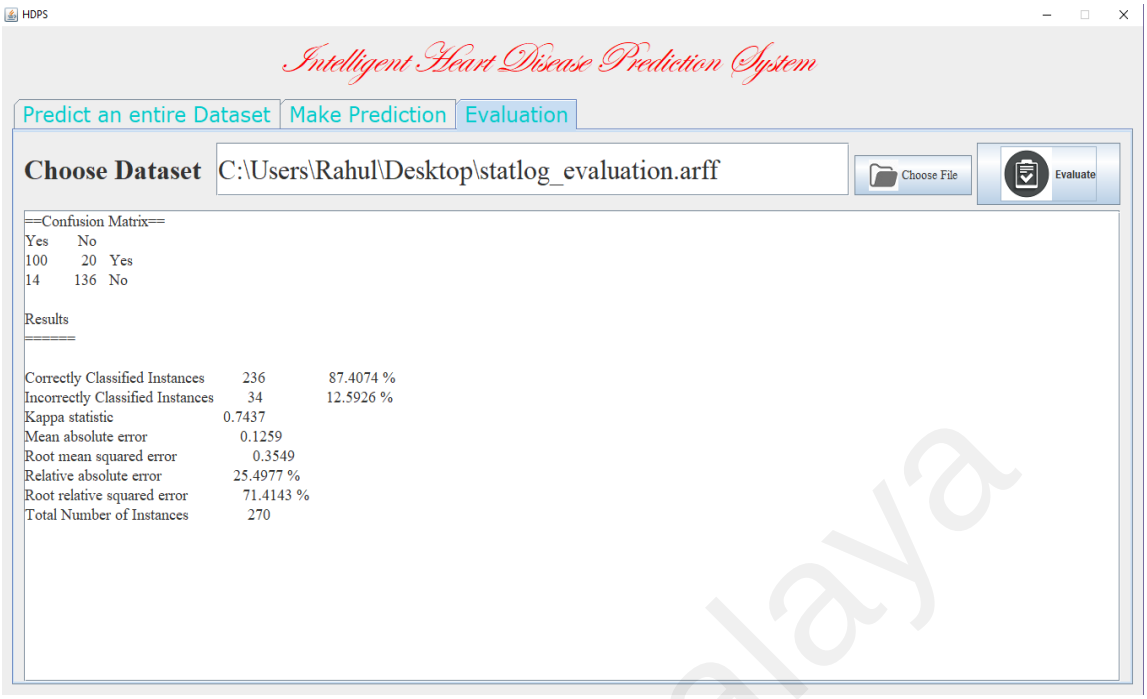

Figure 6.7: Evaluation Interface of the system

6.4 Testing

This section describes the process of testing the intelligent heart disease prediction system. The purpose of the testing is to detect the defects or anomalies and fix all the defects. The testing includes unit testing and system testing to evaluate whether the system has met the requirements and fit for use. Test cases were prepared for system testing.

6.4.1 Unit Testing

In this phase of testing, the system is tested for each of its separate units. The test is designed to verify all the functions are working properly and providing the right system functionalities. White box testing method was used for the purpose of unit testing.

6.4.2 System Testing

In system testing, the software was tested as a complete and integrated system. The purpose of system test is to evaluate whether the system compliance with the

specified requirements and gives the accurate prediction results. The black box testing was used for this purpose.

6.4.3 Test Cases

Test cases were developed to test all the features of the system. To detect as many defects or anomalies as possible during system testing, these test cases were designed and prepared. Test cases were executed to find the defects and fix all the defects. Moreover, the actual outputs of testing are matched with the expected outcomes of test cases. Each test case has its own unique id. Table 6.5 to Table 6.8 shows all the test cases to evaluate this system. All the test cases were executed and passed.

Table 6.5: Test Case for Navigation Between Tabs

Test Case ID	T01
Objective	Navigation Between Tabs
Flow of action	1. Load the System 2. Select different tab than the active one 3. Select tabs that have been left out 4. Select a tab randomly
Expected output	After selection, the subsequent interface will be visible and all another unrelated interface to that selected tab will vanish
Result	Pass

Table 6.6: Test Case for Predicting Single Instance

Test Case ID	T02
Objective	Prediction of heart disease of one patient
Flow of action	1. Select “Make Prediction” tab 2. Fill the information required 3. Click on Predict
Expected output	The prediction will be given in a pop-up dialogue.
Result	Pass

Table 6.7: Test Case for Predicting an Entire Dataset

Test Case ID	T03
Objective	Prediction of heart disease with a dataset
Flow of action	1. Select “Predict an entire dataset” tab 2. Click Choose File to select the dataset 3. Click on Predict
Expected output	The prediction is saved in desktop and a message is displayed in a pop-up dialogue. The prediction of the dataset is shown in the text area of the interface.

Alternate action	Error occurs if the dataset is not chosen Error occurs if the dataset is not in arff format
Result	Pass

Table 6.8: Test Case for Evaluation

Test Case ID	T04
Objective	Evaluation of the model
Flow of action	1. Select "Evaluation" tab 2. Click Choose File to select the dataset 3. Click on Evaluate
Expected output	The evaluation with the confusion matrix is shown in the text area of the interface.
Alternate action	Error if the dataset is not in arff format
Result	Pass

6.5 Summary

This chapter explained the design, development and testing of the system based on the proposed model. The system was developed in JAVA programming language with the help of WEKA API. The system was tested and has met all the defined requirements. The well-integrated system provides the right functionalities and accurate results to predict heart disease.

Chapter 7: Conclusion

7.1 Introduction

The clinical industry has huge patient data that are not processed. Finding a way to process this raw data into a gem of information can save a lot of life. Data mining techniques can be used to analyse the raw data to provide new insights towards the goal of disease prevention with accurate predictions. Heart disease is one of the main causes of death in this world. It is crucial to detect the heart disease in patients as soon as possible to prevent heart disease.

In this research, we have identified significant attributes that improve the accuracy of heart disease prediction using the best performing classification techniques. The research experiment was carried out using the UCI Cleveland dataset and the findings were evaluated using Statlog dataset. The nine significant features identified in this research are sex, cp, fbs, restecg, exang, oldpeak, slope, ca and thal. The top three data mining techniques that produce high accuracy in prediction were identified. These classification techniques are Vote, Naïve Bayes and Support Vector Machine. A two-phase evaluation was conducted to validate the significant features and data mining techniques. The results of the first phase reconfirm the nine selected features are significant. Additionally, among the top three techniques, Vote has outperformed the other two techniques. The best performing model was created using the nine significant attributes and Vote technique. Finally, the proposed model was benchmarked against the models in the existing studies. The outcome of the benchmarking indicates that the proposed classification model has produced higher accuracy in prediction and performed better than other studies.

In this chapter, the outcomes and contribution of this research are summarized. Section 7.2 describes the fulfilment of the objectives of this research while section 7.3

focuses on the contribution of this research. Section 7.4 and 7.5 describes the limitation and future possibilities for this research respectively.

7.2 Objective Fulfillment

Objective 1: To identify significant features in predicting cardiovascular disease

Solution: In this research, we have conducted a thorough experiment using UCI Cleveland dataset and identified 9 significant features. The features are sex, cp, fbs, restecg, exang, oldpeak, slope, ca and thal. During the evaluation, the performance was compared between the models created using the 9 significant features and 13 features. The evaluation results show that the 9 features gave better accuracy than the 13 features.

Objective 2: To identify data mining techniques in predicting cardiovascular disease

Solution: During the experiments, top three data mining techniques were identified. The techniques are Vote, Naïve Bayes and Support Vector Machine. After evaluation, Vote was selected as the best technique. A classification model was proposed using the nine selected significant features and Vote data mining technique.

Objective 3: To evaluate the performance of the prediction models with the identified data mining techniques using the selected significant features

Solution: The proposed model was evaluated using Statlog dataset and also benchmarked with the existing studies. The proposed model has outperformed the existing studies and predicts the heart disease with acceptable accuracy.

Objective 4: To develop a system to predict heart disease with the identified data mining techniques using the selected significant features

Solution: The proposed model was integrated using Java and WEKA API. The system was tested and evaluated. The results were matched with the original experiment.

7.3 Research Contribution

The research contributes in identifying significant attributes and data mining techniques that improve the accuracy of the heart disease prediction. In addition, Vote was identified as the highest performer data mining technique. This has encouraged further research to explore the performance of hybrid techniques. The proposed model has performed well and thus pave the way for further research cardiovascular disease prediction that can support the decision making of clinicians or medical experts. A heart disease prediction system was developed based on the proposed model. This system contributes to predict heart disease easily by providing the nine significant data. Prediction can be done automatically. This research also contributes to future research in many ways. The comprehensive experiments on feature selection make it easier for researchers to understand the heart disease data. Overall this research has shown that the identified significant attributes and data mining techniques have greatly increased the performance of the prediction.

7.4 Limitations

There are some limitations in this research. The limitations are listed as follows:

- The research scope is limited to UCI machine learning repository. The significant features were not tested other real-world datasets.
- The size of the datasets is small. Cleveland dataset only has 303 records and Statlog dataset has 270 records. Further experiments should be conducted on a larger dataset to raise the confidence level of significant features and data mining techniques identified in this research.

- The Vote technique used for the proposed model is a hybrid technique that combines Naïve Bayes and Logistic Regression. Further research can be conducted to test different combination of data mining techniques in heart disease prediction.

7.5 Future Work

There are many ways to enhance this research and address the limitations described earlier. Some of the recommendation for future works are listed as below;

- Extend this research by conducting the same experiment on a large scale real-world dataset.
- Extend this research by experimenting with different combinations of data mining techniques for Vote
- Extend this research by exploring new feature selection methods such as MFS to get a more broader perspective on the significant features to improve accuracy in prediction.

Reference

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. Paper presented at the Acm sigmod record.

Kim, Y., Street, W. N., & Menczer, F. (2003). Feature selection in data mining. Data mining: opportunities and challenges, 3(9), 80-105.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33(1), 135-146.

Srinivas, K., Rao, G. R., & Govardhan, A. (2010a). Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. Paper presented at the Computer Science and Education (ICCSE), 2010 5th International Conference on.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. IADS-DM.

Srinivas, K., Rani, B. K., & Govrdhan, A. (2010b). Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE), 2(02), 250-255.

Karaolis, M. A., Moutiris, J. A., Hadjipanayi, D., & Pattichis, C. S. (2010). Assessment of the risk factors of coronary heart events based on data mining with decision trees. IEEE Transactions on information technology in biomedicine, 14(3), 559-566.

Rajkumar, A., & Reena, G. S. (2010). Diagnosis of heart disease using datamining algorithm. Global journal of computer science and technology, 10(10), 38-43.

Anbarasi, M., Anupriya, E., & Iyengar, N. (2010). Enhanced prediction of heart disease with feature subset selection using genetic algorithm. International Journal of Engineering Science and Technology, 2(10), 5370-5376.

Khatibi, V., & Montazer, G. A. (2010). A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment. Expert Systems with Applications, 37(12), 8536-8542.

Kavitha, K., Ramakrishnan, K., & Singh, M. K. (2010). Modeling and design of evolutionary neural network for heart disease detection. International Journal of Computer Science Issues, 7(5), 272-283.

Liu, G.-P., Li, G.-Z., Wang, Y.-L., & Wang, Y.-Q. (2010). Modelling of inquiry diagnosis for coronary heart disease in traditional Chinese medicine by using multi-label learning. BMC complementary and alternative medicine, 10(1), 37.

Soni, S., & Vyas, O. (2010). Using associative classifiers for predictive analysis in health care data mining. International Journal of Computer Applications, 4(5), 33-37.

Rani, K. U. (2011). Analysis of heart diseases dataset using neural network approach. arXiv preprint arXiv:1110.2626.

Jabbar, M., Chandra, P., & Deekshatulu, B. (2011). Cluster based association rule mining for heart attack prediction. Journal of Theoretical and Applied Information Technology, 32(2), 197-201.

Khaing, H. W. (2011). Data mining based fragmentation and prediction of medical data. Paper presented at the Computer Research and Development (ICCRD), 2011 3rd International Conference on.

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques: Elsevier.

Subbalakshmi, G., Ramesh, K., & Rao, M. C. (2011). Decision support in heart disease prediction system using naive bayes. Indian Journal of Computer Science and Engineering (IJCSE), 2(2), 170-176.

Kumar, D. S., Sathyadevi, G., & Sivanesh, S. (2011). Decision support system for medical diagnosis using data mining. International Journal of Computer Science Issues, 8(3), 147-153.

Parthiban, G., Rajesh, A., & Srivatsa, S. (2011). Diagnosis of heart disease for diabetic patients using naive bayes method. International Journal of Computer Applications, 24(3), 7-11.

Zhao, H., Chen, J., Hou, N., Zhang, P., Wang, Y., Han, J., . . . Wang, W. (2011). Discovery of diagnosis pattern of coronary heart disease with Qi deficiency syndrome by the T-test-based Adaboost algorithm. Evidence-Based Complementary and Alternative Medicine, 2011.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Mendis, S., Puska, P., & Norrving, B. (2011). Global atlas on cardiovascular disease prevention and control. World Health Organization.

Chen, A. H., Huang, S.-Y., Hong, P.-S., Cheng, C.-H., & Lin, E.-J. (2011). HDPS: Heart disease prediction system. Paper presented at the Computing in Cardiology, 2011.

Khemphila, A., & Boonjing, V. (2011). Heart disease classification using neural network and feature selection. Paper presented at the Systems Engineering (ICSEng), 2011 21st International Conference on.

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011a). Intelligent and effective heart disease prediction system using weighted associative classifiers. International Journal on Computer Science and Engineering, 3(6), 2385-2392.

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011b). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.

Acuña, E. (2011). Preprocessing in Data Mining International Encyclopedia of Statistical Science (pp. 1083-1085): Springer.

Shouman, M., Turner, T., & Stocker, R. (2011). Using decision tree for diagnosing heart disease patients. Paper presented at the Proceedings of the Ninth Australasian Data Mining Conference-Volume 121.

Bhatla, N., & Jyoti, K. (2012a). An analysis of heart disease prediction using different data mining techniques. International Journal of Engineering, 1(8), 1-4.

Shouman, M., Turner, T., & Stocker, R. (2012a). Applying k-nearest neighbour in diagnosing heart disease patients. International Journal of Information and Education Technology, 2(3), 220.

Anooj, P. (2012a). Clinical decision support system: risk level prediction of heart disease using decision tree fuzzy rules. Int J Res Rev Comput Sci, 3(3), 1659-1667.

Anooj, P. (2012b). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. Journal of King Saud University-Computer and Information Sciences, 24(1), 27-40.

Atkov, O. Y., Gorokhova, S. G., Sboev, A. G., Generozov, E. V., Muraseyeva, E. V., Moroshkina, S. Y., & Cherniy, N. N. (2012). Coronary heart disease diagnosis by

artificial neural networks including genetic polymorphisms and clinical parameters. Journal of cardiology, 59(2), 190-194.

Dangare, C. S., & Apte, S. S. (2012a). A data mining approach for prediction of heart disease using neural networks. International Journal of Computer Engineering and Technology (IJCET), 3(3), 30-40.

Lahsasna, A., Ainon, R. N., Zainuddin, R., & Bulgiba, A. (2012). Design of a fuzzy-based decision support system for coronary heart disease diagnosis. Journal of medical systems, 36(5), 3293-3306.

Devi, T., & Saravanan, N. (2012). Development of a data clustering algorithm for predicting heart. International Journal of Computer Applications, 48(7).

Rajeswari, K., Vaithiyanathan, V., & Neelakantan, T. (2012). Feature selection in ischemic heart disease identification using feed forward neural networks. Procedia Engineering, 41, 1818-1823.

Muthukaruppan, S., & Er, M. J. (2012). A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. Expert Systems with Applications, 39(14), 11657-11665.

Dangare, C. S., & Apte, S. S. (2012b). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-48.

Shouman, M., Turner, T., & Stocker, R. (2012b). Integrating decision tree and K-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients. Paper presented at the Proceedings of the International Conference on Data Mining (DMIN).

Ephzibah, E., & Sundarapandian, V. (2012). A neuro fuzzy expert system for heart disease diagnosis. Computer Science & Engineering, 2(1), 17.

Bhatla, N., & Jyoti, K. (2012b). A Novel Approach for heart disease diagnosis using Data Mining and Fuzzy logic. International Journal of Computer Applications, 54(17).

Sundar, N. A., Latha, P. P., & Chandra, M. R. (2012). Performance analysis of classification data mining techniques over heart disease database. IJESAT] International Journal of engineering science & advanced technology ISSN, 2250-3676.

Huang, F., Wang, S., & Chan, C.-C. (2012). Predicting disease by using data mining based on healthcare information system. Paper presented at the IEEE International Conference on Granular Computing (GrC), 2012 Khan, M. T., Qamar, S., & Massin, L. F. (2012). A prototype of cancer/heart disease prediction model using data mining. International Journal of Applied Engineering Research, 7(11), 1-6.

Peter, T. J., & Somasundaram, K. (2012). Study and development of novel feature selection framework for heart disease prediction. International Journal of Scientific and Research Publications, 2(10), 1-7.

Shouman, M., Turner, T., & Stocker, R. (2012c). Using data mining techniques in heart disease diagnosis and treatment. Paper presented at the Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on.

Nahar, J., Imam, T., Tickle, K. S., & Chen, Y.-P. P. (2013a). Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications, 40(4), 1086-1093.

Nahar, J., Imam, T., Tickle, K. S., & Chen, Y.-P. P. (2013b). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications, 40(1), 96-104.

Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., . . . Sani, Z. A. (2013). A data mining approach for diagnosis of coronary artery disease. Computer methods and programs in biomedicine, 111(1), 52-61.

Sen, A. K., Patel, S. B., & Shukla, D. (2013). A data mining technique for prediction of coronary heart disease using neuro-fuzzy integrated approach two level. International Journal of Engineering and Computer Science, 2(9), 1663-1671.

Chaurasia, V., & Pal, S. (2013). Early prediction of heart diseases using data mining techniques. Carib. j. SciTech, 1, 208-217.

Amin, S. U., Agarwal, K., & Beg, R. (2013). Genetic neural network based data mining in prediction of heart disease using risk factors. Paper presented at the Information & Communication Technologies (ICT), 2013 IEEE Conference on.

Taneja, A. (2013). Heart disease prediction system using data mining techniques. Oriental Journal of Computer science and technology, 6(4), 457-466.

Medhekar, D. S., Bote, M. P., & Deshmukh, S. D. (2013). Heart disease prediction system using naive bayes. Int. J. Enhanced Res. Sci. Technol. Eng, 2(3).

Chitra, R., & Seenivasagam, V. (2013a). Heart disease prediction system using supervised learning classifier. Bonfring International Journal of Software Engineering and Soft Computing, 3(1), 1.

Shouman, M., Turner, T., & Stocker, R. (2013). Integrating clustering with different data mining techniques in the diagnosis of heart disease. J. Comput. Sci. Eng, 20(1).

Ishtake, S. & Sanap, S. (2013). Intelligent heart disease prediction system using data mining techniques. International Journal of Healthcare & Biomedical Research, 1(3), 94-101.

Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013a). Heart disease prediction system using associative classification and genetic algorithm. arXiv preprint arXiv:1303.5919.

Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013b). Knowledge discovery using associative classification for heart disease prediction Intelligent Informatics (pp. 29-39): Springer.

Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013c). Classification of heart disease using k-nearest neighbor and genetic algorithm. Procedia Technology, 10, 85-94.

Manikantan, V., & Latha, S. (2013). Predicting the analysis of heart disease symptoms using medicinal data mining methods. International Journal of Advanced Computer Theory and Engineering, 2, 46-51.

Chitra, R., & Seenivasagam, V. (2013b). Review of heart disease prediction system using data mining and hybrid intelligent techniques. ICTACT journal on soft computing, 3(04), 605-609.

Blake, C. L., & Merz, C. J. (1998). UCI Repository of machine learning databases [http://www. ics. uci. edu/~ mlearn/MLRepository. html]. Irvine, CA: University of California. Department of Information and Computer Science, 55.

Hachesu, P. R., Ahmadi, M., Alizadeh, S., & Sadoughi, F. (2013). Use of data mining techniques to determine and predict length of stay of cardiac patients. Healthcare informatics research, 19(2), 121-129.

Chaurasia, V., & Pal, S. (2014). Data mining approach to detect heart diseases.

Chandna, D. (2014). Diagnosis of heart disease using data mining algorithm. (IJCSIT) International Journal of Computer Science and Information Technologies, 5(2), 1678-1680.

Banu, M. N., & Gomathy, B. (2014). Disease forecasting system using data mining methods. Paper presented at the Intelligent Computing Applications (ICICA), 2014 International Conference on.

Methaila, A., Kansal, P., Arya, H., & Kumar, P. (2014). Early heart disease prediction using data mining techniques. Computer Science & Information Technology Journal, 53-59.

Tomar, D., & Agarwal, S. (2014). Feature selection based least square twin support vector machine for diagnosis of heart disease. International Journal of Bio-Science and Bio-Technology, 6(2), 69-82.

Masethe, H. D., & Masethe, M. A. (2014). Prediction of heart disease using classification algorithms. Paper presented at the Proceedings of the world congress on engineering and computer science.

Bajaj, P., & Gupta, P. (2014). Review on Heart Disease Diagnosis Based on Data Mining Techniques. International Journal of Science and Research (IJSR), 3(5).

Kaur, B., & Singh, W. (2014). Review on heart disease prediction system using data mining techniques. International journal on recent and innovation trends in computing and communication, 2(10), 3003-3008.

Sudhakar, K., & Manimekalai, D. M. (2014). Study of heart disease prediction using data mining. International Journal of Advanced Research in Computer Science and Software Engineering, 4(1).

Asmi, S. P., & Samuel, S. J. (2015). An analysis and accuracy prediction of heart disease with association rule and other data mining techniques. Journal of Theoretical and Applied Information Technology, 79(2), 254.

Kim, J., Lee, J., & Lee, Y. (2015). Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree. Healthcare informatics research, 21(3), 167-174.

Bandyopadhyay, S., Wolfson, J., Vock, D. M., Vazquez-Benitez, G., Adomavicius, G., Elidrisi, M., . . . O'Connor, P. J. (2015). Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data. Data Mining and Knowledge Discovery, 29(4), 1033-1069.

Naghavi, M., Wang, H., Lozano, R., Davis, A., Liang, X., Zhou, M., . . . Abd-Allah, F. (2015). Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet, 385(9963), 117-171.

Nahato, K. B., Harichandran, K. N., & Arputharaj, K. (2015). Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Computational and mathematical methods in medicine, 2015.

Bahrami, B., & Shirvani, M. H. (2015). Prediction and Diagnosis of Heart Disease by Data Mining Techniques. Journal of Multidisciplinary Engineering Science and Technology (JMEST), 2(2), 164-168.

Kalaiselvi, C., & Nasira, G. (2015). Prediction of heart diseases and cancer in diabetic patients using data mining techniques. Indian Journal of Science and Technology, 8(14).

Purusothaman, G., & Krishnakumari, P. (2015). A survey of data mining techniques on risk prediction: Heart disease. Indian Journal of Science and Technology, 8(12).

Dey, A., Singh, J., & Singh, N. (2016). Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis. Analysis, 140(2).

Thanigaivel, R., & Kumar, K. R. (2016). Boosted Apriori: an Effective Data Mining Association Rules for Heart Disease Prediction System. Middle-East Journal of Scientific Research, 24(1), 192-200. doi: 10.5829/idosi.mejsr.2016.24.01.22944

Kavitha, R., & Kannan, E. (2016). An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. Paper presented at the Emerging Trends in Engineering, Technology and Science (ICETETS), International Conference on.

Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System. Procedia Computer Science, 85, 962-969.

Paul, A. K., Shill, P. C., Rabin, M. R. I., & Akhand, M. (2016). Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. Paper presented at the Informatics, Electronics and Vision (ICIEV), 2016 5th International Conference on.

Kaur, K., & Singh, L. M. (2016). HEART DISEASE PREDICTION SYSTEM USING ANOVA, PCA AND SVM CLASSIFICATION.

Mokashi, A. R., Tambe, M. N., & Walke, P. T. (2016). Heart Disease Prediction Using ANN and Improved K-Means. Heart Disease, 4(4).

Turabieh, H. (2016). A Hybrid ANN-GWO Algorithm for Prediction of Heart Disease. American Journal of Operations Research, 6(02), 136.

Verma, L., Srivastava, S., & Negi, P. (2016). A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. Journal of medical systems, 40(7), 1-7.

Abdar, M., Zomorodi-Moghadam, M., Das, R., & Ting, I.-H. (2017). Performance analysis of classification algorithms on early detection of liver disease. Expert Systems with Applications, 67, 239-251.

Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications–A decade review from 2000 to 2011. Expert systems with applications, 39(12), 11303-11311.

Oracle. (n.d.). Oracle Advanced Analytics' Machine Learning Algorithms SQL Functions. Retrieved February 27, 2017, from http://www.oracle.com/technetwork/database/enterprise-edition/odm-techniques-algorithms-097163.html

World Health Organization. (2017). Cardiovascular diseases (CVDs). Retrieved March 10, 2017, http://www.who.int/mediacentre/factsheets/fs317/en/.

Rajkumar, A., & Reena, G. S. (2010). Diagnosis of heart disease using datamining algorithm. Global journal of computer science and technology, 10(10), 38-43.

Weka 3: Data Mining Software in Java. (n.d.). Retrieved February 05, 2017, from https://www.cs.waikato.ac.nz/ml/weka/

Data Analytics. (n.d.). Retrieved February 05, 2017, from https://www.mathworks.com/solutions/data-analytics.html

TANAGRA - A free DATA MINING software for teaching and research. (n.d.). Retrieved February 05, 2017, from https://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html

IBM SPSS Modeler. (n.d.). Retrieved February 05, 2017, from https://www.ibm.com/us-en/marketplace/spss-modeler

Data Mining Extensions (DMX) Reference. (n.d.). Retrieved February 05, 2017, from https://docs.microsoft.com/en-us/sql/dmx/data-mining-extensions-dmx-reference

RapidMiner: Data Science Platform. (n.d.). Retrieved February 05, 2017, from https://rapidminer.com/

KEEL: Software tool. Evolutionary algorithms for Data Mining. (n.d.). Retrieved February 05, 2017, from http://www.keel.es/