ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF HEALTH INFORMATICS

MINING ECHOCARDIOGRAPHY DATA TO PREDICT HEART DISEASE (MEDPHD)

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN HEALTH INFORMATICS

BY
THOMAS HABTE ABRHA

JUNE, 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF HEALTH INFORMATICS

MINING ECHOCARDIOGRAPHY DATA TO PREDICT HEART DISEASE (MEDPHD)

By
THOMAS HABTE ABRHA

Name and Signature of Members of the Examining Board

Dr Million Meshesha, Advisor _____

Dr Jemal Hader, Advisor _____

Dr Solomone Tefera, Internal Examiner _____

Dr. Fasil, External Examiner _____

DECLARATION

The thesis is my original, has not been presented for a degree in any other university andthat all sources of material used for the thesis have been duly acknowledged.

Thomas Habte Abrha

June 2012

The thesis has been submitted for examination with our approval as university advisors.

Dr Million Meshesha

Dr Jemal Hader

ACKNOWLEDGEMENTS

Above all, I would like to glorify the almighty GOD for giving me the ability to be where I am. Oh, GOD through very difficult time, you have done so mach for me

I am strongly indebted to my advisor Dr Million Meshesa from the faculty of Informatics, Addis Abeba University, for his unreserved advice, keen insight, guidance, meticulous comment, and skilful pushes I received throughout my thesis work. Without his advice, the accomplishment of this research would have been impossible. I also would like to express my heartfelt gratitude to Dr. Jemal Hadar for his all rounded guidance and support in the course of writing this Thesis. My sincere applause goes to Dr Dejuma Yadota from Internal Medicine cardiology unit, Addis Abeba University.

I am deeply grateful to international Cardiovascular Hospital and Abel Damtaw, for allowing me to access echocardiography datasets and assisting me in any way when I look for and especially Sister Tesday Gethanu and Ato Getachew H/Selease, who are staff members of the Hospitals, for their unlimited support and understanding.

.

I would like to express my appreciation to the wonderful friends I have had at the Department of Health Informatics most of whose company I can count on and enjoyed. Namely Abraham G/Giorigis, Messay Ketmbo, Biniyam Ayele, Beuzyahu Getnet, Akale Regessa, Tesfahun H/Mariam and Antenh Akelilu.

Finally yet importantly, I would like to thank the coordinator of Health Informatics, W/ro Mesert Ayanaw , very deeply. Ato Haile Getachew(Tom) as well, without whose good will, I would not have accomplished this research.

TABLE OF CONTENTS

LIST OF ACRONYMS AND ABBREVIATIONS

ANN	Artificial Neural Network
BNN	Bayesian Neural Network
BPNN	Back Propagation Neural Network
CART	Classification and Regression Trees
CHD	Coronary Heart Diseases
CRISP-DM	Cross Industry Standard Process for Data Mining
CSV	Comma-Separated Value
CVD	Cardiovascular Diseases
ECG	Electrocardiography
FN	False Negative
FP	False Positive
IHD	Ischemic Heart Disease
IHDPS	Intelligent Heart Disease Prediction System
IQR	Inter Quartile Range
KBS	Knowledge Based System
KDD	Knowledge Discovery in Database
KDP	Knowledge Discovery Process
MCG	Magnetocardiogram
MMH	Maximum Marginal Hyperplane
PNN	Probabilistic Neural Network
SMO	Sequential Minimal Optimization
SPSS	Statistical Package for Social Science

SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TTE	Transthoracic Echocardiography
WEKA	Waikato Environment for Knowledge Learning
WHO	World Health Organization

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Background: These days, a major challenge of health care is reaching to correct diagnosis of specific disease condition. Poor clinical decision leads to catastrophic consequences which are unacceptable. Decision making process at the health care setting needs to be supported with more advanced technology including a computer based information system.

Objective: This study aims at extracting hidden knowledge (patterns and relationships) associated with echocardiography datasets and designing a predictive model for heart disease detection using data mining techniques.

Methods: A Hybrid Data Mining methodology is followed, which is a six-step knowledge discovery process. The data for this research obtained from International Cardiovascular Hospital in Addis Abeba, Ethiopia.

This study investigates the use of different data mining techniques, Decision tree, neural network, and support vector machine for classification tasks. On Transthoracic Echocardiography report datasets, descriptive data summarization and visualization were taken to gain understanding of the data. Moreover, missing values, outliers data, data integration and transformation were managed at preprocess stage of hybrid process model.

Results: The results show that all the models performed well, though J48 Decision tree algorithms outperforms support vector machine, Multilayer Perceptron Neural Network, registering 96.73%. The best attributes selected by J48 decision tree are Left Atrium Systole Diameter, LV ejection fraction, and Tricuspid velocity. As per discussion made with the cardiologist, one of the interesting rule, a patient with Left atrium systole diameter less than or equal to 40 millimeter and LV ejection fraction less than or equal to 51% blood pumped out of ventricles and Tricuspid velocity is greater than 2.5 centimeter per second results Left Ventricle dysfunction and Pulmonary hypertensive disorder.

Conclusion: The result thus obtained in this study is promising to apply data mining for heart disease detection. To make usable the knowledge extracted in this study, an attempt has made to design a knowledge-based system that shows the potential to integration. It is a further research direction.

Keywords: Echocardiography, Knowledge discovery process, Decision tree, neural network, Support vector machine

CHAPTER ONE
INTRODUCTION

1.1. Background

Healthcare industry today generates huge amounts of data about patients, hospitals resources, disease diagnosis, electronic patient records, and medical devices. This large amount of data is a key resource to be analyzed and processed to extract hidden information and knowledge. A major challenge of health care facilities is the provision of correct patients diagnosis and administering treatments effectively. Poor clinical decisions lead to catastrophic consequences, which are unacceptable. A decision making process at the health care setting needs to be supported with more advanced technology including a computer based information system.

Clinical decisions rest with health care professionals. This decision-making action can be integrated with decisions support system which reduce medical errors significantly. Typically, data mining brings a set of tools and techniques that can be applied to discover hidden patterns and knowledge. This knowledge provides health care professionals an additional source of information to make intelligent clinical decisions. This enhances patient safety, increase patient quality of life, and improve patient diagnosis outcome.

Around the world chronic disease are mainly heart disease, stroke, cancer, chronic respiratory disease, and diabetes [16]. According to World Health Organization (WHO), those chronic diseases took about 35 million people worldwide including many young and middle age groups. From factsheet of WHO the total number of people dying from chronic diseases is double that of all infectious disease including HIV/AIDS, tuberculosis and malaria [16]. Another interesting fact is that of 80% of chronic disease deaths occur in low and middle-income countries. And without addressing the causes, deaths from chronic disease will increase by 17% between 2005 and 2015 [16].

Basically, there are three major types of heart disease. These are cardiovascular disease (CVD), Coronary heart disease (CHD), and ischemic heart diseases.

Data mining is the analysis observational data set to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [8]. Data mining are typically applicable to heart disease diagnosis. For instance, a study [1] conducted in 2008 have disclosed that using medical variables such as age, sex, blood pressure, blood sugar and other variables it is possible to predict the likelihood of patients getting a heart disease. Similarly, another recent study conducted in 2010 [2] have indicated data mining application to Magnetocardiogram (MCG) tool. It is a tool for detecting electro-physiological activity of the heart. MCG is a fully non-contact method, which avoids the problems of skin-electrode contact in the Electrocardiogram (ECG). By applying different classification models, they have proposed the use of machine learning for identification of Ischemic heart disease (IHD) patient. The researchers also state that the analysis of MCG data by experts is time-consuming along with the shortage of experts who possess the knowledge on the analysis of MCG data. Therefore, methods are proposed to automate the interpretation of MCG recordings by minimizing human efforts [2].

Many researches are published on the heart disease using data mining technology classifications and predication methods. But most of these researches are conducted based on medical and socio-demographic variables [1], [3], [4], [5].

1.2. Statement of the Problem

Heart disease is one of the popular chronic diseases that took many lives worldwide. Cardiovascular disease (CVD) has typically been limited to economic developed countries. Nevertheless, due to rapid changes in consumption of excessive amounts of animal-derived fat, urbanization, increasing life expectancy, and because of control of communicable disease, a sedentary life style and smoking, the problem has spread to Latin America, Eastern Europe, Asia, and to the populations of the poor African countries [22].

At 21st century, CVD is responsible for 30 per cent of all death worldwide, with about 80 per cent of the burden of CVD death occurring in low-income and middle-income countries [23]. Specifically, in sub-Saharan Africa death due to CVD are project to more than double between the years 1990 and 2020 [23].

In order to diagnosis heart disease, professionals take patient history, physical findings, order laboratory investigation and other Scio-demographic characters of the patient like eating habit, smoking , exercise etc. After assessing the result of those tests, the physician makes decisions. While doing these processes, echocardiography test mostly ordered to the patients to determine the size, shape and movement of heart's valves and its chambers. Each parameters of echocardiography is crucial to permit a full assessment of the heart and accurate diagnosis of certain cardiovascular diseases.

Interpreting the output of echocardiography takes about 30 minutes up to 1 hour based on the severity of the cardiac structures. Secondly, health professionals may interpret wrongly because echocardiography measurements take many tedious variables and error prone. Moreover, in our country there is scarcity of echo cardiographer who specializes on this machine.

Taking into consideration of the above problems, this research primarily intend to develop a model by taking Transthoracic Echocardiography reports; determine the status of a patient (i.e. healthy or sick person). This will save a great deal of time; it adds value to quality decisions to physicians which in turn favor the patient to heal and can be applied where there is shortage of professionals.

To this end, this study attempt to answer the following research questions:

- Which attributes are more important to heart disease prediction?
- Which classification algorithm is more suitable to create a predictive model for heart disease detection?

1.3. Objectives of the Study

1.3.1. General Objective

The general objective of this research is to extract hidden knowledge (patterns and relationships) associated with echocardiography datasets and design a predictive model for heart disease detection using data mining techniques.

1.3.2. Specific Objectives

In order to achieve the stated general objective, the research work has carried the following specific objectives

- To conduct an exhaustive review of literature on the existing data mining techniques and methods in general, and their application in the healthcare sector particularly in heart disease treatment.
- To generate good quality datasets by applying data preprocessing techniques
- To select suitable classification algorithms and build prediction model
- To evaluate the performance and the result of the selected models

1.4. Scope and Limitations of the Study

The term “heart disease” refers to several types of heart conditions. The most common type is cardiovascular disease, Coronary artery disease, and ischemic heart diseases. Echocardiography machine determine the presence of many types of heart disease [6] and accordingly all the cases are considered in this study. Using echocardiography datasets, this research attempts to apply data mining techniques, specifically classification algorithms that enable to construct predictive model and determine whether the person is healthy or sick of heart disease.

The major limitation of this study is the inability to design Knowledge Based System that integrates the hidden knowledge extracted from the given datasets. This is due to time constraints and lack of enough support from cardiology experts.

1.5. Research Methodology

A research methodology is an arrangement of condition for collocation and analysis of data in a manner that aim to address the research problem.

1.5.1. Research Design

To realize a model that yields optimum classifier of heart disease status of an individual, a Hybrid data mining process model were applied that consumes Knowledge Discovery Process (KDD) and Cross-industry Standard Process (CRISP-DM) models.

Hybrid data mining model is selected for the present study since:

- It provides more general, research-oriented description of the steps
- It does emphasize the iterative aspects of the process, drawing experience from previous models.
- It support academia and industrial data mining projects

The Hybrid DM (see figure 1.1) consists of six-step Knowledge Discovery Process (KDP); i.e. understanding the problem domain, understanding data, preparation of data, data mining, evaluation of the discovered knowledge, use of the discovered knowledge.

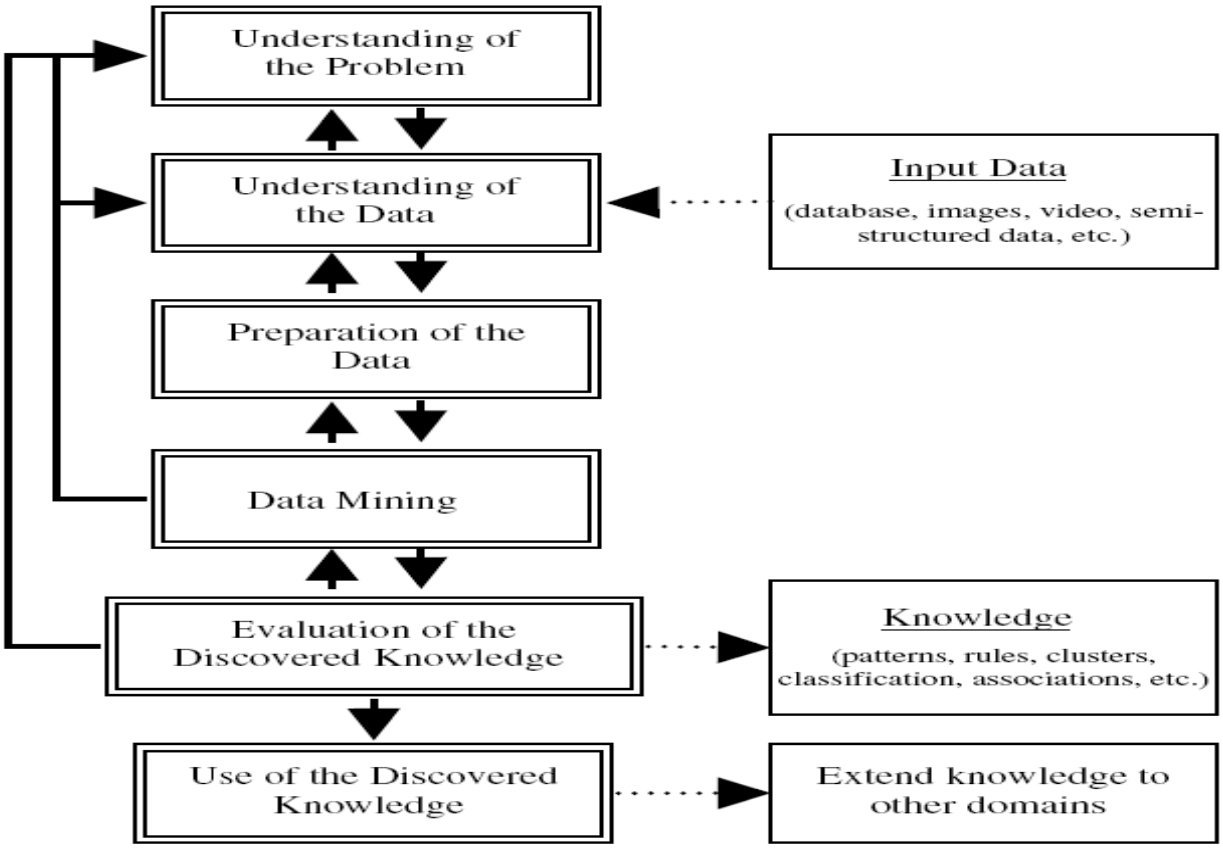

Figure 1.1 Hybrid Process Model

1.5.2. Understanding of the Problem

This stage of hybrid model involves working closely with domain experts to define the problem and corresponding solutions and determine the research goals. It also involves learning about domain-specific terminology, selection of DM tools to be used in the process. Therefore, literatures were consulted to learn about domain specific terminology and DM tools selection as well.

The study area was in Addis Ababa city administration, which is the capital city of the Ethiopia. With Average Elevation of 2500m above sea level, the city administration has a geographic and territorial possession with an area of 540 sq. km and a total population of about 3 million [18].

The distribution of public health facilities in the city administration is; Hospitals; owned by Addis Ababa Health Bureau 5, Ministry of Health 4, Addis Ababa University 1, Ministry of defense 2, and Police force 1 total government Hospitals 13. Total governmental Health Centers owned by Addis Ababa Health Bureau 23. Total governmental Clinics owned by Addis Ababa Health Bureau 9. Total Health Posts owned by Addis Ababa Health Bureau 34. Total health facilities at Addis Ababa city is 79 [19]. Unfortunately, for convenience and unavailability of data in government hospitals, Problem understanding of Echocardiography datasets were taken in private hospital called International Cardiovascular hospital situated in Addis Abeba, Ethiopia.

1.5.3. Understanding of the Data

High quality data is a prerequisite for any data mining technique. The sources of data for this research are result of echocardiography machine that acquired from private hospital of International Cardiovascular Hospital situated in Addis Abeba, Ethiopia.

To gain understanding of echocardiography data, discussion were taken with domain experts and other personnel who interact with the data. Moreover, descriptive summarization and visualizations of data conducted using statistical software of SPSS. SPSS is an acronym for Statistical Package for Social Science. This application software were used to create a database for echocardiography dataset.

Thus, for this research, total amounts of 7339 dataset utilized. The datasets for this study have a scale of measurement of numeric (11 attributes), Nominal (3 attributes including target class), and ordinal (1 attribute). This datasets partitioned and used for training the model and testing the model accuracy.

1.5.4. Preparation of the Data

The datasets were undergoes data preparation steps to confirm for completeness, redundancy, missing values and plausibility of attribute. The collected data preprocessed and cleaned in a way to fulfill the requirement of data mining software. After the selection of relevant features/factors of echocardiography machine datasets based on goal of the study, the next step was data preparation. In this step of data preparation, tasks like handling missing values, handling outliers' data, transformation of data, and data reduction were taken place. Feature selection and

extraction algorithms process handled to acquire cleaned data. The results are data that meet the specific input requirements for DM tools.

1.5.5. Data mining Techniques

Here the data miner uses various classification Data Mining (DM) methods to derive knowledge from the preprocessed data. Among the available algorithms in WEKA machine learning software; Decision Tree, Artificial Neural Network (ANN), and Support Vector Machine were used in this research. These models were selected in this research due to their popularity in the recently published documents. The subsequent chapters are a brief introduction to the three classification algorithms and parameter setting of each model.

Two data mining tools have used namely, SPSS and WEKA to implement the KDP.
WEKA, formally called Waikato Environment for Knowledge Learning developed at the University of Waikato in New Zealand, is open-source data mining software in java. It provides implementations of learning algorithms that can be applied to a given dataset and analyze its output to learn more about the data, and use learned models to generate predictions on new instances [17]. Another possible way to apply WEKA is that of to use the learned models to generate predictions on new instances and compare the performance of the models in order to select the best for prediction [17].

1.5.6. Evaluation of the Discovered Knowledge

Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts. Only approved models are retained. Thus after the development of the model based on the training dataset, the accuracy of the model were tested using test datasets. A confusion matrix and the result of the model were assessed to determine the impact of the discovered knowledge.

Using confusion matrix, accuracy, sensitivity, specificity, and precision were calculated to evaluate the performance of each models.

1.5.7. Use of the Discovered Knowledge

This final step consists of planning where and how to use the discovered knowledge. This last step determines the success of the entire knowledge discovery process. A plan to monitor the implementation of the discovered knowledge is created and the entire project is documented. The results of this thesis work will be disseminated to the following stakeholders and to any interested parties. For the reason this, interested readers can get access to the research results so as to support the decisions making process, or use it for further research in the area or for any other applicable reasons.

- It will be presented to school of information science and to school of public health.
- A hardcopy of this thesis results will be available to bibliographic library of information science.
- A softcopy of this thesis will upload to Addis Abeba University e-resource official website.
- Maximum effort will be exerted to publish the result in different journals.

1.5.8. Ethical Consideration

The research entitled “Mining Echocardiography Data to Predict Heart Disease” acquired ethical clearance and approval from the department of Community Health, Faculty of Medicine, Addis Abeba University. The necessary verbal permission was obtained from International Cardiovascular Hospital Director. In addition, a letter of cooperation wrote for communicating the concerned officials about the thesis from Faculty of Informatics. Furthermore, during analysis, a study subject’s sensitive data kept anonymous and confidential.

1.6. Significances of the Research

This study attempt to extract useful predictive model from echocardiography dataset, and hence physician can make use of these model in their day-to-day diagnosis of heart diseases. Particularly echocardiography measurements are less successful and inconsistent relative to other imaging techniques, sometimes perceived as unreliable [33]. Therefore, decisions support system in diagnosis of heart disease would be fundamental.

Physicians, regarding to heart disease, can make use of the results of this study in order to make optimal decisions about the presence or absence of heart disease. It can be deployed where sonographer personal is not there; so that General practitioners (GP) can utilize the model application to determine the patients have a heart disease or not. Moreover, it adds quality to decision-making process for quality health care service provisions to the society at large.

In addition, health institutions and interested researchers use this result as a stepping stone towards the application of data mining on Echocardiography to create a model that can help fighting heart diseases endeavor.

1.7. Organization of the Research

This thesis contains six chapters. The first chapter deals with the general overview of the study including background, statement of the problem, research objectives and methodology of the research. The second chapter is devoted to literature review of data mining technology, heart disease, echocardiography machine, and review of related literature.

Chapter Three is about the data mining algorithms that are used to build the model based on echocardiography machine report. Chapter Four is devoted to Business Understanding, Data understanding and Data preprocessing of the data for generating good quality datasets for the classification task.

Chapter Five reports the experiment of the research. It comprises training, building and validation of the models. Results of the experiment were also analyzed and interpreted. The last, presents chapter six, concluding remarks and recommendations of the study.

CHAPTER TWO

LITERATURE REVIEW

Due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge, data mining has attracted a great deal of attention in the information industry and in society as a whole. This huge volume of data can be accumulated beyond database and data warehouses. The abundance of data, coupled with the need for powerful data analysis tools, has been described as a “data rich but information poor situation”.

Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research. Data mining refers to extracting or “mining” knowledge from large amounts of data [8].

In fields of business, science, and engineering a need to understand large, complex, information-rich data sets is common to all. A method to extract useful knowledge hidden in these data and to act on the knowledge is becoming increasingly important in today’s competitive world. The entire process of applying a computer-based methodology for discovering knowledge from data is called data mining [18].

2.1. Multidisciplinary Nature of Data Mining

Han et al [7] mentioned, data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science. Moreover, depending on the data mining approach used, techniques from other disciplines may be applied, such as neural networks, fuzzy and/or rough set theory, knowledge representation, inductive logic programming, or high-performance computing.

Han et al [7] explained that because of the diversity of disciplines contributing to data mining, data mining research is expected to generate a large variety of data mining systems. Therefore, it is necessary to provide a clear classification of data mining

systems, which may help potential users distinguish between such systems and identify those that best match their needs.

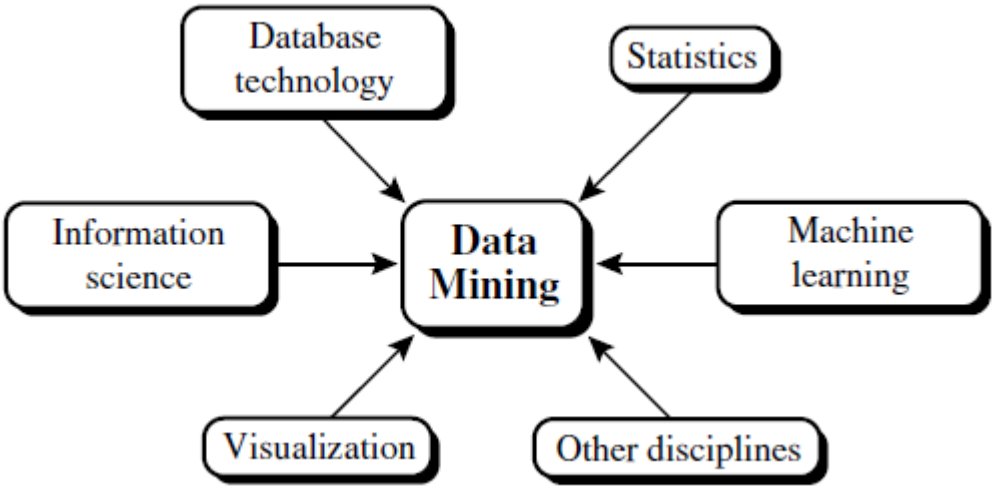

Figure 2.1 Data mining a confluence of multiple disciplines

As Figure 2.1 picture portray, data mining is an interdisciplinary field, the intersection of a set of disciplines, i.e. database technology, statistics, information science, information retrieval and others. Even, depending on the application, data mining system may also integrate techniques from spatial data analysis, pattern recognition and image analysis, signal processing, computer graphics, web technology, economics, business, bioinformatics, or psychology [7].

2.2. Data Mining Modeling Methodology

The two primary goals of data mining are prediction and description. Predictive modeling consists of several types of model such as classification, regression and predictive models and is referred to as supervised learning, since calculated or estimated values are compared with known results. Descriptive techniques, on the other hand deal with association mining, clustering analysis, pattern recognition models, and visualization methods. Descriptive techniques are referred to as unsupervised learning which interrogate the database to identify patterns and relationship in the data.

These days, there are a number of tools for developing predictive and descriptive models. Some use statistical methods such as linear regression and logistic regression. Others use non-statistical or blended methods like neural networks, genetic algorithms, classification trees, and regression trees. Olivid [9] articulates, the steps surrounding the model processing are more critical to the overall success of the project than technique used to build the model.

2.2.1. Classification Tasks

2.2.1.1. Decision tree/Classification Trees

During the late 1970s and early 1980s, J. Ross Quinlan, a researcher in machine learning, developed a decision tree algorithm known as ID3 (Iterative Dichotomiser) [7]. Quinlan later presented C4.5 (a successor of ID3), which became a benchmark to which newer supervised learning algorithms are often compared [7]. In 1984, Breiman et al. [7] published the book *Classification and Regression Trees* (CART). ID3 and CART were invented independently of one another at around the same time, yet follow a similar approach for learning decision trees from training tuples [7].

The most widely used algorithms of decision tree are ID3, C4.5, and in statistical field Classification and Regression Trees(CART) adopt a top down recursive divide-and-conquer manner [28]. Most algorithms for decision tree induction also follow such a top-down approach [7]. Induction is a type of reasoning which is generalization based on repeated observations. After many observations of X and Y occurring together, learn the rule if x then y [27].

According Larose [24], there must be certain requirements that met before decision tree algorithms applied; this are

1. Decision tree algorithms represent supervised learning. A training data set must be supplied with pre-classified target variables.
2. The training data set should be rich and varied. Decision trees learns by example and if examples are lacking certain subset of records, classification for this subset will create problematic.

3. Finally, the target attributes classes must be discrete. Decision tree cannot accept a continuous target variable.

The true purpose of a classification tree is to *classify* the data into distinct groups or branches that create the strongest separation in the values of the dependent variable [9].Classification trees are developed through a series of steps and rules that offer great flexibility. The top node represents the performance of the overall campaign.

Each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can easily be converted to classification rules [7].

Han [7] notes that if a medical researcher wants to analyze breast cancer data in order to predict specific treatments a patient receive; data analysis task is classification, where a model or classifier is constructed to predict *categorical labels*, such as “treatment A,” “treatment B,” or “treatment C” for the medical data. And then these categories can be represented by discrete values, where the ordering among values has no meaning. For example, the values 1, 2, and 3 may be used to represent treatments A, B, and C, where there is no ordering implied among this group of treatment regimes. This classifier model typically constructed with decision tree model.

Data classification is a two-step process. In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or “learning from” a training set made up of database tuples and their associated class labels. Because the class label of each training tuple *is provided*, this step is also known as supervised learning. The classifier is “supervised” in that it is told to which class each training tuple belongs. On the other hand, the class labels of each training tuple is not known, and the number or set of classes to be learned may not be known in advance then the selected task is unsupervised learning [7].

In the second step, the predictive accuracy of the classifier is estimated. If we were to use the training set to measure the accuracy of the classifier, this estimate would likely be optimistic, because the classifier tends to overfit the data. Therefore, a test set is used, made up of test tuples and their associated class labels. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier [7].

2.2.1.2. Neural Networks

A neural network, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units [7]. It is modeled after the function of the human brain [9].

According to Olivid [9], neural networks are made up of nodes that are arranged in layers. This construction varies depending on the type and complexity of the neural network. Before the process begins, the data is split into training and testing data sets and a third group is held out for final validation. Then weights or ''inputs'' are assigned to each of the nodes in the first layer. During each iteration, the inputs are processed through the system and compared to the actual value. The error is measured and fed back through the system to adjust the weights. In most cases, the weights get better at predicting the actual values. The process ends when a predetermined minimum error level is reached.

Typically, the strength of neural network is its ability to pick up nonlinear relationships in the data. This can allow users to fit some types of data that would be difficult to fit using regression. But the drawback is its tendency to over-fit the data. This can cause the model to deteriorate more quickly when applied to new data. If this is the method of choice, one must be sure to validate carefully. Another disadvantage to consider is that the results of a neural network are often difficult to interpret [9].

2.3. Data Mining Functionalities

2.3.1. Characterization & Discrimination

Data can be associated with classes or concepts. It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. According to Han et al [7], these concepts and class descriptions can be derived using data characterization and data discrimination. Data characterization is a summarization of the general characteristics or features of a target class of data and data discrimination, by comparison of the target class with one or a set of comparative classes.

2.3.2. Frequent Patterns, Associations, and Correlations

Frequent patterns, as the name implies, are patterns that occur frequently in data. It discusses many kinds of frequent patterns, subsequences and substructures [7]. A frequent item set typically refers to a set of items that frequently appear together in a transactional data set. And a frequent occurring subsequence, such as the pattern that customers tend to purchase first a personal computer(PC), followed by a digital camera, and then memory card. Such phenomena described as frequent subsequence.

Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

2.3.3. Classification and Prediction

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data whose data objects whose class label is known. David Hand [8] argues data mining model is data-driven and the discovery of a highly predictive model should not be taken to mean that there is a causal relationship. For example, an analysis of customer records may show that customers who buy high-quality wines are also more likely to buy designer clothes; in this case clearly one's tendency is not causally related to the other propensity. The derived model may be represented in various forms, such as *classification (IF-THEN) rules*, *decision trees*, *mathematical formulae*, or *neural networks [7].*

In a predictive model, one of the variables is expressed as a function of the others. This permits the value of the *response* variable to be predicted from given values of the *explanatory* or *predictor* variables [8].

2.3.4. Application of Data Mining

According to De Veaux[11], data mining applied at Customer relation management (CRM) to help improve their entire business process including new customer acquisition, customer retention, and also up selling opportunities and customer optimization across different areas. The writer also mentions the points regarding to marketing experiments.

Another area of application as mentioned by Han et al [7], data mining is also applicable for biological and DNA data analysis. Biomedical researches include the development of new pharmaceuticals and advances in cancer therapies to the identification and study of human gene by discovering large-scale sequencing patterns. The authors added that recent research of DNA analysis has led to the discovery of new medicine and approaches for disease diagnosis, prevention, and treatment.

2.4. Heart and Echocardiography

2.4.1. Anatomy and Physiology of Heart

Heart is hollow muscular organ that pumps blood through the body. The heart, blood, and blood vessels make up the circulatory system, which is responsible for distributing oxygen and nutrients to the body and carrying away carbon dioxide and other waste products [12].

Four valves within the heart prevent blood from flowing backward in the heart. The valves open easily in the direction of blood flow, but when blood pushes against the valves in the opposite direction, the valves close. Two valves, tricuspid valve and bicuspid or mitral valve, known as atrioventricular valves located between the atria and ventricles. The other two heart valves are located between the ventricles and arteries. Commonly known as pulmonary valve and the aortic valve [12].

2.4.2. Cardiovascular Disease

Cardiovascular disease (CVD) is any of a number of specific diseases that affect the heart itself and/or the blood vessel system, especially the veins and arteries leading to and from the heart (Wikipedia).

A. Coronary Heart Disease

As stated in Eugene [12], the most common type of heart disease in most industrialized countries. In United States alone, account over 515,000 deaths per year. It is caused by atherosclerosis, the buildup of fatty material called plaque on the inside of the coronary arteries. Through time, the arteries narrow so that less oxygen reaches the heart muscle.

Coronary artery disease is a common cause of systolic dysfunction. As a result, the ventricles pump out less than 40 to 50% of the blood. At a normal circumstance, the ventricles are pumped out about 60% of the blood [13].

B. Heart Valve Disorders

Malfunction of one of the four valves within the heart can cause problems that affect the entire circulatory system. According to J. Malcolm [13], narrowing (stenosis) of a valve, which hinders blood flow through the heart, or leakage of blood backward (regurgitation) through a valve will cause heart failure. Both stenosis and regurgitation of a valve can severely stress the heart, so that over time, the heart enlarges and cannot pump adequately. A well-known type of valve malfunction is mitral valve prolapse. ”In this condition, the leaflets of the mitral valve fail to close properly and bulge backward like a parachute into the left atrium” [12].

C. Myocarditis/Inflammation of Heart Muscle

Myocarditis caused by a bacterial, viral, or other infection that can damage all or part of the heart muscle, impairing its pumping ability [13]. Eugene [12] pronounces that myocarditis leads to permanent damage of the heart muscle, reducing the heart’s ability to pump blood and making it prone to developing abnormal rhythms.

D. Heart Failure

The final stage in almost any type of heart disease is heart failure in which the heart muscle weakens and is unable to pump enough blood to the body [12]. Heart failure develops in about 1 of 100 people [13].

According to Malcolm [13], heart failure does not mean that the heart has stopped rather it means that the heart cannot keep up with the work required to pump adequate blood to all parts of the body. However, this definition is simplistic, Heart failure is extremely complex, and no simple definition can encompass its many causes, aspects, forms, and consequences.

2.4.3. What is Echocardiography?

As Joseph [14] wrote, Echocardiography is a unique noninvasive method for imaging the living heart. It is based on detection of echoes produced by a beam of ultrasound pulses transmitted into the heart.

According to Aruro et al. [6] echocardiography is the most widely used imaging technique in clinical cardiology practice since it provides comprehensive evaluation of cardiac and vascular structures and function. Two (2D) and three-dimensional (3D) imaging can accurately assess cardiac chamber size, wall thickness, ventricular function, valvular anatomy and the size of great vessels. A complete adult echocardiographic report have a measurements of left ventricle, Mitral valve, left atrium, Aortic valve, Aorta, Right ventricle, Tricuspid valve, Right atrium, Pulmonary valve, Pulmonary artery, Pericardium, Inferior vena cava, and Pulmonary veins [6]. This echocardiography machine emits several different frequency waves called ultrasound.

2.4.4. Ultrasound

Ultrasound is sound made up of several different frequency waves. The very high frequency range is inaudible to the human ear and is known as ultrasound [15].

The transducer, a microphone-shaped device, transmits ultrasound and constantly receives waves that are reflected back every time the beam travels from one density to another. The reflected ultrasound waves are collected and analyzed by the machine. Determining the amount of time it took the beam to travel from and to the

transducer (plus the consistency and changes in position of the different structures that it passed through), the ultrasound machine can determine the shape, size, density and movement of all objects that lay in the path of the ultrasound beam. The information is presented on a monitor screen and can also printed on paper or, recorded on tape, a Compact Disk (CD). This is how an obstetrician evaluates the fetus of a pregnant woman, and a cardiologist examines the heart of a patient [15].

2.5. Review of Related Literature

Numerous works in Heart disease related researches using data mining and DM algorithms have been conducted.

A survey research to predict heart morbidity conducted by *Srinivas et al [21],*disclose that using attributes with dichotomous self reported measures whether respondents were ever diagnosed with morbidity and other 15 medical attribute used for prediction of heart attack. Age is one of the attributes which was coded in years and ranged from 18 to 99; other covariates included smoking, coded as a three level variable; current, former and/or life time smoker. Srinivas et al. [21], have used a decision tree, Naïve bayes, and neural network algorithms. As a result, a classification accuracy of each algorithm is 82%, 82%, and 92% have achieved respectively. The typical reason for good neural network accuracy is that of syndrome is identified by human brain and neural network is considered as best modeler of human brain [21].

Another researcher called Yosawin et al [2] have conducted a research on MCG measurement and model for identification of ischemic heart disease patients. This is different from those researchers who model the pattern of heart disease using medical profile and socio-demographic variables. Experiments were carried out on MCG data acquired by sequential measurement of myocardium using J-T interval of 125 cases. The machine learning algorithms of Back-propagation neural network (BPNN), the Bayesian neural network (BNN), the probabilistic neural network (PNN), and the support vector machine (SVM) were applied to develop classification models for identifying IHD patients.

Related research called Intelligent Heart Disease Prediction System (IHDPS) built with the aid of data mining techniques like Decision Trees, Naïve Bayes and Sellappan P. et al. [1] proposed Neural Network. Naïve Bayes model appear to perform better as it gives the highest accuracy/ prediction of 86.12%, followed by neural network (85.68%) and Decision Trees (80.4%). The results illustrated the peculiar strength of each of the methodologies in comprehending the objectives of the specified mining objectives. IHDPS was capable of answering queries that the conventional decision support systems were not able to. It facilitated the establishment of vital knowledge, e.g. patterns, relationships amid medical factors connected with heart disease. IHDPS subsists well being web-based, user-friendly, scalable, reliable and expandable.

Finally, another very similar research conducted by Abel [37] conducted a research with the aim of designing a predictive model for heart disease detection using data mining techniques. The finding of this study suggests that the J48 model perform better in predicting heart disease with classification accuracy of 95.56% than Naïve Bayes and Neural network classifier. As per the recommendation made by Abel [37] in this study, an extensive experimentation is done to create better predictive model, based which an attempt is made to design a knowledge-based system. This will show the potential of integrating the knowledge extracted using data mining to knowledge based system to ease its usability.

CHAPTER THREE

ALGORITHMS USED FOR KNOWLEDGE DISCOVERY

One fundamental classification techniques used in data mining is decision tree. They are tree-like structures used for classification, clustering, feature selection, and prediction [28]. Furthermore, decision tree are easily interpretable and well suited for high-dimensional applications. Another classification method used is that of Artificial Neural network (ANN) which is useful for learning complex data. The last algorithm used in this study is support vector machine (SVM). This is a new method for the classification of linear and nonlinear data. SVM map the original training data into a higher dimension; within this dimension, it searches for a decision boundary separating the tuples of one class from another [7].

3.1. Decision Trees

Decisions tree algorithm have three parameters. This are: D (a complete set of training tuples and their associated class labels), attribute list (a list of attributes describing the tuples), and attribute selection method which select best discriminates attribute given tuples according to class. This method employs an attribute selection measure, such as information gain or gini index [7]. The typical difference between information gain and gini index is that of gini index enforce the resulting tree to be binary where as information gain allow a multiway splits or it allow two or more branches to be grown from a node.

3.1.1. J48 Algorithms

The basic algorithms for decision tree induction is a greedy algorithm which constructs decision trees in a top down approach dividing each node recursively until a leaf node is encountered. The following algorithm shows the generating of a decision tree from a training tuples of data partition [7].

Input

- *Data partition, D, which is a set of training tuples and their associated class labels;*
- *Attribute list, the set of candidate attributes;*

- *Attribute selection method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting attribute and, possibly, either a split point or splitting subset.*

***Output**: A decision tree*

***Method**;*

Create a node N;

If tuples in D are all the same class, C then

Return N as a leaf labelled with the class C;

If attribute list is empty then

Return N as a leaf node labelled with the majority class in D;

Apply attribute selection method (D, attribute list) to find the "best" splitting criterion;

Label node N with splitting criterion;

If splitting attribute is discrete valued and multiway splits allowed then

Create list ← attribute list splitting attribute;// remove splitting attribute

For each outcome j of splitting criterion

Let Dj be the set of data tuples in D satisfying outcome j;

If Dj is empty then

Attach a leaf labelled with the majority class in D to node N;

Else attach the node returned by Generate decision tree (Dj, attribute list) to node N;

End for

Return N

To construct optimal decision tree, Entropy and Information Gain needs to be calculated. The information gain measure enables to select the test attribute at each node in the tree and the attribute with the highest information gain or greatest entropy reduction is chosen as the test attribute for the current node [29].

Let S be a set consisting of s data samples. Suppose the class label attribute has m distinct values defining m distinct classes, C_i (C_1, C_2, C_3 . . . , C_m). Let Si be the number of sample of S in class C_i. The expected information needed to classify a given sample is given by:

$$I(s_1, s_2, \ldots\ldots\ldots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad \text{..... (3.1)}$$

Where Pi is the probability that an arbitrary sample belongs to class Ci and a log function is to the base 2 is used because the information is encoded in bits.

The entropy, or expected information based on the partitioning into subsets by A is given by

$$D(n_+, n_-) = -\frac{n_+}{n}\log_2\frac{n_+}{n} - \frac{n_-}{n} - \frac{n_-}{n}\log_2\frac{n_-}{n} = \text{Entropy(A)} \quad \text{...(3.2)}$$

The smaller the entropy value is, the greater the purity of the subset partitions [29]. The information that would be gained by branching on A is given by the following formula;

$$\text{Gain (A)} = (s_1, s_2, \ldots\ldots\ldots, s_m) - \text{Entropy(A)} \quad \text{..... (3.3)}$$

This algorithm computes the information gain of each attribute. The attribute with the highest information gain is chosen as the test attribute for the given set S. A node is created and labeled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly [29].

Decision tree use the above formulas to determine which attribute to split on. The highest information gain considered as the test attribute for the given database. Then, a node is created and branches out. The same procedure followed for the next attribute to split.

3.2. Artificial Neural Network

An artificial neural network (ANN) tries to capture the brain problem solving ability and apply them to information systems. The human brain provides proof of the

existence of massive neural networks that can succeed at those cognitive, perceptual, and control tasks in which humans are successful [20]. Humans are better than computers in performing complex tasks such as speech and image recognition, but computers are faster in performing mathematical computations. ANN attempts to replicate the human massive processing capacity to problem solving information systems.

The basic architecture for an ANN is depicted in Figure 3.1. Each node receives inputs, does processing, and generates output. Whether this output will be transferred to other nodes will depend on its strength. As illustrated in Figure 3.1, a Multilayer Feed-forward Neural Network have three types of layers in the network; input, hidden, and output layer. The input layer corresponds to the number of attributes, which would be weighted and fed simultaneously for hidden layer, and the weighted outputs of the last hidden layer are input to output layer [7].

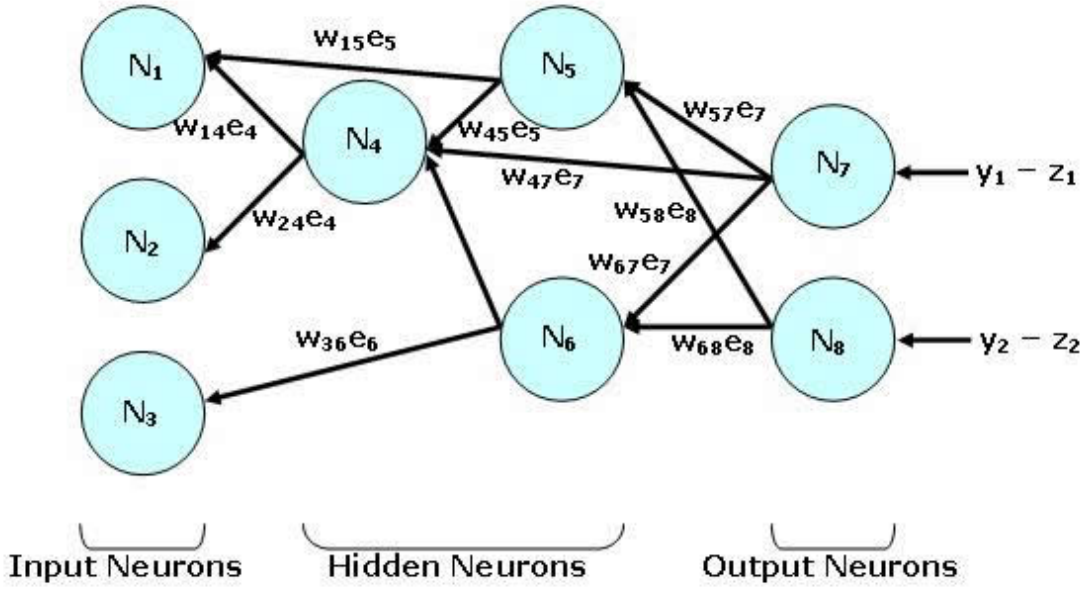

Figure 3.1 Multilayered artificial neural network

One drawback of Neural network is that all attribute values must be encoded in a standardized manner, taking values between zero and one, even for categorical variables [24]. To standardize the attribute values between zeros and ones of continous variable and categorical variables there are a number techniques. Among those techniques we may apply the min-max normalization, Z-score normalization,

and normalization by decimal scaling [25]. The min-max normalization works for continous variable as long as the minimum and maximum values are known[24].

$$X^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad \text{.....(3.4)}$$

But for categorical variable the above formula are problematic. For example, in a given dataset gender attribute, containing values female, male would have to transformed since neural network could not handle in their present forms.

Those categorical values could have a value of 1 for female and 0 for male to record female, while records for males would have a values of 0 for female and 1 for male. Another example set by T. Larose [24] that the data set contains information on a marital status attribute. It is possible to code the attribute values divorced, married, separated, single, widowed, and unknown, as 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0, respectively.

In general, the data given for ANN needs normalization to enforce the attributes to fall in specific range of values such that 0.0 to 1. Normalization is particularly useful for classification algorithms involving neural networks or distance measurements [25].

3.2.1. Backpropagation Algorithms

Among many neural networks algorithms; In 1980s an algorithms called backpropagation gained reputation [7]. The backpropagation algorithm performs learning on a Multilayer Feed Forward-Neural Network; it iteratively learns a set of weights for prediction of the class label of tuples. The network is feed-forward in that none of the weights cycles back to an input unit and it's fully connected in that each unit provides input to each unit in the next forward layer.

Backpropagation learns iteratively by processing a data set of training tuples. At each observation from the training set is processed through the network, an output

value is produced from the output node. This output value is compared to the actual value of the target variable for this training set observation, and the error is calculated [24]. As a result weights are modified so as to minimize the mean squared error [7].

The steps for computing the output of a single neuron are as follows:

1. Compute the weighted sum of inputs to the neuron and add the bias to the sum

$$(I_j) = \sum_i W_{ij} O_i + \theta_j \quad \text{..... (3.5)}$$

Where wij is the weight of the connection from unit j in the previous layer to unit j; Oj is the output of unit i from the previous layer, $\theta j$is the bias of unit.

2. Each unit in the hidden and output layers takes its net input and then applies an activation function [7]. The output of the activation function is defined to be the output of the neuron.

$$O_j = \frac{1}{1+e^{-Ij}} \quad \text{..... (3.6)}$$

This function is called logistic or sigmoid function or also is referred to as a squashing function, because it maps a large input domain onto the smaller range of 0 to 1 [7].

3.3. Support Vector Machine

Until 1992 Support Vector Machine (SVM) were largely unnoticed due to widespread belief in the statistical and/or machine learning community, despite being theoretically appealing [30]. They were taken seriously only when excellent results achieved in numeral recognition, computer vision, and text categorization; today

SVM show better results than Neural network comparable outcome and other statistical models [30].

SVM is a promising new method for the classification of both linear and nonlinear data; it uses a nonlinear mapping to transform the original training data into a higher dimension [7].

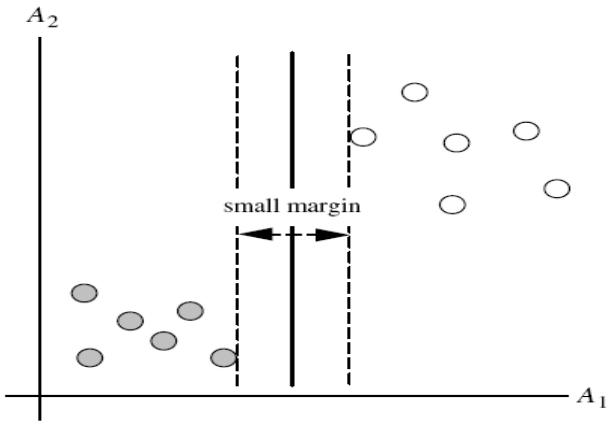

Figure 3.2 The 2-D training data are linearly separable.

To explain SVM; if it is a two-class problem where the classes are linearly separable, an algorithm is implemented to finds a special kind of linear models. Let $x_i \in R_n$, (i= 1, 2, 3… m) represents the vectors and $y_i \in \{1,-1\}$. The term $f(x_i)$ can be represented by a linear functions of the form by $y_i = f(x_i)$

$$f(x_i) = (W.\ X) + b \quad \text{..... (3.7)}$$

Where W is a weight vector namely, $W = \{W_1, W_2, W_3 \ldots W_n\}$ and b is a scalar, often referred to as bias [7]. There is infinite number of hyperplane/ separating lines that could be drawn for classifying the two-classes [7]. To find the optimal linear model or Hyperplane(n dimensions) that will have the minimum classification error on previously unseen tuples, SVM search for maximum marginal hyperplane [17].

Maximum marginal hyperplane is the one that gives the greatest separation between the classes [17]. Figure 3.2 depict that hyperplanes that can correctly classify all of the given data tuples. But larger margin are likely to be more accurate at classifying future data tuples than the hyperplane with the smaller margin.

During learning phases, SVM searches for the hyperplane with largest margin, which is the maximum marginal hyperplane (MMH) [7]. When dealing with the MMH, this distance is the shortest distance from the MMH to the closest training tuple of either class [7]. A hyperplane separating the two classes' decision boundary may be written as

$$X = b + \sum \alpha_i y_i \, \mathrm{a(i).a} \quad \text{..... (3.8)}$$

Here, y_i is the class value of training instance a(i); while b and α_i are numeric parameters that have to be determined by the learning algorithms. a(i) and a represent the vectors. The vector a represent the test instances and a(i) are the training instances.

3.4. Evaluation of Credibility

Mostly for classification problem, it is common to measure a classifiers performance in terms of the *error rate or misclassification rate*[17]. The classifier predicts class label of each instances and if it is correct, counted as success, else counted as error. Evaluating the accuracy using training datasets derive a classifier or predictor to be likely misleading due to overspecialization of the learning algorithms to the data [7], [17]. For the reason this we need to assess the error rate based on independent test dataset that have no role in classifier datasets. Both, training data and the test data, needs to be representative sample of the problem [17].

For measuring accuracy there are a number of techniques such as the holdout (train amount is held for training and testing), random sub-sampling, k-fold cross-validation, and bootstrap, which select datasets with replacement for training and testing purpose, method [7]. In K-fold cross validation techniques, one decide the number of fold (partitions of the data) and then the data is split in K approximately

equal partitions; each partitions, two-thirds and one-third, in turn used for training and testing respectively [17]. 10-Fold cross validation was implemented for this research.

For this research to evaluate the creditability of algorithms performance, confusion matrix, which tells the correctly classified positive and negative records and incorrectly classified records as well are used.

3.4.1. Confusion Matrix

Confusion matrix is a tool for analyzing how well the classifier can recognize tuples of different classes [7]. Give m classes, a confusion matrix is a table of at least m by m. For instance, in tow class case with classes yes and no, sick or healthy, cancer or not cancer, or lend or not lend, a single prediction has the four different possible outcomes.

Table 1.1 Different outcomes of a two-class prediction

		Predicted class	
		Yes	No
Actual class	Yes	true positive	False negative
	No	false positive	True negative

The above table display the four possible outcomes of a two-class prediction that is true positive (TP), true negative (TN) which are the correct classification and the false positive (FP) occurs when the outcome is incorrectly predicted as yes when it is actually no. False negative (FN) are the positive tuples that were incorrectly predicted as no when it is actually yes [7], [17].

Confusion matrix enable to access how well the classifier can recognize “yes” tuples and how well it can recognize “no” tuples, the sensitivity (the true positive rate) and specificity (the true negative rate) measure can be used respectively [7].

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.9)$$

$$\text{Specificity} = \frac{TN}{FP+TN} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.10)$$

The overall success of rate is the number of correct classification divided by the total number of classification .i.e.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.10)$$

In this study, to measure the performance of each model, sensitivity, specificity, and Accuracy were implemented to each learning algorithms. Then, the best selected learning algorithm was discussed in brief.

CHAPTER FOUR

DATA PREPARATION

Data Mining is a technology that uses various techniques to discover hidden knowledge from a data stored in large databases, data warehouses and other massive information repositories. To discover non-trivial knowledge and patterns, the database needs to undergo effective business understanding, data understanding and data preparations steps of a model/process.

For this study, Hybrid DM model/process selected. Hybrid data mining process embrace six steps; understanding of the problems, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and use of the discovered knowledge. Data understanding and preprocessing usually consumes the majority of the effort in the entire data mining process [7].

4.1. Business Understanding

Sharma and Osei-Bryson[32] argue that business understanding phase is somewhat more important than other phases like that of modeling, data preparation, data understanding, evaluation, etc. Lack of thorough discussions during Business understanding phase lead to the DM project may take a completely different direction than what was intended. Further, it creates inefficiencies with regard to time and resources as these decisions have to be dealt with in later phases [32].

Business understanding mainly concerned with determination of business objectives, assessment of the situations, and determination of data mining goals and this section modulated as per those mentioned points.

This study was consulted with domain experts and non-medical staffs to have insight into the problem domain and physical observation was conducted in the hospital as well. The domain experts constitute individuals from International cardiovascular hospital that are in charge of Nursing and treating patients with heart disease.

4.1.1. Determination of Business Objectives

Business objectives determination requires discussion among either responsible business personnel who interact with the system in horizontally or vertically manner to finalize the business objective.

International Cardiovascular Hospital situated in Addis Abeba and established in February 2001 as an outpatient clinic. It was the first specialized private hospital in cardiovascular disease. Currently, the hospital employed more than 35 permanent medical staff and many part-time staff. The medical personnel include cardiologist (3), nurse (21), Radiologist (1), Pharmacist (5), and Laboratory technologist (2).

The primary objective of International Cardiovascular Hospital is to provide quality cardiac diagnosis in a nationwide and in the East Africa as well. In addition, it intended to provide every piece of Cardiac service locally where many patients look for abroad. For example, peacemaker of heart was not available sometimes ago but now it has been started.

Regarding to echocardiography department, the primary objective, the physician is to be able to identify correctly heart disease patients and non heart disease ones at all times. Unfortunately, in some cases though many years of experience, echocardiography report may not signify heart status of the patients.

4.1.2. Echocardiography Report

Patients at international Cardiovascular Hospital may be new, referred patients, and old patients who had medical examination at the hospital. New patients directly request service at the reception desk and the receptionist fills all the required information on the medical chart and issue an index card. Next, the receptionist send medical chart to the nurse stations. The nurse took vital signs including pulse rate, Oxygen saturations, Body Mass Index (BMI), and blood pressure. After vital sign taken, the patient goes to the physician along with its medical chart. Moreover, all the patients need to take ECG (Electrocardiography) before going in to examination room with the physician.

When the physician is ready to examine the patients, one of the nurses accompanies the patients to the examination room for diagnosis. The physician record the patient history, physical examination and the physician may order laboratory test mostly including echocardiography.

Echocardiography is the dominant cardiac imaging technique used in emergency, operating, and intensive care departments mainly because of its portability and versatility [33]. Furthermore, echocardiography can provide predicting information and assist in the management of patients with acute, chronic, and end stage Heart failure [34]. Echocardiography is a type of ultrasound examination, which bounces sound waves off an object and records the returning sound waves. The reflections are processed by special instrument and powerful computer that subsequently measure and create a visual image of the organ [35]. This image will be shown on the screen of the machine and the machine also have different shaped transducers (probes) attached to it. This ultrasound can be used for imaging a number of areas including Cardiology (heart), Breast, Obstetrics, Small parts (testicles, eyes, thyroid), Urology (bladder, urethra, kidney), Vascular (arteries, vein, intravascular ultrasound), Musculoskeletal and many more areas [35].

Then, echocardiography results of a patient will be given to the physician from echo room (laboratory). The physician assesses all the information of patient history, physical examination, and laboratory results to make decisions. The decision may be to prescribe drugs and discharge the patients with follow up instructions, or admission of the patients depending on the severity of the problem, or to refer patients for further laboratory test if sufficient information is not avail.

As it's seen on appendix I, 20 input attributes are noticed from Full Transthoracic Echocardiography Report. This are patient name, reason for echo, date of exam, referred by, card no, age, sex, Aortic root, Left atrium, Left ventricle in diastole, left ventricle in systole, posterior wall of LV, Interventricular septum, LV- ejection fraction, Main pulmonary artery diameter, Pericardial effusion, TR velocity, E_m/A_m velocity ratio, Rhythm, and target class of Diagnosis are input attributes.

For this study, 14 attributes and including a target class that were used to predict heart disease (see Table 4.1).

Table 4.1 Identified List of attribute from TTE report

List of Identified Attributes			
1	Age	8	Sex
2	Aortic root – diameter	9	LV- ejection fraction
3	Left atrium: (sys) diam.	10	Main Pulmonary Artery diameter
4	Left ventricle in: diastole	11	Pericardial effusion
5	Left ventricle in systole	12	TR Velocity
6	Posterior wall of LV	13	E_m/A_m velocity ratio
7	Interventricular septum	14	Rhythm

4.1.3. Data Mining Goals

The principal investigator takes on the task of translating the business objective to data mining objective. The data-mining Goal 1; given Transthoracic Echocardiography report, predict those who are likely to be diagnosed with heart disease and who are not. The data- mining goal 2; Identify important variable and relationship between input variable of echocardiography associated with the absence or presence of heart disease.

4.2. Data Understanding

In order to meet the general objective of this research, collecting representative subset of Echocardiographic data is the prerequisite. Data understanding begins with collection of initial data. The data collection process was carried out from International Cardiovascular Hospital which is situated in Addis Abeba, Ethiopia. Echocardiography data from International Cardiovascular Hospital have already collected by Abel [37] a total amount of 7339 patient's echocardiography report covering from October 2008 to March 2011.

However, in this research, from original datasets, data preprocessing which was conducted by the researcher eliminate 352 instances and left with 6987 datasets for training and testing purpose.

Series discussions were conducted with domain experts and health background follow classmate in the area of heart disease diagnosis before selecting the target dataset attributes. Because of these discussions, the researcher selected 15 input attributes for training and testing of decisions tree, neural network, and support vector machine algorithms. Table 4.2 has shown the 15 input attributes list along with their data description and data types. These attributes were selected due to their relevance or significance in diagnoses. Or in other words, patient name, reason for echo, date of exam, referred by, and card no were discarded due to less importance for heart diagnosis and privacy issues also considered, typically for patient name.

Table 4.2 List of selected attributes along with their description and data types

S. No	**Attribute Name**	**Description**	**Data types**
1.	Age	Age of the patient in years	Numeric
2.	Sex	Sex of the patient	Nominal
3.	Aortic root – diameter	Size of Aortic root – diameter in mm	Numeric
4	Left atrium: (sys) diameter	Size of left atrium: systole diameter in mm	Numeric
5.	Left ventricle in: diastole	Size of the left ventricle: diastole in mm	Numeric
6.	Left ventricle: systole	Size of the left ventricle: systole in mm	Numeric
7.	Posterior wall of LV	Size of posterior wall of LV in mm	Numeric
8.	Interventricular septum	Size of Interventricular septum in mm	Numeric
9.	LV- ejection fraction	Fraction of blood pumped out of ventricles with each beat in %	Numeric
10.	Main Pulmonary Artery diameter	Size of Main Pulmonary Artery diameter in mm	Numeric
11.	Pericardial effusion	Presence of an abnormal amount or character of fluid in the pericardial space	Ordinal
12	TR Velocity	Tricuspid Regurgitation velocity in cm/sec	Numeric
13.	E_m/A_m velocity ratio	The ratio between myocardial early and atrial peak velocities	Numeric
14.	Rhythm	Type of heart rhythm observed	Nominal
15.	Diagnosis	Either presence or absence of heart disease	Nominal

4.2.1. Descriptive Data Summarization and Visualizations

Descriptive data summarization provides the general statistics summarization and visualization for characteristics of the data and identifies the presence of noise or outliers. Data characteristics regarding to central tendency include mean, median, mode and mid range, while regarding to measure of data dispersion include quartiles, inter quartile range (IQR), and variance. Table 4.3 displays the valid number of instance, minimum, maximum, the mean and the standard deviation of Transthoracic echocardiography numerical datasets.

Table 4.3 Descriptive Data summarization of Attributes

Descriptive Statistics						
S.No	**Attributes**	**N**	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
1	Age	6987	4.00	95.00	47.7203	17.60771
2	Aortic Root Diam.	6987	17.00	44.00	28.8030	3.88041
3	Left Atrium Sys Diam.	6987	16.00	74.00	35.5953	6.88232
4	Left Ventricle Diam.	6987	24.00	80.00	47.4043	6.67716
5	Left Ventricle Systole	6987	12.00	69.00	30.6151	6.41677
6	Posterior Wall of LV	6987	6.00	19.00	10.0506	1.66734
7	Interventricular Septum	6987	6.00	19.00	9.9153	1.73828
8	LV Ejection Fraction	6987	15.00	74.00	53.4433	8.14633
9	Main Pulmo. Artery Diam.	6987	1.0000	3.1000	1.853588	.2400615
10	Tricusped Re. velocity	6987	.0000	5.9000	1.591813	.8918248
11	Em/Am velocity ratio	6987	.30	1.90	.9576	.25275

The remaining Transthoracic echocardiography attributes other than listed in table 4.3, namely Sex, Pericardial effusion, Rhythm, Diagnosis have a scale of measurement nominal and ordinal.

Typically, the target class of diagnosis needs to have balance level of observations. If not so, WEKA 3.5 have a functionality of 'SMOTE' to resample and balance the number of class labels for those attributes used as target class in model building process. Diagnosis classes had two class labels with balanced size which optimize the training process. As Figure 4.1 exhibit, the number of observations with diagnosis yes is 2700(38.6%) and the number of diagnosis with no is 4287 (61.4%). From this one can conclude, the class diagnosis have an observations of approximated balanced size

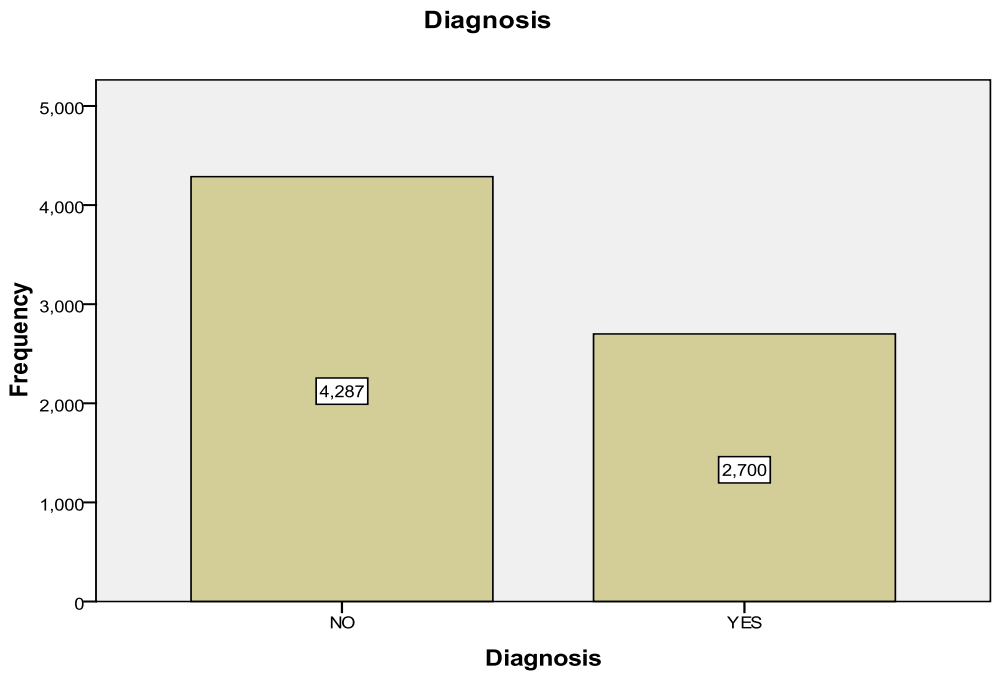

Figure 4.1 Statistic about class labels distribution in a datasets based on Diagnosis as a target class

Table 4.4 Diagnosis * Sex Crosstablulation

Diagnosis * Sex Cross tabulation				
Sex		Diagnosis		Total
	Female	No	Yes	
		2377 (55.44%)	1228 (45.48%)	3605
	Male	1910 (44.55%)	1472 (54.51%)	3382
Total		4287	2700	6987

Table 4.4 and Figure 4.2 shows that the distribution of heart disease between female and male. In the datasets, the fact revealed that Male who took the test is more likely to have the disease than female who took the test. Out of 4287 “NO” diagnosis, female accounts 55.44% while male accounts 44.55%. In “YES” diagnosis, this fact reversed. Out of 2700, “YES” diagnosis, 45.48% was female and 54.51% was male.

From the above paragraph and Table 4.4, a hypothesis can be setups like men are more favorable than women to heart disease. Moreover, this hypothesis needs further research to prove it.

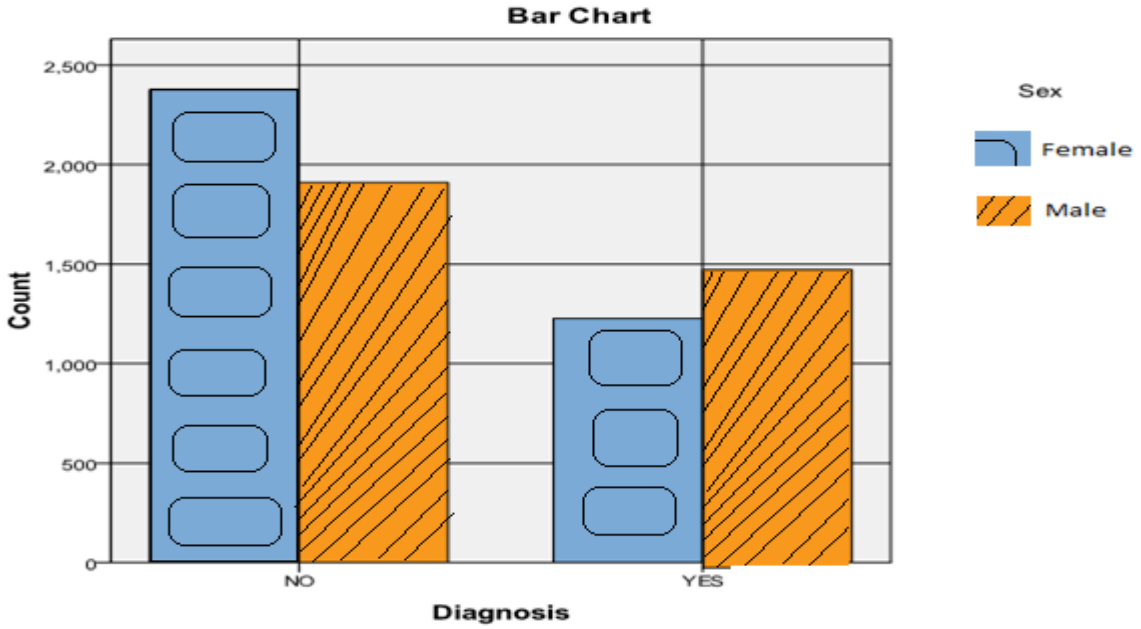

Figure 4.2 Heart Disease Distribution in a Gender attributes

4.3. Data Preprocessing

Today real word data are incomplete, inconsistent, noisy, redundant, missing due to their large size (gigabytes or more) and in some cases due to multiple sources. High quality data will lead to high-quality mining results and vice verse. Consequently, real world data of low quality needs preprocessing.

There are a number of data preprocessing tasks involved in this study such as data cleaning, handling outliers data, data integration and transformation, and data reduction techniques. Data integration is another major task, which merges data from multiple sources into a single and coherent data sources. Data transformation, such as normalization and discritization, may applied to optimize the accuracy and efficiency of mining algorithms. Data reduction is another major task which can reduce the data size to obtain quick processing time and save memory.

The objective of data processing at this stage is two-fold; to obtain data prepared in the form required by the data mining algorithms and to expose as much information as possible for data modeling.

4.3.1. Data Cleaning

Data cleaning which involve fill in missing values, smooth noisy data, identify or remove outliers etc.

4.3.1.1. Handling Missing Values

Most data encountered in practice, such as Transthoracic Echocardiography data as shown in Table 4.5, contains missing values. In the data missing values are frequently indicated by blank spaces, and sometimes unknown values placed in numeric field.

Missing values may occur for several reasons. Such as malfunctioning measurement equipment, lack of consistency with other recorded data and thus deleted, or respondent in a survey may refuse to answer certain questions such as age or income and data may not be recorded due to misunderstanding. But those missing values needs to be given significant attention. For example in this Transthoracic Echocardiography data age is the most frequent missing value; what do this things mean about the presence or absence of heart disease? Or does missing age have

something to do with heart disease or it is just because of some random events? To make an informed judgment about the missing values, it would be very appropriate to consult with someone who is domain expert. Thus, in consultation with domain expert, the missing values have no as such significant implication on heart disease diagnosis. Those missing values are ordinary or random.

To deal with missing values, alternatives have suggested by T. Larose[24] and Chakrabarti et al [25]: this are

- Ignore the missing value
- Replace the missing value manually
- Replace the missing value with a global constant to fill in the missing value
- Replace the missing value with some constant, specified by the analyst
- Replace the missing value with the field mean(for numerical variables) or the mode (for categorical variables)
- Replace the missing values with a value generated at random from the variable distribution observed.

In the data of Transthoracic Echocardiography (TTE) reports the missing values attributes are Age, Posterior Wall of LV, LV Ejection Fraction, Main Pulmo. Artery Diam, Tricusped Re. Velocity, Em/Am velocity ratio and rhythm with 262, 3, 9, 17, 12, 31, and 11 respectively. The frequencies of missing value are shown in Table 4.5. The Em/Am velocity ratio velocities variable were numeric but values like AF, NR, and RV characters was filled with.

Thus for this study, except Em/Am velocity ratio, the missing Value were replaced globally before applying learning algorithms. This deed takes placed in SPSS 17 for each missing value with the mean for numeric attributes and the mode for nominal ones. As observed in table 4.5, the missing value for categorical variables is Rhythm attribute only; as a result mode of rhythm variables were replaced for missed one. This decision taken because of the proportions of missing values for variables is small, and likely not to have more than a small effect on the results derived from the data.

For Em/Am velocity ratio attribute, the given values were considered as an outlier (typographical error) and a decision taken to remove all the records having a value AF (321), NR (30), and RV (1). This is mainly because of instances having AF, NR, or RV values are in turn incorporate unknown/outliers value in their other attributes and also since there is enough amount of data left for training and testing, it seem more appropriate the remove it .

The remaining data, that is 7339 - (321 + 30 + 1) = 6987, prepared in a way WEKA 3.7.5 required and were available for training and testing purpose.

Table 4.5 Statistics of Echocardiography datasets for Missing, Mean, and Mode of attributes

Statistics					
S.No	**Attributes Name**	**N**		**Mean**	**Mode**
		Valid	**Missing**		
1	Age	6725	262	47.72	60
2	Sex	6987			
3	Aortic Root Diam.	6987		28.83	29
4	Left Atrium Sys Diam.	6987		35.65	33
5	Left Ventricle Diam.	6987		47.43	46
6	Left Ventricle Systole	6987		30.70	28
7	Posterior Wall of LV	6984	3	10.06	11
8	Interventricular Septum	6987		9.93	10
9	LV Ejection Fraction	6978	9	53.33	55
10	Main Pulmo. Artery Diam.	6970	17	1.856714	2.0000
11	Pericardial Effusion	6987			
12	Tricusped Re. velocity	6975	12	1.604363	1.0000
13	Em/Am velocity ratio	6956	31	.96	1
14	Rhythm	6976	11		
15	Diagnosis	6987			

4.3.1.2. Handling Outliers Data

Outliers are extreme values that lie near the limits of the data range or go against the trend of the remaining data. This outlier's data may arise due to typographic or measurement error like the value of a nominal attributes is misspelled or in a numeric attributes correct ones can be possibly filled with incorrect values which results outliers that can be easily figured out by graphing one variable at a time. Further cause to noisy data can be faulty data collection instruments, data entry problems, data transmission problems, and inconsistencies in naming convention are the typical ones for noisy data to happen.

One graphical method for identifying outliers for numeric variables is to examine a histogram of the variable [24] and for categorical variable one rectangle is drawn for each know value and called commonly to as a bar chart [25].

According to Chakrabarti et al [25], a common rule of thumb for identifying suspected outliers is to single out values falling at least 1.5 * IQR above the third quartile or below the first quartile. Or outlier's detection defined as [24]

1. It is lower than Q1 – 1.5(IQR)
2. It is higher than Q3 + 1.5(IQR)

The Five-number summary (see Table 4.6) can be obtained by providing the lowest and the highest data values; this summary of a distribution consists of the median, the quartiles, Q1 and Q3, and the smallest and the highest values, written in the order minimum, Q1, Median, Q3, and Maximum[25]. To detect outliers IQR is a measure of variability that is mach more robust than the standard deviation [24].

Table 4.6 Statistics of Echocardiography datasets for Minimum, Maximum, and Percentiles of numeric attributes.

Statistics							
S.No	Attributes Name	N	Minimum	Maximum	Percentiles		
		Valid			25	50	75
1	Age	6987	4.00	102.00	33.0000	48.0000	60.0000
2	Aortic Root Diam.	6987	17.00	44.00	26.0000	29.0000	31.0000
3	Left Atrium Sys Diam.	6987	16.00	74.00	31.0000	35.0000	40.0000
4	Left Ventricle Diam.	6987	24.00	80.00	43.0000	47.0000	51.0000
5	Left Ventricle Systole	6987	12.00	69.00	27.0000	29.0000	33.0000
6	Posterior Wall of LV	6987	6.00	19.00	9.0000	10.0000	11.0000
7	Interventricular Septum	6987	6.00	19.00	9.0000	10.0000	11.0000
8	LV Ejection Fraction	6987	15.00	74.00	55.0000	55.0000	55.0000
9	Main Pulmo. Artery Diam.	6987	1.0000	3.1000	1.700000	1.900000	2.000000
10	Tricusped Re. velocity	6987	.0000	5.9000	1.000000	1.000000	2.100000
11	Em/Am velocity ratio	6987	.30	1.90	.8000	.9000	1.2000

Boxplots is a popular way of visualizing a distribution and incorporates the IQR (the box length), Median (a line within the box), the smallest (minimum) and the largest (maximum) observations [25].

In this study to administer the outlier's data, Boxplots are used. As it is revealed in Figure 4.3, the dark line in the middle of the boxes is the median attribute and half of the instances have a value greater than the median, and half have a value lower. The bottom of each figure box indicates the 25^{th} percentile and the top of the box represents the 75^{th} percentile. This means that 50% of the instances lie within the box. For example Figure 4.3, the box is mach shorter for Aortic root diameter attribute than age attribute. This is one clue that aortic root diameter attribute values varies less than age values.

The T-bars that extends from the boxes are called whiskers or inner fences. These extend to 1.5 times the height of the box. In Figure 4.3, the inner fences extend mach less for LV Ejection Fraction attribute compared to other attributes, this is an indication that LV Ejection Fraction varies less than all other attributes.

The points that do not fall in the inner fences are outliers. The asterisks denote this outlier's value. As all Figure 4.3 showed, there are several outliers for all picture except age attribute which have single outliers. As a result, the principal investigator had made a decision to replace those outliers value by the mean of the corresponding attribute as per the statistics of Table 4.6 which comprise Minimum, Maximum, First quartile, Second quartiles, and Third quartiles.

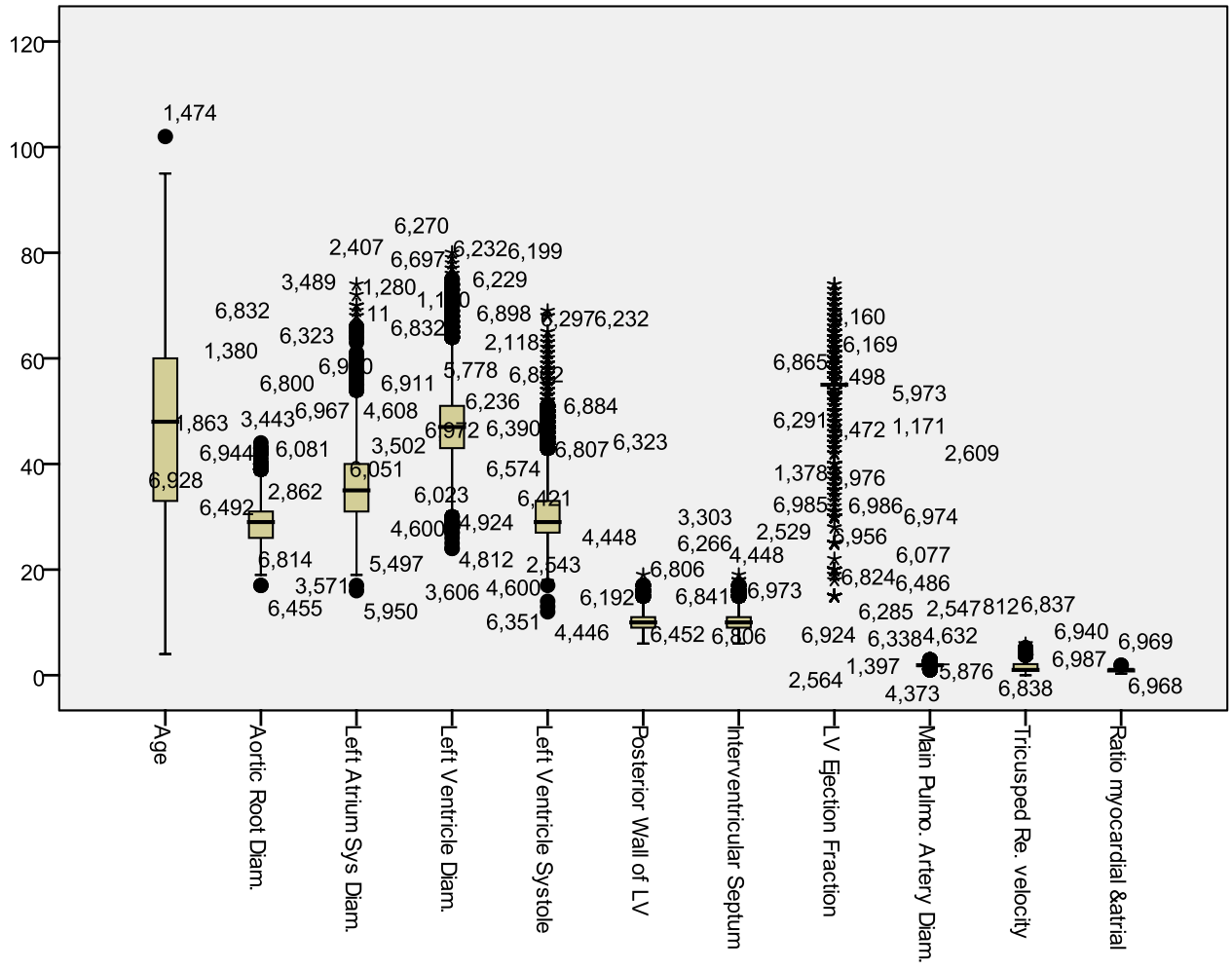

Figure 4.3 Boxplot for numeric attributes in the dataset to identify outliers.

4.3.1.3. Data Integration and Transformation

When working on a data mining problems, it is first necessary to bring the data together into the same platform. Integrating data from different sources usually pause many challenges; these are different department may use different style of record keeping, different time periods, different conventions, different primary key and many others. Fortunately, Echocardiography datasets have been stored as a single Ms-Word file such that each file brought to Ms-Excel file format to create database of echocardiography examinations. Then, from Ms-Excel datasets exported to SPSS 17 for descriptive analysis.

As per Han et al [7], Data transformation is about transforming or consolidating the data to make it appropriate for mining. Data transformation can involve the following operations:

- Smoothing; this techniques include binning, regression, and clustering which works to remove noise from the data.
- Generalization of the data; where low-level raw data are replaced by higher-level concepts. For example numeric attribute of age may be generalized to youth, middle age, and old age.
- Normalization; this operations performed on attribute to scaled the value to fall within a small specified range.

Among those techniques of transformation, Normalization the input values for Artificial Neural Network learning algorithms were takes place to speed up the learning process [7]. Typically, input values are normalized so as to fall between 0:0 and 1:0. Thus, Transthoracic Echocardiography (TTE) reports datasets was Normalize at ANN.

4.3.1.4. Data Reduction

Data reduction techniques are applied to obtain reduced datasets and yet maintain the integrity of the original data. Strategies of data reduction include data cube aggregation, attribute subset selections, dimensionality reduction, numerosity reduction, and discretization and concept hierarchy generations [7, 25]. In this study attribute subsets selections and dimensionality reduction applied.

4.3.1.5. Attribute Subsets Selections

Data sets for a given domain may contain dozen of attributes which reckon as irrelevant to the mining task. For example, the target variable for this study is Diagnosis, with potential patients being classified as to healthy or sick of heart disease; attributes such as the patient name or address are likely to be irrelevant. Attribute subset selection is one way to reduce the data size by removing irrelevant or redundant attributes. The fundamental reason of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes.

Methods of attribute subset selection include the following techniques [7, 25]. This are:

- ***Stepwise forward selection*** ; this procedure starts with null set of attributes as the reduced set and then the best original attributes is selected and added to the reduced set.
- ***Stepwise backward elimination***; this procedure follows exact reverse of stepwise forward selection. It starts with the full set of attributes and removes the worst attribute remaining in the set.
- ***Combination of forward selection and backward elimination***; at each step, the procedure selects the best attribute and removes the worst among the remaining attributes.
- ***Decision tree induction***; this method select subset attribute assume all attributes that do not appear in the tree are irrelevant.

The researcher therefore reduced the number of attributes from 15 to 8 by eliminating the redundant or less relevant variables using step forward selection. This selection technique assume empty set of attribute as the reduced set, then the best single is picked first and next best attribute condition to the first. The benefit of attribute subset selection is that the dimensionality of the solution space is also reduced, so that, for certain data mining algorithms become more efficient [24]. Besides, many irrelevant attributes will place unnecessary computational overhead on any data mining algorithms; even may cause the algorithms to give poor result [27]. Best First method selected 8 attributes from a total of 15 attributes. This are

Left Atrium systole diameter, Posterior wall of LV, LV Ejection Fraction, Main pulmonary artery diameter, pericardial effusion, Tricuspids R. Velocity, Em/Am velocity ratio, Rhythm, and Diagnosis attributes.

4.3.1.6. Dimensionality Reduction

According to Chakrabarti et al [25] in data reduction, data encoding or transformation are applied to reduce or compress representation of the original data. Data compressions are in two kinds, lossless and lossy. Lossless are if the original data can be reconstructed from the compressed data without any loss of information we called it lossless. Whereas lossy is if we reconstruct only an approximation of the original data, then the data reduction is called lossy.

The principal investigator in this research applied the lossless which does not lose any data in the compression process. In echocardiography datasets the attributes of gender (male, female), pericardial effusion (normal, mild, moderate, large, small, very small, significant), Rhythm (Sinus, AF), and diagnosis (Yes, No) have a categorical variable. Those attributes value are changed to numeric format and followed with data size reduction. The encoding techniques handled with SPSS 17, Transform menu of recode into same variable futures; this enables to change the existing variables value with the given numeric value without creating additional variables. Table 4.7 portrayed the detail of recoding with regard to attributes name, their old value and their new numeric value as well.

Table 4.7 Attribute Encoding New value for replacement of old value.

S.No	Attributes name	Original Value	Respective New Value
1	Sex	{Male, Female}	{1 ,0}
2	Pericardial effusion	{normal, Mild, Moderate, Large, small, very small, significant}	{1,2,3,4,5,6,7}
3	Rhythm	{Sinus, AF}	{1,2}
4	Diagnosis	{yes, no}	{1,0}

CHAPTER FIVE
DATA MINING AND EVALUATIONS OF MODELS

Once a good quality data is generated, a predictive model is created using classification algorithms such as decisions tree, artificial neural network, and support vector machine. The test option on those algorithms was 10 fold cross validation is used to split the data set into training and test set. Test results and corresponding discussions are presented on each experiment.

5.1. Experimental Setup

The echocardiography datasets was presented in a spreadsheet format. However, WEKA natively store data as an ARFF format; as a result the datasets are converted from a spreadsheet to ARFF. The ARFF file consists of a list of the instances, and attribute values for each instances separated by commas.

Mostly spreadsheet and database programs allow you to export data into a file in comma-separated value (CSV) format as list of records with commas between. Then you only load the file into text editor or word processor; add the dataset's name using the @relation tag, the attribute information using @attribute, and a data information with @data tag and save the file as raw text with .arff file format [17]. Figure 5.1 depicts the sample of machine understandable format of the dataset in WEKA employed for this study.

WEKA 3-7-5 supports many types of classification algorithms. Among the those algorithms, J48 is the classic classifier, which is the reimplementation of C4.5, and multilayer perceptron of neural network and Sequential minimal optimization of support vector machine are the available classifier on WEKA 3-7-5 [17].

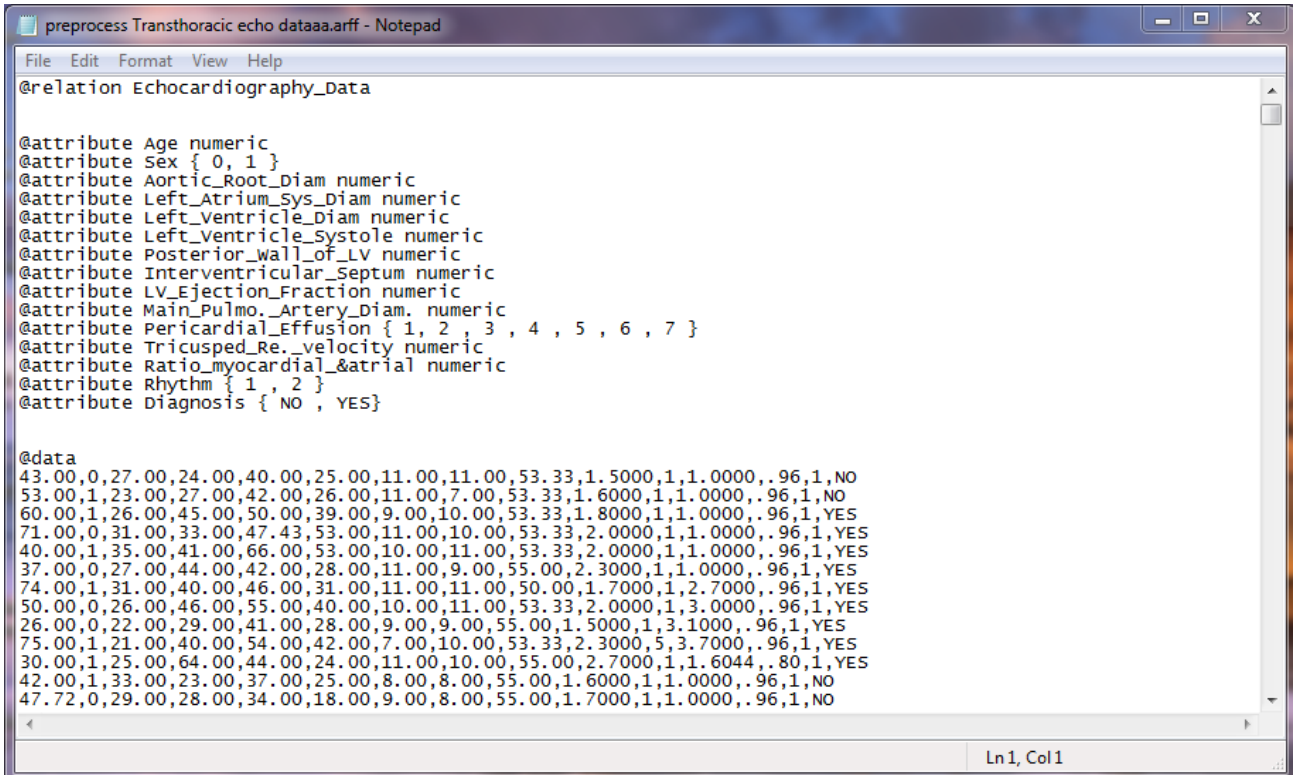

Figure 5.1 Sample Machine understandable format of the data set in WEKA

During data preprocessing stage, 352 instances removed and left with 6987 datasets for training and testing purpose. Hence, 6987 is the valid instances number for model building. On 6987 datasets, three supervised machine-learning algorithms applied.

The modeling phase in the data mining process of this investigation was carried out in two sub-phases

- Using all the 15 attributes of the TTE datasets
- Using 8 attributes of the TTE datasets based on algorithm that selects the best attributes

Then supervised machine learning algorithms of J48 (decisions Tree), SMO (Support Vector Machine) and Multilayer perceptron (Artificial Neural Network) were used for model building.

In this research, seven experiments conducted with the following research design of data mining algorithms, parameters, and attribute selection.

- ✓ Experiment I using J48 decisions tree with parameters of pruned and unpruned tree and 15 attributes.
- ✓ Experiment II using J48 decision tree with parameters of pruned and unpruned tree and 8 best selected attributes.
- ✓ Experiment III using Sequential Minimal Optimization algorithm of SVM with default parameters and 15 attributes.
- ✓ Experiment IV using Sequential Minimal Optimization algorithm of SVM with default parameters and 8 best selected attributes.
- ✓ Experiment V using Multilayer perceptron of ANN with default parameters and 15 attributes.
- ✓ Experiment VI using Multilayer perceptron of ANN with default parameters and 8 best selected attributes.
- ✓ Experiment VII using Multilayer perceptron of ANN with parameters adjustments of hidden units, learning rate, and smoothing factor and 15 attributes.

In this research, many experiments are put into test with different filter preprocessing techniques to derive a model with better accuracy rate. Hence, experiments with the application of resample results better accuracy. Regarding to test options, 10 fold cross-validations were employed to partition the dataset into training set and test set; this gives 10 approximate partitions of the datasets and from each partition 2/3 used for training and 1/3 used for testing purpose. This testing optimize that the random sampling is done in such a way as to guarantee that each class is properly represented in both training and test sets. In addition to this, there are other parameters under the more options button and in this case all are set to their default values.

5.2. J48 Experiments

In the first step, a predetermined training data set is analyzed by a classification algorithm to construct the model. This model is shown as classification rules or the decision tree. The second step is the classification. In this step, test data is used to verify the classification rules or the decision tree. If verification is an admissible rate then the rules are used to classify the new data. The accuracy of the model applied

to the test data is the ratio of its accurate classification to the all classes in the test data.

Ahead of the experimentations, the datasets instances are resample which produces a random subsample of a dataset using either sampling with replacement or without replacement. In the resample object editor box sampling with replacement or without replacement parameters can be set either false or true. Moreover, Decision trees are usually built in two steps. First, an initial tree is built till the leaf nodes belongs to a single class only; second, pruning is done to remove any overfitting to the training data [31]. In this paper, although, the tree generation took place in pruned and unpruned futures to realize the difference; factually, Table 5.1 results of decision tree accuracy rate have shown insignificant difference in either pruned or unpruned case. Nevertheless, significant difference observed with regard to number of leaves and size of tree.

By making use of WEKA 3-7-5, four experiments were carried out with different input parameters to construct the decision trees. To display the run parameters and the outputs of the respective experiments **Table 5.1 and 5.2** are used. As displayed in the tables, the different experiments were carried out based on generating a pruned and unpruned C4.5 decision tree using all 15 attributes and using best selected 8 attributes. This enables to realize the effect of tree pruning methods on classification accuracy, size of decision tree and model complexity when building a decision tree model versus unpruned tree.

All in all, a total of four experiments were took place with default value for all model except unpruned parameters are set to false and true value; this are

- Unpruned tree with 15 attributes and 8 best selected attributes
- Pruned tree with 15 attributes and 8 best selected attributes

Table 5.1 Experiment 1 Decision tree Results

Experiment 1	N	Attribute No	No of Leaves	Size of Tree	Time (Sec)	Accuracy %
Pruned Tree	6987	15	123	230	2.11	96.72
Unpruned Tree	6987	15	212	398	1.42	96.80

The first experiment provided with better accuracy as compared to the second. However, the decision tree generated from the two experiments was large, complex and difficult to generate understandable rules. Pruned tree of experiment 1 have produced a size of tree 230 and number of leaves with 123 was obtained from the training. Unpruned tree of experiment 1 have a size of tree and number of leaves produced were 398 and 212 respectively. If one needs to pass through all the nodes of this tree in order to come out with valid rule sets, it's cumbersome and difficult.

Thus, the researcher therefore reduced the number of attributes from 15 to 8 by eliminating the less relevant variables using step forward selection. Hoping, this elimination results more efficient algorithms and also remove unnecessary computational overhead on any data mining algorithms as section 4.5.4.1 explains. Based on this scheme an experiment was conducted and shown in the following Table 5.2. Unfortunately, after comparing results, the effect of best attributes selection was not there. However, tree-pruning effect was pronounced.

Table 5.2 Experiment 2 Decision tree Results

Experiment 2	N	Attribute No	No of Leaves	Size of Tree	Time (Sec)	Accuracy %
Pruned Tree	6987	8	91	171	1.11	95.89
UnprunedTree	6987	8	200	379	1.08	96.17

As table 5.2 of experiment 2 demonstrates, reduced attribute of Pruned tree, a size of tree and number of leaves produced from the training was 171 and 91 respectively. Similarly, Unpruned tree of experiment 2 have a size of tree and number of leaves produced were 379 and 200 respectively which can be seen the effect of tree pruning.

Generally, as the above Table 5.1 and 5.2 demonstrates, four decision tree experiments were conducted to TTE dataset. Experiment 1 performed well relative to the experiment 2 with respect to accuracy level of 96.80%. But with regard to number of leaves and size of tree experiment 2 highly favorable. The number of leaves and the corresponding sizes of the trees constructed from experiments 2 are less than those found from experiments 1. Most probably, this incidence happened on experiment 2 due to the fact that just having 8 attributes.

5.3. Decision Tree Evaluation

Four experiments were demonstrated based on all 15 attributes and 8 best selected attributes. Improved results achieved by experiment 1(Table 5.1) of pruned tree and this experiment is chosen for the following reason. First, better accuracy rate scored relative to experiment 2 (Table 5.2) results of reduced attribute. Secondly, within its own experiments, lower number of leaves and size of trees produced. The following confusion matrix (Table 5.3 of experiment 1 pruned tree) demonstrates the correctly classified instances versus incorrectly classified instances. From this table one can observe sensitivity, specificity and accuracy rate.

Sensitivity is ability of the test to correctly identify persons with the diseases (true positive) or the proportion of people with the disease who have a POSTIVE test for the disease; accordingly, pruned tree of experiment 1, sensitivity measures (2493/ (2493 + 142))* 100 = 94.61%. Specificity is ability of a test to correctly identify persons without the disease or proportion of people without a disease who have a NEGATIVE test for the disease. Table 5.3 figure of specificity measures (4265/ (4265+87))*100 = 98%

Table 5.3 Confusion Matrix Outcomes of a Decision Tree Prediction

		Predicted class	
		Yes	No
Actual class	Yes	2493 (TP)	142 (FN)
	No	87 (FP)	4265 (TN)

Table 5.3 indicates that the total test set (6987 records) provided to the program; the accuracy rate calculated, chapter 3(sections 3.5.1), as (TP +TN)/ (TP+TN+FP+FN). Accuracy scored was 96.72%; 6758 (96.72%) instances were correctly classified while the remaining 229 (3.28%) instances were classified incorrectly.

5.4. Sequential Minimal Optimization (SMO) Experiments

The WEKA is a comprehensive suite of java class libraries that implement many machine learning and data mining algorithms. In this study, SVM is the second data mining algorithms used. Support vector machine algorithms find the maximum separating hyperplane that have maximum distance between the nearest training tuples [7].

The experiments for SVM was conducted with same dataset size (N= 6987) as the same as for others algorithms and test options was also the default 10 fold cross-validation pertained. What makes this experiment peculiar from that of decision tree and neural network is with respect to resample preprocessing techniques. With the application of resample futures, the rate of accuracy was 90.63%. On the other hand, without resample futures, accuracy level gets better a little bit (90.88%). As results, resample techniques did not considered. The following table discloses the experimental results in brief.

Table 5.4 Experiment 3 Support Vector Machine Results

SMO Experiments 3	Instances	Attribute No	Time (Sec)	Accuracy %
SMO (1)	6987	15	1.69	90.88
SMO (2)	6987	8	0.38	90.56

Table 5.4 of Support Vector Machine experiments result have conducted with two scenarios; the first case is the experiment conducted with all 15 attribute of TTE and the second case was based on 8 selected attribute with DM search method of Best First.

As a result, accuracy scored for attributes number 15 and 8 is not as such significant especially in contrast with decision tree and neural network rate.

5.5. Support Vector Machine Evaluation

In this machine learning, two experiments conducted using the default parameters of Sequential Minimal Optimization (SMO) algorithms for training a support vector classifier. These two experiments schemas are; using all 15 attributes of TTE datasets and 8 best selected attributes.

TTE datasets was already preprocessed in chapter four for this algorithm as well. Nothing more or less preprocessing actions were taken. In fact, SMO globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default.

Between two experiments of support vector machine, the first experiments of SMO (1)(Table 5.4) selected. This experiments is the which uses all 15 attributes and it have less numbers of instances incorrectly classified relative to SMO (2) experiments. In other words, SMO (1) experiments have higher rate of accuracy as compared to SMO (2) experiment. The next table 5.5 discloses the confusion matrix

outcomes of the experiments. As results, sensitivity, specificity, and accuracy rate calculated.

Table 5.5 Confusion Matrix Outcomes of a Support Vector Machine Prediction

		Predicted class	
		Yes	No
Actual class	Yes	2280 (TP)	420(FN)
	No	217 (FP)	4070(TN)

Sensitivity (true positive rate) of the two-way table measures (2280/ (2280+420))*100 = 84.44% and the specificity (true negative rate) support vector machine experiments is (4070/ (4070+217))*100= 94.93%. The overall accuracy of this training algorithm was 90.88% which is significantly lower than the other two algorithms used in this study.

5.6. Multilayer Perceptron Experiments

The third data mining technique engaged in this study is neural networks. WEKA software package is used to build the neural network model and it employed back propagation algorithm. Much of the preprocessing tasks that mentioned in chapter four are also applicable for neural networks algorithms with a little addition of tasks like normalization of attributes.

ANN needs Normalization for each attribute in the training instances and can only process datasets when the values of attributes are numeric. This will help to speed up the learning phase. Therefore, to make the dataset suitable for neural network, values of all attributes were normalize to fall in the range of 0.0 – 1.0. The min-max normalization method applied as section 3.3 discussed.

The experiments for neural network were conducted based on the same data sets used as same as for others algorithms. The datasets instances are resample which results produce better accuracy rate. Neural network experiment use 6987 valid records and 10 fold cross-validation test options was employed as the same for other algorithms.

The following Table 5.6 discloses the experiments results conducted on TTE datasets. Two scenarios were assumed i.e. this are using all 15 attributes and 8 best selected attributes.

Table 5.6 Experiment 4 Artificial Neural Network Results

ANN Experiment 4	Instances	Attribute No	Test Mode	Time (Sec)	Accuracy
Mulitlayer Perceptron (1)	6987	15	10 Fold	48.63	95.21
Mulitlayer Perceptron (2)	6987	8	10 Fold	28.47	94.97

Among the most important parameters in training neural network is that of learning rate and smoothing factor (momentum) are archetypal one [36]. Backpropagation learns by iteratively processing a data set, comparing the network prediction for each tuple with the actual known target value [7]. For each training tuple, the weights are adjusted depending on the mean squared error. Learning rate determines how big a change must be made towards the correct value. That means, to adjust the weights do we take a giant step towards the correct value (large learning rate) or small step (small learning rate) [36]. A very high learning rate is not preferred since there would be giant oscillation as the network makes large adjustments for one pattern and another large change for the next pattern. Another important parameter during the training phase of neural network is smoothing factor (momentum); momentum tends to keep the change in the same direction from one iteration to the next.

Thus, in order to improve the performance of the model, an attempt was made to modify some of the parameters. With the intention of devising astonishing performance of a model, the researcher modified the number of hidden units to 7, the learning rate and smoothing factor (momentum) modified to 0.5 and 0.4 which were 0.3 and 0.2 in the default values respectively. As results, time taken to build model and the accuracy rate obtained from this experiment 5 was 39.38 seconds and **95.53**% respectively which execute superior than default parameters value. Moreover, further attempt was made to set the hidden units at different level and others parameters as well to improve the performance of the model and accuracy rate. Yet, it was not possible.

5.7. Artificial Neural Network Evaluation

The third and the last algorithms for this study evaluation taken on neural network. As the same manner as others algorithms; sensitivity, specificity and rate of accuracy were used as comparison techniques among learning algorithms.

During training, many experiments were taken to make sure the best possible results achieved; but the one with the best results sought was an experiment with the modified parameters. These changes happened to be on hidden units, learning rate, and smoothing factor (momentum) parameters; those change values were from a, 0.3, and 0.2 to 7, 0.5 and 0.4 values respectively. These changes of parameters for multilayer perceprton occur after many experiments.

Table 5.7 Confusion Matrix Outcomes of a Neural Network Prediction

		Predicted class	
		Yes	No
Actual class	Yes	2458 (TP)	177 (FN)
	No	135 (FP)	4217 (TN)

Sensitivity (true positive rate) of the two-way table of neural network prediction measures (2458 / (2458+117))*100 = 93.28% and the specificity (true negative rate) of neural network experiments was (4217 / (4217+135))*100= 96.89%. The overall accuracy of this training algorithm was 95.53% that is the second best accuracy rate after decision tree learning algorithms.

5.8. Discussion

5.8.1. Generated Rules from Decision Trees

Typical advantages of classification using decisions tree classifiers is to generate set of rules. Although in some cases the decision tree may be large and complex, it is generally fairly easy to follow anyone branch through the tree. It is possible to find out a set of rules simply by traversing the decision tree and generating a rule for each leaf and making a combination of all the tests found on the path from the root to the leaf. This produces a clear-cut rule.

Among decision tree experiments a model that was selected as a best one generated 123 rules for heart disease detection. The principal researcher selects and discusses a rule that only satisfy the assumptions of giving a correct result for at least 100 cases. Based on this assumption, eight rules discussed below and those rules represent more than 80% (5725) of the echocardiography datasets.

Among 8 rules to detect heart disease 6 of them were acceptable by the domain expert.

Rule 1: *if Left_Atrium_Sys_Diam < = 40 AND LV_Ejection_Fraction <=51 AND Tricusped_Re._velocity < = 2.5 AND Left_Ventricle_Systole > 38 : Diagnosis = YES (110.0/1.0)*

The above rule generated by Tree.J48 algorithms gave a correct result for 110 of the 111 cases it covers. The domain expert accepted this rule. A patients having left atrium systolic Diameter less than or equal to 40 millimeter and LV ejection fraction less than or equal to 51% blood pumped out of ventricles and

Tricuspid Velocity less than or equal to 2.5 centimeter per second and left ventricle systole greater than 38 millimeter would result LV systolic dysfunction disorder. This means left ventricle have a problem of contraction.

Rule 2: *if Left_atrium_sys_diam < = 40 and LV_ejection_fraction < = 51 and Tricusped_re._velocity > 2.5 : diagnosis = yes (177.0)*

This rule also accepted by domain expert. The second rule state a patient having a left atrium systolic diameter less than or equal to 40 millimeter and LV ejection fraction having less than or equal to 51% blood pumped out of ventricles and Tricuspid velocity is greater than 2.5 centimeter per second would result LV systolic dysfunction and may also result pulmonary hypertensive. Pulmonary hypertensive is an increase in blood pressure in the pulmonary arteries, pulmonary veins, or pulmonary capillaries (Lung vacillator) which lead to shortness of breath, dizziness, or other symptoms.

Rule 3: *If Left_Atrium_Sys_Diam < = 40 AND LV_Ejection_Fraction > 51 AND Interventricular_Septum > 11 AND Posterior_Wall_of_LV > 12 : Diagnosis = YES (163.0/2.0)*

This rule also accepted by domain expert. A patient having a disorder of left atrium systolic diameter less than or equal to 40 millimeters and LV ejection fraction is greater than 51% blood pumped out of ventricles and Interventricular septum is greater than 11 millimeter and posterior wall of LV greater than 12 millimeter definitely results disorder of Left Ventricular Hypertrophy (LVH). This is maladaptive response to chronic pressure overload

and an important risk factors for atrial fibrillation, and sudden death in patients with hypertension.

Rule 4: *If Left_Atrium_Sys_Diam > 40 AND LV_Ejection_Fraction < = 53: Diagnosis = YES (673.0/1.0)*

The above rule is not acceptable by domain expert. It is very difficult to judge based on these two variables only.

Rule 5: *If Left_Atrium_Sys_Diam > 40 AND LV_Ejection_Fraction > 53 AND Em/Am velocity ratio < = 0.7 AND Age > 27: Diagnosis = YES (427.0)*

As the same to rule four, this rule also is not acceptable or not interesting to domain experts. The normal range for Em/Am velocity ratio is between 0.75 – 1.5. if Em/Am velocity ratio were > 1.5 and if age of the patient greater than 60 would results Grade I diastolic dysfunction disorder. Diastolic dysfunction or diastolic heart failure refers to decline in performance of one or both ventricles of heart during the time phase of diastole.

Rule 6: *If Left_Atrium_Sys_Diam > 40 AND LV_Ejection_Fraction > 53 AND Em/Am velocity ratio > 0.7 AND Tricusped_Re._velocity > 2.6 AND Interventricular_Septum > 7: Diagnosis = YES (169.0/1.0)*

The above rule accepted by domain expert with a set of conditions. This are if Em/Am velocity ratio attribute have a value greater than 1.5 and interventricular septum have a value greater than 11 millimeter and all the

other attribute retaining there values; it results hypertensive heart disease and also cause pulmonary hypertensive.

Rule 7: *If Left_Atrium_Sys_Diam < = 40 AND LV_Ejection_Fraction > 51 AND Interventricular_Septum < = 11 AND Tricusped_Re._velocity < = 2.6 AND Main_Pulmo._Artery_Diam < = 2 AND Pericardial_Effusion = 1 AND Posterior_Wall_of_LV < = 11 AND Em/Am velocity ratio > 0.7 AND Aortic_Root_Diam < = 26: Diagnosis = NO (145.0/4.0)*

Rule number seven, also an acceptable rule that would be more suitable if Aortic root diam have a value less than or equal to 37. This case may depend on the physician or the hospital whose trend on aortic root diam variable followed.

Rule 8 : If Left_Atrium_Sys_Diam < = 40 AND LV_Ejection_Fraction > 51 AND Interventricular_Septum < = 11 AND Tricusped_Re._velocity < = 2.6 AND Main_Pulmo._Artery_Diam < 2 AND Pericardial_Effusion = 1 AND Posterior_Wall_of_LV < = 11 AND Em/Am velocity ratio > 0.7 AND Age > 17: Diagnosis = NO (3861.0/61.0)

According to domain experts, this is the perfect rule to diagnose a patient with NO heart disease.

5.8.2. Evaluation of Mining Goals

Two mining goals were defined based on detection of heart disease and objectives of this research. Best accomplish algorithms of J48 decision tree show that all the stated goals achieved.

Goal 1: given Transthoracic Echocardiography report, predict those who are likely to be diagnosed with heart disease and who are not. Providing echocardiography datasets, one can diagnose patients with heart disease. For example, Left Atrium Diam < = 10 and LV ejection fraction < = 51 and Tricuspid velocity <= 2.5 and Left ventricle systole > 38 would definitely diagnose a patient with hart disease with 110 supporting cases out of 111. As these values obtained from echocardiography datasets of patients, doctors could carry out further heart examination.

Goal 2: Identify important variable and relationship between input variable of echocardiography associated with the absence or presence of heart disease. The Tree.J48 algorithms identify most significant factors in predicting heart disease is Left atrium systole diameter, LV ejection fraction, Tricuspid velocity, Interventricular Septum, and Em/Am velocity ratio are the prominent ones. On the other hand, less important variables for detection of heart disease are Sex, left ventricle in diastole, and rhythm are the typical one.

5.8.3. Model Comparison

Initially, this research papers have set out general and specific objectives. Among those objectives, one of this is to identify which data mining techniques outperform better. Accordingly, each experiment carried out in this research has employed decision tree, support vector machine and artificial neural network learning techniques.

The model comparison presented in Table 5.8 focuses on learning algorithms of which have considered as better experiments outcomes. These are; decision tree with an accuracy rate 96.72% on experiment 1, support vector machine with an accuracy rate of 90.88% on experiment 3, and on experiment seven neural network scored an accuracy rate of 95.53%.

As depicted in Table 5.8, the comparison parameters of machine learning algorithms is TP Rate (correctly classified), FP Rate (incorrectly classified), Precision, and Recall were used. When sensitivity increased, the specificity decreases and vice verse. There is a need to weigh consequences of leaving cases undetected (false negatives) against erroneously classifying healthy persons as diseases (false positives). For example, looking at column of Table 5.8, Trees.J48 have a specificity of 0.98 (98%) and relatively decreased sensitivity of .946 (94.6%).

Table 5.8: Model Comparison for experiments with superior performance only

No	DM algorithms	Accuracy	TP Rate	FP Rate	Precision	Recall	Class
1	Decision Tree	96.72%	0.98	0.054	0.968	0.98	NO
			0.946	0.02	0.966	0.946	YES
2	Multilayer perceptron ANN	95.53%	0.969	0.067	0.96	0.969	NO
			0.933	0.031	0.948	0.933	YES
3	SMO SVM	90.88%	0.949	0.156	0.906	0.949	NO
			0.844	0.051	0.913	0.844	YES

The FP Rate column of Table 5.8 entangled with TP Rate. Looking at algorithms of Trees.J48, .98 (98%) is patients without a disease correctly identified as NEGATIVE; this means the remaining 0.02 (2%) are patients who was identified incorrectly as positive while they were NEGATIVE.

The other two column of Table 5.8, Precision and Recall, are other parameters that used in many information retrieval researches [17]. Given a query, a web search engine produces a list of hits that represent relevant documents; recall represent, number of documents retrieved that are relevant divided by total number of documents that are relevant. Precision also represent, number of documents retrieved that are relevant divided by total number of document that are retrieved [17].

Positive precision of Tree.J48, Function.Multilayer perceptron, and Function.SMO respectively are 0.966, 0.948, and 0.913. Therefore, Tree.J48 algorithms labeled patients as belonging to class YES and actually patient labeled as class YES are 96.6% successful which is exceeds the other algorithms. This means, from the total prediction of 2493 patients are actually diagnosed as YES out of 2580. Only 87 patients are labeled as YES while they were actually NO (see Table 5.3). Negative Precision of Tree.J48, Function.Multilayer perceptron, and Funtion.SMO are 0.968, 0.96, and 0.906 respectively. Still, Tree.J48 done well. Therefore, Tree.J48 model labeled patients as belonging to class NO and actually patient labeled as class NO are 96.8% doing well. 4265 NO prediction of patients are actually diagnosed as NO out of the total NO prediction of 4407 (see Table 5.3).

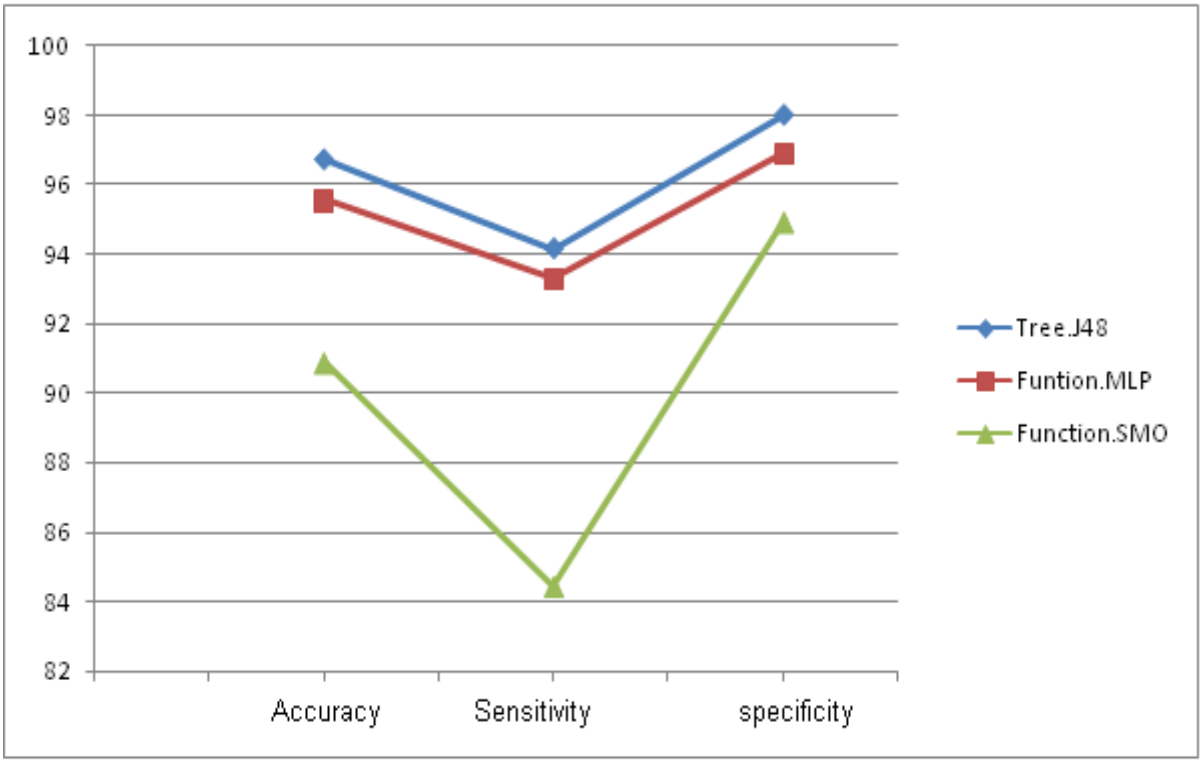

Figure 5.2 Comparison of three models, Tree.J48 providing highest accuracy, Sensitivity, and Specificity

In conclusion, among those three algorithms in Table 5.8, to select the model which outperforms others, the researcher picks specificity and sensitivity measurements. Therefore, specificity results for Trees.J48, Function.Multilayer perceptron, and Function.SMO are 0.98, 0.969, and 0.949 respectively; so that Trees.J48 has incorrectly classified very little (2%) patients as if they were POSTIVE relative to the

other algorithms. By the same token, sensitivity results for Trees.J48, Function.Multilayer perceptron, and Function.SMO are 0.946, 0.933, and 0.844 respectively; still Trees.J48 has incorrectly classified very little (5.4%) patients as if they were NEGATIVE. Decision tree (J48) is proven to be the best with the given datasets.

5.9. KBS Prototype Development

All trained models in this research could be used to detection of heart disease. Though, J48 decision tree algorithms perform classification and prediction of heart disease with minimal rate of error. As results, the KBS development attempt based on the extracted knowledge of J48 decision tree.

This prototype is based on the rule identified in section 5.8.1. These rules represent more than 80% of the training and testing dataset used in this study.

The Heart Detection program contains MS access database, the MS visual basic program hosting the classification rule. This is tools selected due to the convenience and familiarity of the researcher to those tools.

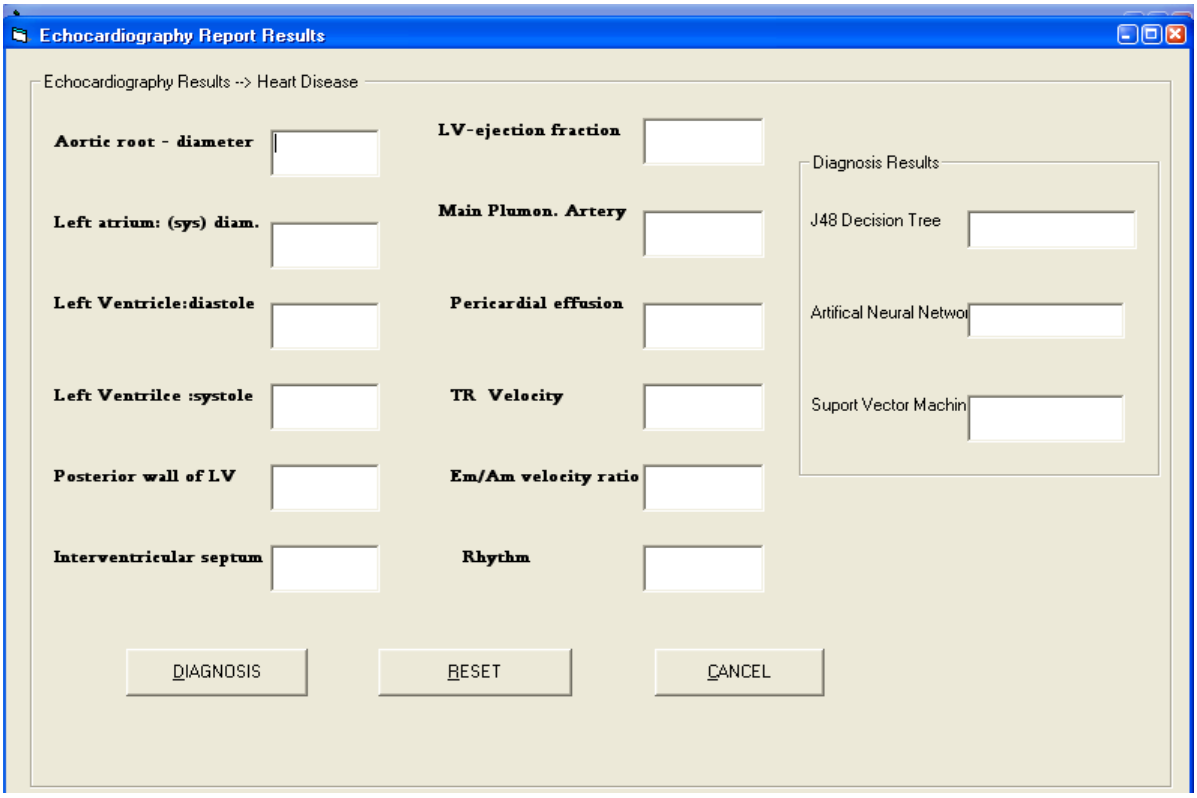

Figure 5.3 Prototype of the interface to predict Heart disease

CHAPTER SIX
CONCLUSIONS AND RECOMMENDATIONS

6.1. Conclusions

The objective of this research was to extract hidden knowledge from echocardiography datasets using data mining techniques. The experiments carried out in this research using decision trees, support vector machine, and neural network. Besides, the results were promising and encouraging especially relative to literature of similar research.

In data mining application, first the data in hand and the business problem to be solved must be analyzed and understood very well. Suitable mining techniques also play an important role for successful data mining application along with data preprocessing in the present study. Mach emphasis is given for business understanding and data understanding to make sure the best possible results obtained. To clean the data in this study, missing values and outliers data were handled. Moreover, data integration and transformation were managed.

In this research, the methodology employed was Hybrid Data mining process model; it involves six steps and the principal researcher rigorously passes through all the steps and iterated as needed. A total of 6987 patient's records were used to build and test data mining algorithms. In order to build the models that can predict a patient's status of heart either healthy or sick, several experiments were conducted with diverse parameters values. The best performing learning algorithms, found in this research, is J48 decision tree with an accuracy rate of 96.72%, sensitivity 94.6%, and with specificity of 98%.

In general, encouraging result is obtained by employing J48 decision tree as compared to artificial neural network and support vector machine learning. It is also observed decision tree outputs sought more applicable and appropriate to the

problem domain since its simplicity and easily understandable rules that can be used even by non-technical health care professional and policy makers as well.

In this study, though socio-demographic variables have a great effect on heart disease, we are unable to get the necessary records to come up with better and all rounded prediction of heart disease.

6.2. Recommendations

The researcher makes the following recommendations based on the result of this study.

- In this research, an attempt has been made to apply data mining technology to heart disease diagnosis using Echocardiography report variables only. However, there are a number of other socio-demographic variable, patient history, and physical examination which prominently determine heart disease. So that it remains to investigate those variables along with echocardiography report dataset to come up with generic predictive model.
- Decision tree, neural network, support vector machine were the selected model in this study. However, more machine learning algorithms along with much larger data size needs to be taken to realize the effects and optimize the prediction of heart disease.
- In this study J48, decision tree achieved promising results. Hence, an attempt has been made to develop/design Knowledge Based System (KBS). However, there is a need to further development of KBS along with domain experts. It is a research direction.
- At normal circumstances, each patient's echocardiography examination results stored as a single file. This makes it difficult to utilize the information in any way. For that reason, creating a database for echocardiography examination results and other examination results would be helpful for application of data mining techniques and for other purpose as well.
- In data mining projects or study, data preparation is the major challenge and consumes majority time of the process. Therefore, statistical and data mining techniques should be implemented thoroughly to acquire cleaned and quality data.

REFERENCES

1. Sellapan palaniappan and Rafiah Awang (2008). Intelligent Heart disease predication system using data mining techniques, IJCSNS International Journal of Computer Science and Network Security. 8(8)
2. Yosawin Kangwanariyakul, Chanin Natasenamat, Tanawut Tantimongcolwat, Thankakorn Naenna. (2010). Data mining of Magnetocardiograms for prediction of Ischemic heart Disease. EXCLI Journal 9:82-95
3. Latha Parthiban and R. Subramanian. (2007). Intelligent Heart disease prediction system using CANFIS and Genetic Algorithm, International Journal of Biological and Life Sciences 3(3)
4. Tahseen A. Jilani, Huda Yasin, Madiha Yasin, Cemal Ardil. (2009). Acute Coronary Syndrome Predication Using Data mining techniques - An Application, International Journal of Information and Mathematical Sciences. 5(4)
5. Shantakumar B.patil and Y.S. Kumaraswamy. (2009). Intelligent and Effective heart Attack prediction system using Data mining and Artifical Neural network, European Journal of Scientific Research. 31(4)
6. Aruro Evangelista, Frank Flachskampf, Patrizio Lancellotti, Luigi Badano, Rio Aguilar, Mark Monaghan, Jose Zamorano, and Petros Nihoyannopoulos. (2008). European Associations of echocardiography recommendations for standardization of performance, digital storage and reporting of echocardiographic studies, European Journal of echocardiography 9:438-448
7. Jiawei Han and Micheline Kamber. (2006). Data Mining: Concepts and Techniques. San Francisco. Morgan Kaufmann Publishers.
8. David Hand, Heikk Mannila and Padhraic Smyth . (2001). Principles of Data Mining. The MIT Press
9. Olivid Parr Rud. (2001). Data mining cookbook: Modeling data for Marketing, Risk, and customer relationship management. Wiley Computer publishing
10. Allan G. Bluman. (1998). Elementary statistics ; A step by step approach, WCB McGraw-Hill
11. Richard D. De Veaux. (2009). Successful Data mining in practice. Williams College.

12. Passamani, Eugene. (2008). "Heart" Microsoft Encarta. Redmond, WA: Microsoft Corporation.
13. J. Malcolm O. Arnold (MD). (2008). Heart Failure. Internet URL: http://www.merckmanuals.com/home/heart_and_blood_vessel_disorders/heart_failure/heart_failure.html (Accessed data: February 2012)
14. Joseph A. Kisslo, David B. Adams, Graham J. Leech. Essentials of echocardiography #1: Two-Dimensional Echocardiography in the Normal Heart
15. Dr. Abdulla M. Abdulla. (2010). Echocardiogram. Internet URL: http://www.heartsite.com/html/echocardiogram.html. (Accessed date: January 2012)
16. WHO (2011), Fact Sheet: Chronic diseases and their common risk factors. World Health Organization, URL: www.who.int/chp
17. Ian H. Witten, Eibe Frank. (2005). Data Mining practical Machine Learning tools and Techniques. Second Edition, Morgan Kaufmann publishers
18. Overview of Addis Ababa city solid waste management system. (2010). Addis Ababa, Ethiopia.
19. The city government of Addis Ababa the PEFA assessment report FWC beneficiaries. (2009). 11.
20. Ajith Abraham. Artificial Neural Networks. *Oklahoma State University, Stillwater. USA*
21. K. Srinivas, Dr. G. Raghavendra Rao, and Dr. A. Govardhan. (2011). "Survey on Prediction of Heart morbidity using Data mining techniques", International Journal of Data Mining and Knowledge Management process. 1(3)
22. Internet source: http://www.addiscardiac.com/research.php (Accesses date: march 2012)
23. Thomas A Gaziano, September. (2007). "Economic burden and the cost-effectiveness of treatment of cardiovascular diseases in Africa., group.bmj.com publisher
24. Daniel T. Larose. (2005). Discovering Knowledge in Data; An introduction to Data Mining. A John Wiley & Sons, Inc. Publication
25. Soumen Chakrabarti, Earl Cox, Eibe Frank, Ralf Hartmut Güting, Jaiwei Han, Xia Jiang, Micheline Kamber, Sam S. Lightstone, Thomas P. Nadeau, Richard E. Neapolitan, Dorian Pyle, Mamdouh Refaat, Markus Schneider, Toby J.

Teorey, Ian H. Witten. (2009). Data Mining Know it All. Morgan Kaufmann Publishers

26. Xindong Wu , Vipin Kumar , J. Ross Quinlan , Joydeep Ghosh , Qiang Yang, Hiroshi Motoda , Geoffrey J. McLachlan , Angus Ng , Bing Liu , Philip S. Yu, Zhi-Hua Zhou , Michael Steinbach , David J. Hand , Dan Steinberg. (2007). Top 10 algorithms in data mining. Springer-Verlag London Limited
27. Max Bramer. (2007). Undergraduate Topics in Computer Science; Principles of Data Mining. Springer-Verlag London Limited
28. Michaelw Berry, Murray Browne. (2006). Lecture Notes in Data Mining. World Scientific Publishing Co. Pte. Ltd. *University Of Tennessee, USA*
29. Jiawei Han and Micheline Kamber. (2000). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers
30. Te-Ming Huang, Vojislav Kecman, IvicaKopriva. (2006). Kernel Based Algorithms for Mining Huge Data Sets. Springer-Verlag Berlin Heidelberg
31. Yike Guo, Robert Grossman. (2002). High Performance Data Mining Scaling Algorithms, Applications and Systems. Kluwer Academic Publishers
32. Sumana Sharma, Kweku-MuataOsei-Bryson. (2008). Framework for formal implementation of the business understanding phase of data mining projects; Virginia Commonwealth University, United States, Elsevier Ltd
33. Roberto M. Lang, Michelle Bierig, Richard B. Devereux, Frank A. Flachskampf, Elyse Foster, Patricia A. Pellikka, Michael H. Picard, Mary J. Roman, James Seward, Jack S. Shanewise, Scott D. Solomon, Kirk T. Spencer, Martin St John Sutton, and William J. Stewart. (2005). Recommendations for Chamber Quantification: A Report from the American Society of Echocardiography's Guidelines and Standards Committee and the Chamber Quantification Writing Group, Developed in Conjunction with the European Association of Echocardiography, a Branch of the European Society of Cardiology, Journal of the American Society of Echocardiography 18(12)
34. Ernesto E. Salcedo. (2006). Echocardiography in Heart Failure—Current Applications, University of Colorado Health Sciences Center
35. Overview of Ultrasound Births Medical Ultrasound Society (BMUS), www.BMUS.ORG Accessed Date: February 2012

36. Shegaw Anagaw. 2002. Application Of Data Mining Technology To Predict Child Mortality Patterns: The Case Of Butajira Rural Health Project (BRRHP); M.Sc. Thesis, School of Information Science, Addis Abeba University, Addis Abeba, Ethiopia.
37. Abel Damtew. (2011). Designing A Predictive Model For Heart Disease Detection Using Data Mining Techniques; M.Sc. Thesis, Health informatics, Addis Abeba University, Addis Abeba, Ethiopia

APPENDICES

Appendix I: Transthoracic Echocardiography Report

Name: Age: 89yrs Gender: Female

Reason for echo:		Date of exam: 18/01/10
Referred By:	**Card No: 19782**	

Measurements **Ref Feigenbaum/Otto**

	Findings		Findings
Aortic root – diameter	27	LV- ejection fraction	55%
Left atrium: (sys) diam.	**42**	Main Pulmonary Artery diameter	**2.4**
Left ventricle in: diastole	47	Pericardial effusion	Normal
Left ventricle in systole	31	TR Velocity	**3.2**
Posterior wall of LV	**13**	E_m/A_m velocity ratio	**0.7**
Interventricular septum	**12**	Rhythm	Sinus

Summary

Normal sized aortic root and main pulmonary artery. Internal cardiac cavity dimensions are within normal range, except for the LA. No regional or global wall motion abnormality observed. Normal LVEF (55%). The LVPW is thickened. No intracavitary mass or vegetation noted. The Pericardium looks normal, with no effusion or abnormal thickening.Both inter ventricular septum and inter atrial septum are intact with no distorted shape or abnormal motion.

Except for severely calcified aortic valve, the cardiac valves leaflets, annular portion and subvalvular structures are normal, with adequate cusp separation and excursion.

CW, PW and color flow Doppler examination revealed mild to moderate AR and TR.

Transmitral Doppler inflow velocity ratio is reversed indicating inadequate LV compliance.

Conclusion:- *The above findings are consistent with:*

Pulmonary hypertension
LV diastolic dysfunction.
Concentric LV hypertrophy.
Mild AR.

Gizaw Erena; MD

Consultant Cardiologist

Appendix II: Rules Generated from J48 Classifier

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: Echocardiography_Data-weka.filters.unsupervised.instance.Resample-S1-Z100.0

Instances: 6987

Attributes: 15

- Age
- Sex
- Aortic_Root_Diam
- Left_Atrium_Sys_Diam
- Left_Ventricle_Diam
- Left_Ventricle_Systole
- Posterior_Wall_of_LV
- Interventricular_Septum
- LV_Ejection_Fraction
- Main_Pulmo._Artery_Diam.
- Pericardial_Effusion
- Tricusped_Re._velocity
- Ratio_myocardial_&atrial
- Rhythm
- Diagnosis

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

Left_Atrium_Sys_Diam <= 40

| LV_Ejection_Fraction <= 51

| | Tricusped_Re._velocity <= 2.5

| | | Left_Ventricle_Systole <= 38

| | | | LV_Ejection_Fraction <= 25

| | | | | Ratio_myocardial_&atrial <= 0.7: YES (4.0)

| | | | | Ratio_myocardial_&atrial > 0.7: NO (12.0/1.0)

| | | | LV_Ejection_Fraction > 25

| | | | | Left_Ventricle_Diam <= 38

| | | | | | Ratio_myocardial_&atrial <= 0.9: YES (3.0)

| | | | | | Ratio_myocardial_&atrial > 0.9: NO (11.0/1.0)

| | | | | Left_Ventricle_Diam > 38

| | | | | | Posterior_Wall_of_LV <= 12

| | | | | | | Main_Pulmo._Artery_Diam. <= 1.6

| | | | | | | | Ratio_myocardial_&atrial <= 0.7: YES (4.0)

| | | | | | | | Ratio_myocardial_&atrial > 0.7: NO (9.0/1.0)

| | | | | | | Main_Pulmo._Artery_Diam. > 1.6

| | | | | | | | Left_Ventricle_Diam <= 55

| | | | | | | | | LV_Ejection_Fraction <= 49: YES (95.0/1.0)

| | | | | | | | | LV_Ejection_Fraction > 49

| | | | | | | | | | Ratio_myocardial_&atrial <= 0.6: YES (26.0)

| | | | | | | | | | Ratio_myocardial_&atrial > 0.6

| | | | | | | | | | | Interventricular_Septum <= 8: YES (14.0)

| | | | | | | | | | | Interventricular_Septum > 8

| | | | | | | | | | | | Main_Pulmo._Artery_Diam. <= 2

| | | | | | | | | | | | Interventricular_Septum <= 11

| | | | | | | | | | | | | Left_Ventricle_Diam <= 43: YES (5.0)

| | | | | | | | | | | | | Left_Ventricle_Diam > 43

| | | | | | | | | | | | | | Ratio_myocardial_&atrial <= 0.8

| | | | | | | | | | | | | | | Posterior_Wall_of_LV <= 10.06: NO (4.0/1.0)

| | | | | | | | | | | | | | | Posterior_Wall_of_LV > 10.06: YES (4.0)

| | | | | | | | | | | | | | Ratio_myocardial_&atrial > 0.8: NO (13.0)

| | | | | | | | | | | | Interventricular_Septum > 11: YES (4.0)

| | | | | | | | | | | | Main_Pulmo._Artery_Diam. > 2: YES (11.0)

| | | | | | | | Left_Ventricle_Diam > 55

| | | | | | | | | Tricusped_Re._velocity <= 1.5

| | | | | | | | | | Left_Ventricle_Systole <= 37: YES (7.0)

| | | | | | | | | | Left_Ventricle_Systole > 37: NO (2.0)

| | | | | | | | | Tricusped_Re._velocity > 1.5: NO (6.0)

| | | | | | Posterior_Wall_of_LV > 12: YES (59.0)

| | | Left_Ventricle_Systole > 38: YES (110.0/1.0)

| | Tricusped_Re._velocity > 2.5: YES (177.0)

| LV_Ejection_Fraction > 51

| | Interventricular_Septum <= 11

| | | Tricusped_Re._velocity <= 2.6

| | | | Main_Pulmo._Artery_Diam. <= 2

| | | | | Pericardial_Effusion = 1

| | | | | | Posterior_Wall_of_LV <= 11

| | | | | | | Ratio_myocardial_&atrial <= 0.7

| | | | | | | | Age <= 38

| | | | | | | | | Left_Atrium_Sys_Diam <= 34

| | | | | | | | | | Main_Pulmo._Artery_Diam. <= 1.6

| | | | | | | | | | | Age <= 22: NO (2.0)

| | | | | | | | | | | Age > 22: YES (4.0)

| | | | | | | | | | Main_Pulmo._Artery_Diam. > 1.6: NO (11.0)

| | | | | | | | | Left_Atrium_Sys_Diam > 34: YES (14.0)

| | | | | | | | Age > 38

| | | | | | | | | Ratio_myocardial_&atrial <= 0.6

| | | | | | | | | | Posterior_Wall_of_LV <= 9

| | | | | | | | | | | Sex = 0: NO (5.0/1.0)

| | | | | | | | | | | Sex = 1: YES (8.0)

| | | | | | | | | | Posterior_Wall_of_LV > 9

| | | | | | | | | | | LV_Ejection_Fraction <= 54: YES (2.0)

| | | | | | | | | | | LV_Ejection_Fraction > 54

| | | | | | | | | | | | Sex = 0: NO (25.0/1.0)

| | | | | | | | | | | | Sex = 1

| | | | | | | | | | | | | Left_Ventricle_Systole <= 27: NO (10.0)

| | | | | | | | | | | | | Left_Ventricle_Systole > 27

| | | | | | | | | | | | | | Left_Ventricle_Diam <= 48: YES (4.0)

| | | | | | | | | | | | | | Left_Ventricle_Diam > 48

| | | | | | | | | | | | | | | Posterior_Wall_of_LV <= 10.06: NO (9.0/1.0)

| | | | | | | | | | | | | | | Posterior_Wall_of_LV > 10.06: YES (3.0/1.0)

| | | | | | | | | Ratio_myocardial_&atrial > 0.6

| | | | | | | | | | LV_Ejection_Fraction <= 53.33

| | | | | | | | | | | LV_Ejection_Fraction <= 52: YES (3.0)

| | | | | | | | | | | LV_Ejection_Fraction > 52: NO (6.0)

| | | | | | | | | | LV_Ejection_Fraction > 53.33: NO (94.0/2.0)

| | | | | | | Ratio_myocardial_&atrial > 0.7

| | | | | | | | Age <= 17

| | | | | | | | | Aortic_Root_Diam <= 26: NO (145.0/4.0)

| | | | | | | | | Aortic_Root_Diam > 26

| | | | | | | | | | Left_Ventricle_Diam <= 44: NO (12.0)

| | | | | | | | | | Left_Ventricle_Diam > 44

| | | | | | | | | | | Aortic_Root_Diam <= 30

| | | | | | | | | | | | Main_Pulmo._Artery_Diam. <= 1.6: NO (2.0)

| | | | | | | | | | | | Main_Pulmo._Artery_Diam. > 1.6

| | | | | | | | | | | | | Tricusped_Re._velocity <= 2.4: YES (11.0)

| | | | | | | | | | | | | Tricusped_Re._velocity > 2.4: NO (2.0)

| | | | | | | | | | | Aortic_Root_Diam > 30: NO (7.0)

| | | | | | | | | Age > 17: NO (3861.0/61.0)

| | | | | | | Posterior_Wall_of_LV > 11

| | | | | | | | Aortic_Root_Diam <= 32

| | | | | | | | Main_Pulmo._Artery_Diam. <= 1.5: NO (6.0)

| | | | | | | | Main_Pulmo._Artery_Diam. > 1.5

| | | | | | | | | Sex = 0

| | | | | | | | | | Tricusped_Re._velocity <= 1.5: YES (4.0)

| | | | | | | | | | Tricusped_Re._velocity > 1.5: NO (2.0)

| | | | | | | | | Sex = 1: YES (7.0)

| | | | | | | | Aortic_Root_Diam > 32: NO (8.0)

| | | | | Pericardial_Effusion = 2: YES (5.0/1.0)

| | | | | Pericardial_Effusion = 3: YES (10.0)

| | | | | Pericardial_Effusion = 4: YES (2.0)

| | | | | Pericardial_Effusion = 5

| | | | | | Main_Pulmo._Artery_Diam. <= 1.7: NO (2.0)

| | | | | | Main_Pulmo._Artery_Diam. > 1.7: YES (8.0)

| | | | | Pericardial_Effusion = 6: NO (3.0)

| | | | | Pericardial_Effusion = 7: NO (2.0)

| | | | Main_Pulmo._Artery_Diam. > 2

| | | | | Main_Pulmo._Artery_Diam. <= 2.4

| | | | | | Posterior_Wall_of_LV <= 7: NO (2.0)

| | | | | | Posterior_Wall_of_LV > 7

| | | | | | | Left_Atrium_Sys_Diam <= 33

| | | | | | | | Ratio_myocardial_&atrial <= 0.9: NO (7.0/2.0)

| | | | | | | | Ratio_myocardial_&atrial > 0.9: YES (12.0/2.0)

| | | | | | | Left_Atrium_Sys_Diam > 33

| | | | | | | | Left_Ventricle_Diam <= 53: YES (21.0)

| | | | | | | | Left_Ventricle_Diam > 53: NO (2.0)

| | | | | Main_Pulmo._Artery_Diam. > 2.4: YES (18.0)

| | | Tricusped_Re._velocity > 2.6

| | | | Tricusped_Re._velocity <= 3.1

| | | | | Main_Pulmo._Artery_Diam. <= 2

| | | | | | Pericardial_Effusion = 1

| | | | | | | Tricusped_Re._velocity <= 2.7

| | | | | | | | Age <= 68

| | | | | | | | | LV_Ejection_Fraction <= 61: NO (25.0/4.0)

| | | | | | | | | LV_Ejection_Fraction > 61: YES (2.0)

| | | | | | | | Age > 68: YES (4.0)

| | | | | | | Tricusped_Re._velocity > 2.7

| | | | | | | | LV_Ejection_Fraction <= 63

| | | | | | | | | Aortic_Root_Diam <= 29

| | | | | | | | | | Left_Ventricle_Systole <= 28

| | | | | | | | | | | Left_Ventricle_Diam <= 45

| | | | | | | | | | | | Posterior_Wall_of_LV <= 8

| | | | | | | | | | | | | Interventricular_Septum <= 8: NO (3.0)

| | | | | | | | | | | | | Interventricular_Septum > 8: YES (3.0/1.0)

| | | | | | | | | | | | Posterior_Wall_of_LV > 8: YES (11.0/1.0)

| | | | | | | | | | | Left_Ventricle_Diam > 45: NO (9.0)

| | | | | | | | | | Left_Ventricle_Systole > 28: YES (10.0)

| | | | | | | | | Aortic_Root_Diam > 29: YES (25.0)

| | | | | | | | LV_Ejection_Fraction > 63: NO (4.0)

| | | | | | Pericardial_Effusion = 2: YES (2.0)

| | | | | | Pericardial_Effusion = 3: YES (1.0)

| | | | | | Pericardial_Effusion = 4: YES (2.0)

| | | | | | Pericardial_Effusion = 5: YES (4.0)

| | | | | | Pericardial_Effusion = 6: YES (0.0)

| | | | | | Pericardial_Effusion = 7: YES (0.0)

| | | | | Main_Pulmo._Artery_Diam. > 2: YES (18.0)

| | | | Tricusped_Re._velocity > 3.1: YES (93.0/1.0)

| | Interventricular_Septum > 11

| | | Posterior_Wall_of_LV <= 12

| | | | Main_Pulmo._Artery_Diam. <= 2

| | | | | Sex = 0

| | | | | | LV_Ejection_Fraction <= 54: NO (3.0)

| | | | | | LV_Ejection_Fraction > 54

| | | | | | | Main_Pulmo._Artery_Diam. <= 1.7

| | | | | | | | Left_Ventricle_Diam <= 36: NO (3.0)

| | | | | | | | Left_Ventricle_Diam > 36

| | | | | | | | | Main_Pulmo._Artery_Diam. <= 1.5: YES (6.0)

| | | | | | | | | Main_Pulmo._Artery_Diam. > 1.5

| | | | | | | | | | Left_Atrium_Sys_Diam <= 34: NO (3.0)

| | | | | | | | | | Left_Atrium_Sys_Diam > 34: YES (10.0/1.0)

| | | | | | | Main_Pulmo._Artery_Diam. > 1.7: YES (27.0/1.0)

| | | | | Sex = 1

| | | | | | Main_Pulmo._Artery_Diam. <= 1.7

| | | | | | | Left_Atrium_Sys_Diam <= 39: NO (7.0)

| | | | | | | Left_Atrium_Sys_Diam > 39: YES (3.0/1.0)

| | | | | | Main_Pulmo._Artery_Diam. > 1.7

| | | | | | | Left_Ventricle_Diam <= 48

| | | | | | | | Age <= 66

| | | | | | | | | Left_Ventricle_Diam <= 46

| | | | | | | | | | Left_Atrium_Sys_Diam <= 34: NO (6.0/1.0)

| | | | | | | | | | Left_Atrium_Sys_Diam > 34: YES (4.0)

| | | | | | | | | Left_Ventricle_Diam > 46: NO (8.0)

| | | | | | | | Age > 66: YES (8.0)

| | | | | | | Left_Ventricle_Diam > 48: YES (10.0)

| | | | Main_Pulmo._Artery_Diam. > 2: YES (21.0)

| | | Posterior_Wall_of_LV > 12: YES (163.0/2.0)

Left_Atrium_Sys_Diam > 40

| LV_Ejection_Fraction <= 53: YES (673.0/1.0)

| LV_Ejection_Fraction > 53

| | Ratio_myocardial_&atrial <= 0.7

| | | Age <= 27

| | | | Posterior_Wall_of_LV <= 11: YES (21.0/1.0)

| | | | Posterior_Wall_of_LV > 11: NO (3.0)

| | | Age > 27: YES (427.0)

| | Ratio_myocardial_&atrial > 0.7

| | | Tricusped_Re._velocity <= 2.6

| | | | Posterior_Wall_of_LV <= 11

| | | | | Main_Pulmo._Artery_Diam. <= 2.2

| | | | | | Left_Ventricle_Systole <= 36

| | | | | | | Pericardial_Effusion = 1

| | | | | | | | Ratio_myocardial_&atrial <= 0.96

| | | | | | | | | Left_Ventricle_Diam <= 53

| | | | | | | | | | Posterior_Wall_of_LV <= 8: YES (9.0)

| | | | | | | | | | Posterior_Wall_of_LV > 8

| | | | | | | | | | | Age <= 54

| | | | | | | | | | | | Left_Atrium_Sys_Diam <= 43: NO (13.0/1.0)

| | | | | | | | | | | | Left_Atrium_Sys_Diam > 43: YES (3.0)

| | | | | | | | | | | Age > 54

| | | | | | | | | | | | Sex = 0

| | | | | | | | | | | | | Aortic_Root_Diam <= 27: NO (3.0)

| | | | | | | | | | | | | Aortic_Root_Diam > 27: YES (3.0)

| | | | | | | | | | | | Sex = 1

| | | | | | | | | | | | | Tricusped_Re._velocity <= 2.1: YES (29.0/1.0)

| | | | | | | | | | | | | Tricusped_Re._velocity > 2.1: NO (2.0)

| | | | | | | | | Left_Ventricle_Diam > 53

| | | | | | | | | | Sex = 0: YES (2.0)

| | | | | | | | | | Sex = 1: NO (8.0)

| | | | | | | | Ratio_myocardial_&atrial > 0.96

| | | | | | | | | Age <= 70

| | | | | | | | | | Left_Atrium_Sys_Diam <= 43: NO (28.0/2.0)

| | | | | | | | | | Left_Atrium_Sys_Diam > 43

| | | | | | | | | | | Left_Ventricle_Diam <= 42: NO (5.0)

| | | | | | | | | | | Left_Ventricle_Diam > 42: YES (11.0)

| | | | | | | | | Age > 70: YES (6.0)

| | | | | | | Pericardial_Effusion = 2: YES (1.0)

| | | | | | | Pericardial_Effusion = 3: YES (0.0)

| | | | | | | Pericardial_Effusion = 4: YES (0.0)

| | | | | | | Pericardial_Effusion = 5: YES (2.0)

| | | | | | | Pericardial_Effusion = 6: YES (0.0)

| | | | | | | Pericardial_Effusion = 7: YES (0.0)

| | | | | | Left_Ventricle_Systole > 36: YES (18.0)

| | | | | Main_Pulmo._Artery_Diam. > 2.2: YES (22.0)

| | | | Posterior_Wall_of_LV > 11: YES (81.0/1.0)

| | | Tricusped_Re._velocity > 2.6

| | | | Interventricular_Septum <= 7

| | | | | Left_Ventricle_Systole <= 27: NO (2.0)

| | | | | Left_Ventricle_Systole > 27: YES (7.0)

| | | | Interventricular_Septum > 7: YES (169.0/1.0)

Number of Leaves: 123

Size of the tree: 230