B.Sc. in Computer Science and Engineering Thesis

Intensive Care Unit Patient Monitoring of Pediatric and Congenital Heart Disease Using Data Mining

Submitted by

Md. Anisuzzaman
ID: 201114018

M. Niaz Mohammad Nazmul Huda Mithu
ID: 201114041

Md. Farhan Shahriar
ID: 201014030

Supervised by

Dr. Hasan Sarwar

Professor

Department of Computer Science and Engineering (CSE)

United International University (UIU), Dhaka

Bangladesh

Department of Computer Science and Engineering
Military Institute of Science and Technology

December 2014

CERTIFICATION

This thesis paper titled **"Intensive Care Unit Patient Monitoring of Pediatric and Congenital Heart Disease Using Data Mining"**, submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in December 2014.

Group Members:

Md. Anisuzzaman

M. Niaz Mohammad Nazmul Huda Mithu

Md. Farhan Shahriar

Supervisor:

Dr. Hasan Sarwar
Professor
Department of Computer Science and Engineering (CSE)
United International University (UIU), Dhaka
Bangladesh

CANDIDATES' DECLARATION

This is to certify that the work presented in this thesis paper, titled, "Intensive Care Unit Patient Monitoring of Pediatric and Congenital Heart Disease Using Data Mining", is the outcome of the investigation and research carried out by the following students under the supervision of Dr. Hasan Sarwar, Professor , Department of Computer Science and Engineering (CSE), United International University (UIU), Dhaka, Bangladesh.

It is also declared that neither this thesis paper nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Md. Anisuzzaman
ID: 201114018

M. Niaz Mohammad Nazmul Huda Mithu
ID: 201114041

Md. Farhan Shahriar
ID: 201014030

ACKNOWLEDGEMENT

We are thankful to Almighty Allah for his blessings for the successful completion of our thesis. Our heartiest gratitude, profound indebtedness and deep respect go to our supervisor, Dr. Hasan Sarwar, Professor , Department of Computer Science and Engineering (CSE), United International University (UIU), Dhaka, Bangladesh, for his constant supervision, affectionate guidance and great encouragement and motivation. His keen interest on the topic and valuable advices throughout the study was of great help in completing thesis.

We are especially grateful to the Department of Computer Science and Engineering (CSE) of Military Institute of Science and Technology (MIST) for providing their all out support during the thesis work.

Finally, we would like to thank our families and our course mates for their appreciable assistance, patience and suggestions during the course of our thesis.

Dhaka
December 2014

Md. Anisuzzaman

M. Niaz Mohammad Nazmul Huda Mithu

Md. Farhan Shahriar

ABSTRACT

Now a day in our world, heart diseases are the number one major cause of death. Statistics says that about 80% of deaths occurred in low- and middle income countries. If current trends are allowed to continue, by 2030 an estimated 23.6 million people will die from cardiovascular disease (mainly from heart attacks and strokes). The health care industry gathers enormous amounts of heart disease data which, unfortunately, are not ”mined” to discover hidden information for effective decision making. The reduction of blood and oxygen supply to the heart leads to heart disease.However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. This research paper intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques which will be useful for medical practitioners to take effective decision. The objective of this research work is to predict more accurately the presence of heart disease with reduced number of attributes.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

CPU : Central Precessing Unit
CVD : Cardiovascular Disease
ETL : Extract Transform Load
GUI : Graphical User Interface
KDD : Knowledge Discovery in Database

LIST OF SYMBOLS

HCO_2	: Hydrogen bi Carbonate
Hb	: Hemoglobin
PO_2	: Partial Pressure of Oxygen
PCO_2	: Partial Pressure of Carbon di oxide
HCO_2	: Carbonic Acid
Ca	: Calcium
Na	: Sodium
Cl	: Chlorine
$PEEP$	: Positive End-Expiratory Pressure
$Lact$	: Lactic acid
$AaDO_2$	: Alveolo-Arterial Oxygen Difference

CHAPTER 1
INTRODUCTION

According to the World Health Organization heart disease is the first leading cause of death in high and low income countries and occurs almost equally in men and women . By the year 2030, about 76% of the deaths in the world will be due to non-communicable diseases (NCDs). Cardiovascular diseases (CVDs), also on the rise, comprise a major portion of non communicable diseases. In 2010, of all projected worldwide deaths, 23 million are expected to be because of cardiovascular diseases. In fact, CVDs would be the single largest cause of death in the world accounting for more than a third of all deaths. For CVDs specifically, in 2005, the age standardized mortality rate for developing nations like India, China, and Brazil was between 300-450 per 100,000, whereas it was around 100-200 per 100,000 for developed countries like USA and Japan4. According to a recent study by the Registrar General of India (RGI) and the Indian Council of Medical Research (ICMR), about 25 percent of deaths in the age group of 25- 69 years occur because of heart diseases. The core functionalities of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data . From the last two decades data mining and knowledge discovery applications have got a rich focus due to its significance in decision making and it has become an essential component in various organizations. The field of data mining have been prospered and posed into new areas of human life with various integrations and advancements in the fields of Statistics, Databases, Machine Learning, Pattern Reorganization and health care. Medical Data mining in health care is regarded as an important yet complicated task that needs to be executed accurately and efficiently. Health care data mining attempts to solve real world health problems in diagnosis and treatment of diseases. This research paper aims to analyze the several data mining techniques proposed in recent years for the diagnosis of heart disease. Many researchers used data mining techniques in the diagnosis of diseases such as tuberculosis, diabetes, cancer and heart disease in which several data mining techniques are used in the diagnosis of heart disease such as KNN, Neural Networks, Bayesian classification, Classification based on clustering, Decision Tree, Genetic Algorithm, Naive Bayes, Decision tree, WAC which are showing accuracy at different levels. Each data mining technique serves a different purpose depending on the modeling objective. Nave Bayes is one of the successful data mining techniques used in the diagnosis of heart disease patients . Naive Bayes classifiers have works well in many complex real-world situations. Naive Bayes or Bayes Rule is the basis for many machine-learning

and data mining methods. The rule is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. It learns from the evidence by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables. By theory, this classifier has minimum error rate but it may not be case always. However, inaccuracies are caused by assumptions due to class conditional independence and the lack of available probability data. Observations show that Nave Bayes performs consistently before and after reduction of number of attributes. Bagging plays an important role in the field of medical diagnosis. Many research works in this aspect is depicted in related work. Bagging algorithms used to improve model stability and accuracy. Bagging works well for unstable base models and can reduce variance in predictions. In our country there are many patients. A major portion of these patients are suffering from heart disease. There are some patient who suffer in heart disease from birth. These types of heart diseases are treated in the department of pediatric and congenital heart disease.

1.1 Background

The use of data mining tools has become widely used in clinical applications for disease diagnosis more effectively. Various data mining techniques such as decision trees, artificial neural networks, Bayesian networks, support vector machines kernel density, bagging algorithm have been actively used in clinical support systems for diagnosis of heart disease. Although there have been promising results in applying data mining techniques in heart disease diagnosis and treatment, the study done in finding out treatment options for patients and particularly heart patients is comparatively elemental. It has been suggested by researchers that application of data mining techniques for proposing suitable treatments options for patients would not only improve patient care but would also reduce investigation time, errors and would also improve the performance of medical practitioners . There has been a lot of investigation for applying different data mining techniques in the diagnosis of heart disease to find out the most accurate technique but there is no study to find out the data mining technique which can increase reliability and accuracy in finding out effective treatment for heart disease patients.

1.2 Objective

Our main target is to develop an system which will be useful for doctors and nurses. This system will remove the need of papers in the intensive care unit. This will reduce the work of the nurses. The nurses will just give the input. we use data mining on the input data in order to create decision based on the condition of the patient. We will give the appropriate

result for the current condition of the patient. This work will predict the patients condition whether it will be good or bad condition. We are using Data mining in order to get the prediction.

1.3 Outline of Work

The following section will cover the elaborate description of the work.Chapter 2 will cover the related works in this field. Chapter 3 will cover the methodology. Methodology means how the process works. Which data is taken as input? And which is the output? We are using data mining process to give output.For data mining we use Rapid Miner. So there is the detail description about the data mining and RapidMiner. Chapter 4 will cover the implementation steps, input data format and data fitting with the software.In this Chapter we have discussed about the whole process. How we work with the data? And how we develop the System. Chapter 5 will cover the result and discussion. In this chapter we show the decision tree we have created in our process. The last chapter, Chapter 6 will be for Conclusion of our work.

CHAPTER 2
RELATED WORK

Carlos Ordonez, introduced an algorithm [1] to reduce the number of rules which used search limitation. The introduced algorithm searches for association rules in a training set and finally validates them on an independent test set. In medical terms, to the degree of disease in four specific arteries, the association rules related heart perfusion measurements and risk factors. Association rules were applied on a real data set containing medical records of patients with heart disease. Search limitations and test set validation importantly reduced the number of association rules and produced a set of rules with high predictive accuracy. Unfortunately, when association rules were applied on a medical data set, they produced an extremely large number of rules. They used the train and test approach which used two disjoint samples from a data set to search and validate rules. To filter rules on the test set, support, confidence and lift have different importance. They opinioned that to validate rules confidence was the most important metric. Based on heart perfusion measurements and risk factors they used association rules to predict the degree of narrowing in four arteries. They presented medically significant rules discovered on medical data set that remain valid in several independent train/test cycles. The two problems were addressed such as large numbers of rules were obtained by the standard association rule algorithm and the validation of rules on an independent set, which was required to eliminate unreliable rules. K.C.
Tan and E.J. Teoh et al, have proposed a hybrid approach consist of two conventional machine learning algorithms. Genetic Algorithms (GAs) and Support Vector Machines (SVMs) were the two proposed algorithm combined effectively based on a wrapper approach. Here, by an evolutionary process genetic algorithm component searches for the best attribute data set. Based on the attribute subset represented by GA, the SVM classified the patterns into reduced data set. This cyclic method was known as wrapper approach. UCI machine learning repository provided 5 set of data and it was checked by the proposed GA and SVM hybrid approach. After that the data was combined with some of the established classifier in the data mining community and showed that the collected result of hybrid approach provided a high average classification. Also the consistency of the GA-SVM hybrid was clearly seen from the histogram analysis and box plots. The hybrid approach included the utilization of a correlation measure to improve the average fitness of a chromosome population and the substitution of weaker chromosomes based on the correlation measure improved the ability of hybrid classification. The analysis demonstrated GA-SVM hybrid as a good classifier when

the irrelevant attributes were removed. The GA-SVM hybrid approach attained an average accuracy of 76.20% which was relatively high. The robustness of the GA-SVM hybrid in the multi-class domain was showed by the obtained average accuracy 84.07%.

Jesmin Nahar and Tasadduq Imam et al, have proposed a computer intelligent based approach for the diagnosis of heart diseases. Apriori, Predictive Apriori and Tertius were the three different rule mining algorithms used to present rule extraction experiment on heart disease data and showed as efficiency algorithm for diagnosis task. Cleveland dataset, a publicly available dataset and widely popular with data mining researchers, have been used for diagnosis because of the privacy problem related to medical data set. Generally diagnosis were costly, time consuming and likely to suffer from error. The analyzed information available on sick and healthy individuals indicted that females have less chance of coronary heart disease than males. Heart disease for both men and women was existed only in the presence of exercise-induced angina and factors such as chest pain were asymptotic. The resting ECG for men and women was different. The risk factors of resting ECG for women were being either normal or hyper and slope being flat. And only a single rule expressing resting ECG being hyper was an important factor for men. Slope being up, number of colored vessels being zero and old peak being less than or equal to 0.56 indicated after comparing the healthy status of men and women.

Kemal Polat and Salih Gunes, [2] have proposed a feature selection method called Kernel F-score Feature Selection (KFFS) which is used as pre-processing step in the classification of medical datasets. The proposed KFFS method has two phases. In first phase by means of Linear (Lin) or Radial Basis Function (RBF) kernel functions, the features of medical datasets have been transformed to kernel space. Using F-score formula, the F-score values of medical datasets with high dimensional feature space have been calculated. The cause of using kernel functions transformed from non-linearly separable medical dataset to a linearly separable feature space. To test the performance of KFFS method the UCI (University California, Irvine) machine learning Gayathri. P et.al / International Journal of Engineering and Technology (IJET) ISSN : 0975-4024 Vol 5 No 3 Jun-Jul 2013 2948 database used were heart disease dataset, SPECT (Single Photon Emission Computed Tomography) images dataset and Escherichia coli Promoter Gene Sequence dataset. The area under ROC curve values (AUC) values obtained from just Least Square Support Vector Machine (LS-SVM) and Artificial Neural Network (LANN) classifiers without KFFS method on the classification of heart disease, SPECT images dataset and E. coli Promoter Gene Sequence dataset were 0.7960.708, 0.5570.596 and 0.6560.679. The AUC values obtained from LS-SVM and ANN classifiers with KFFS (RBF kernel) on the classification of heart disease, SPECT images dataset, and E. coli Promoter Gene Sequence dataset were 0.8310.765, 0.750.634 and 0.6470.730. They compared and found the best expert system based on the classification used in medical data set.

Pasi Luukka and Jouni Lampinen, have applied classification method based on preprocess-

ing the data first with Principal Component Analysis (PCA) and then applying differential evolution classifier to the diagnosis of heart disease. This method was applied here for predictind diagnosis from clinical data sets with chief complaint of chest pain using classical Electronic Medical Record (EMR), heart data sets which contains demographic properties, clinical symptoms, clinical findings, laboratory test results specific Electro- CardioGraphy (ECG), results pertaining to angina and coronary infarction. Individually they computed results for four different heart data sets and also the results for the case when all data sets were combined together. It was to demonstrate and assess the proposed classification approach. They were considered that the main factor resulting in the good classification accuracy in studied cases was the application of an effective global optimizer, differential evolution, for fitting the classification model instead of local optimization based approaches. With diagnosis of heart disease they found the data preprocessing with PCA. Here higher classification accuracy can be achieved than without preprocessing. The result indicated that prepossessing the data before classification might not only help with the curse of increasing data dimensionality, but also provide a further improvement in classification accuracy. They managed to classify the Switzerland data set with 94.5%0.4% mean accuracy. They were combined the data sets and achieved the mean accuracy of 82%0.5%. Here, the classification accuracy yielded by the proposed approach was outperformed when compared with the other corresponding results of several classifiers.

Resul Das and Ibrahim Turkoglu, et al, [3] have proposed several tools and various methodologies to develop effective medical decision support system. Diagnosing of the heart disease was one of the important issue and many researchers investigated to develop intelligent medical decision support systems to improve the ability of the physicians. A method was introduced which uses Statically Analysis System (SAS) base software 9.1.3 for diagnosing of the heart disease. In this method neural networks ensemble model was used which enabled an increase in generalization performance by combining several individual neural networks train on the same task. SAS base software 9.1.3 supported all tasks in a within a single, integrated solution while providing the flexibility for efficient collaborations. For heart disease diagnosis the experimental result obtained 89.01% classification accuracy, 80.95% sensitivity and 95.91% specificity values for heart disease diagnosis.

Hongmei Yan and Jun Zheng, et al, [4] have proposed a real-coded GA based system to select the critical clinical features essential to the heart diseases diagnosis. It has been proposed to select the critical features and assist the diagnosis of five major heart diseases which were hypertension, coronary heart disease, rheumatic valvular heart disease, chronic pulmonale and congenital heart disease. They used heart disease data with 352 cases and for each case 40 diagnostic features were recorded. Among the 352 cases of heart disease data sets 24 critical diagnostic features have been identified and their corresponding diagnosis weights for supporting or denying the diagnosis of each heart disease have been determined. It provided high accuracy in heart disease diagnosis.

Akin Ozcift and Arif Gulten, have constructed a Rotation Forest (RF) ensemble classifier. It was constructed to improve the accuracy of machine learning algorithm. It was absolutely necessary in designing of high performance computer aided diagnosis system. Here rotation forest (RF) ensemble classifiers of 30 machine learning algorithms to evaluate their classification performances using Parkinson's, diabetes and heart diseases data sets. Using correlation based feature selection algorithm three data sets were reduced and then performances of 30 machine learning algorithms were calculated for three data sets and constructed based on RF algorithm. The performance of respective classifier was accessed with the same disease data and 60 algorithms were evaluated using three metrics. Average accuracy for diabetes, heart and Parkinson's data sets were 72.15%, 77.52% and 84.43%. And for the proposed RF ensemble classifiers have accuracy of 74.47%, 80.49% and 87.13%. The accuracy of RF was improved and used for designing advance systems.

Chih-Lin Chi and W. Nick Street, et al, have proposed Optimal Decision Path Finder (ODPF) which was machine learning based expert system. It was used because of informative in terms of diagnostic accuracy in case of minimizing the time and money spent on diagnostic testing. Two tests were considered here. In first the immediate result was obtained in blood pressure test and the second was more costly and the test result was delayed. The proposed method ODPF focused on second test because of the time delayed result in test. It takes pre-test probability, interaction of variables and the cost of each test into account to generate an individualized test sequence to avoid delay test result in costly test. Lazy-learning classifiers, confident Gayathri. P et.al / International Journal of Engineering and Technology (IJET) ISSN : 0975-4024 Vol 5 No 3 Jun-Jul 2013 2949 diagnosis and Locally Sequential Feature Selection (LSFS) were the methods used to identify the sequence of diagnostic test. Among this LSFS provide cost saving accuracy of heart disease, thyroid disease, diabetes and hepatitis datasets and test saving accuracy by combining four different data sets. After comparing the results test saving provide more information and accuracy on patients available information than cost saving.

Yoon-Joo Park and Se-Hak Chun, et al, [5] have suggested Cost-Sensitive Case-Based Reasoning (CSCBR), a new knowledge extraction technique. It included unequal misclassification cost into conventional case based reasoning. To classify the absence and presence of disease genetic algorithm was used. An effort was taken to minimize misclassification error costs into CBR by the best classification of boundary point and number of neighbor. A fixed number of nearest neighbors in CBR was overcome by CSCBR. The absence and presence of disease was classified by adjusting the optimal cut-off classification point and cut-off distance point for selecting best neighbors. The CSCBR technique was applied in five medical data sets and then compared the result with C5.0 and CART. The total misclassification cost of CSCBR was lower than other cost-sensitive methods and was originally designed to classify binary case.

Jesmin Nahar and Tasadduq Imam, et al, [6] have examined the fact of computational in-

telligent techniques in heart disease diagnosis. Because early detection of heart disease was essential to save lives. Cleveland data was used to perform comparison with six well known classifiers. The potential of medical knowledge-driven feature selection was showed by comparing with computational intelligent based technique. And the imbalance data issue created by publicly available cleaved data was identified. For most classifiers and majority data set the performance was improved by the use of Motivated Feature Selection (MFS). It was because of the conversion of Cleveland data set for binary classification. The experimental results demonstrated that the use of MFS noticeably improved the performance, especially in terms of accuracy, for most of the classifiers considered and for majority of the datasets. MFS with Computer Feature Selection (CFS) was a promising technique used in heart disease diagnosis.

Laercio Brito Gonalves and Marley Maria Bernardes Rebuzzi Vellasco, et al, have determined that the Inverted Hierarchical Neuro-Fuzzy Binary Space Partitioning (HNFB-1) was based on the Hierarchical Neuro-Fuzzy Binary Space Partitioning Model (HNFB) which gave an idea that recursive partitioning of the input space. It was able to generate its own structure automatically and allowed a greater number of inputs. The classification task of HNFB-1 has been evaluated with different benchmark databases such as heart disease data sets. They introduced an Inverted Hierarchical Neuro-Fuzzy BSP System. It was a neuro-fuzzy model which has been specifically created for record classification and rule extraction in databases. It allowed the extraction of knowledge in the form of interpretable fuzzy rules. Fuzzy accuracy and Fuzzy coverage were the two fuzzy evaluation measures defined for the process of rule extraction in the HNFB-1 model. The HNFB-1 model had showed better classification performance when compared with several other pattern classification models and algorithms and the processing time converged by HNFB-1 was very less.

Kemal Polat and Salih Gunes, [7] have presented a hybrid approach based on feature selection, fuzzy weighted preprocessing and Artificial Immune Recognition System (AIRS) to medical decision support systems. The hybrid approaches based on feature selection have two stages. The dimensions of heart disease and hepatitis disease datasets were reduced to 9 from 13 and 19 in the feature selection (FS) sub-program by means of C4. 5 decision tree algorithm. The second stage was heart disease and hepatitis disease datasets were normalized in the range of [0, 1] and were weighted via fuzzy weighted pre-processing. The obtained classification accuracies of system were 92.59% and 81.82% using 5050% training-test split for heart disease and hepatitis disease datasets. AIRS have showed an effective performance on several problems such as machine learning benchmark problems and medical classification problems like breast cancer, diabetes and liver disorders classification. They have used the heart disease and hepatitis disease datasets taken from UCI machine learning database as medical dataset.

Kemal Polat and Seral S ahan et al, [8] have applied k-nearest neighbor (k-nn) weighting preprocessing and fuzzy resource allocation mechanism with AIRS on the task of diagnosis

of heart disease. Here, diagnosis of heart disease was conducted with a machine learning system. In this system, a new weighting scheme based on k-nearest neighbour (k-nn) method was utilized as a preprocessing step before the main classifier. It was evident that the usages of machine learning methods in disease diagnosis have been increased gradually. While conducting this study, they first applied the k-nn based weighting process to the dataset and weighted it in the interval [0, 1]. The results strongly suggested that k-nn weighted preprocessing and fuzzy resource allocation mechanism with AIRS can aid in the diagnosis of cardiac arrhythmias. 87% of classification accuracy was obtained by their system.

Humar Kahramanli and Novruz Allahverdi, have developed a hybrid neural network which included Artificial Neural Network (ANN) and Fuzzy Neural Network (FNN). The proposed method accuracy, sensitivity and specificity measures were evaluated which were used commonly in medical classification. The aim of classification was to increase the reliability of the results obtained from the data. Here a new method was presented for classification of data of a medical database. The proposed algorithm achieved the highest accuracy Gayathri. P et.al / International Journal of Engineering and Technology (IJET) ISSN : 0975-4024 Vol 5 No 3 Jun-Jul 2013 2950 rate when comparing the records in the UCI web site and related previous studies for diabetes dataset. The proposed method achieved accuracy values of 84.24% and 86.8% for Pima Indians diabetes dataset and Cleveland heart disease dataset respectively. The classification accuracies obtained by the proposed hybrid neural network were one of the best results compared with the results reported in the literature.

P.K. Anooj, [9] presented a weighted fuzzy rule-based Clinical Decision Support System (CDSS) for computer-aided diagnosis of the heart disease. The proposed CDSS for risk prediction of the heart patients contains two steps such as: generation of weighted fuzzy rules and developing of a fuzzy rule-based decision support system. Here, data preprocessing was applied on the heart disease data set for removing the noisy information and to find missing values. After that using the frequent attribute categories, the deviation range and relevant attributes were computed in this method. According to the deviation range, the attributes were selected if any deviation exists or not and also the deviation range was used to construct the decision rules. Those decision rules were scanned in the learning database to find out its frequency. As per its frequency the weight age was calculated for every decision rule obtained and by the help of fuzzy membership function, the weighted fuzzy rules were obtained. The automatic procedure to generate the fuzzy rules was an advantage of the proposed system and the weighted procedure introduced in the proposed work was additional advantage for effective learning of the fuzzy system. These weighted fuzzy rules were used to build the CDSS using Mamdani fuzzy inference system.

Nazri Mohd Nawi and Rozaida Ghazali et al, [10] have proposed a novel method to improve the efficiency of back propagation neural network algorithms. In Gradient Descent with Momentum and Adaptive Gain (GDM/AG) proposed algorithm, for each node the gain value was changed adaptively to modify the initial search direction. The modification enhanced

the computational efficiency of training process and can be implemented in optimization process. The convergence speed of the proposed algorithm was evaluated using classification matrix. The heart disease of the patient was predicted efficiently. The algorithms were strongly constructed and have the ability to enhance the computational efficiency.

Evanthia E. Tripoliti et al, have proposed a dynamic determination of the number of trees in random forests algorithm, a computerized diagnosis of diseases based on sorting. They have addressed the dynamic purpose of the optimum number of fundamental classifiers making up the random forests. Their proposed technique is different from most of the techniques presented in the literature. They dogged the number of classifiers during the growing procedure of the forest. Their proposed technique produces an ensemble not only accurate but also diverse ensures the two essential properties that distinguish an ensemble classifier. Their technique is derived from on line fitting procedure and it is calculated using eight biomedical datasets and five versions of random forest algorithm.

Zhihua Cui et al, have proposed a training artificial neural network by exploiting Artificial Photosynthesis and Photo tropism Mechanism (APPM). They used a stochastic optimization algorithm that stirs the plant growing process. In their algorithm each entity is called as branch and the sampled points are contemplated as branch growing trajectory. They have applied the APPM algorithm to instruct the connection weights for artificial neural network. They have used two real world issues which are Cleveland heart disease categorization issue and sunspot number foreseeing issue to evaluate the performance of their APPM trained ANN. Their outcome showed that their technique increased the performance significantly contrast to other sophisticated machine learning techniques.

P.K. Anooj, [11] has proposed a weighted fuzzy rule-based CDSS for the diagnosis of heart disease. It automatically obtains the knowledge from the patients clinical data. The proposed CDSS for risk prediction of heart patients consists of two phases such as automated approach for generation of weighted fuzzy rules and decision tree rules and the second is, developing a fuzzy rule-based decision support system. An example of a medical domain application was a detection system for heart disease based on computer-aided diagnosis methods, where the data was obtained from some other sources and was evaluated by computer based applications. Up to now, computers have usually been used to build knowledge based clinical decision support systems which used the knowledge from medical experts and transferring this knowledge into computer algorithms was done manually. The performance of the proposed CDSS improved the risk prediction when compared with the neural network-based clinical support system.

CHAPTER 3
METHODOLOGY

3.1 Data Mining

Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data. Data mining also known as knowledge discovery in data bases (KDD) is the process of automatically discovering useful information in large data repositories. A formal definition of Knowledge discovery in databases is given as follows: ”Data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data”. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data. Data mining uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used. For example K- means clustering is unsupervised.

3.2 RapidMiner

RapidMiner is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process including results visualization, validation and optimization. RapidMiner is developed on a business source model.

RapidMiner uses a client/server model with the server offered as Software as a Service or on cloud infrastructures. According to Bloor Research, RapidMiner provides 99% of an advanced analytical solution through template-based frameworks that speed delivery and reduce errors by nearly eliminating the need to write code. RapidMiner provides data mining and machine learning procedures including: data loading and transformation (Extract, transform, load (ETL)), data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment. RapidMiner is written in the Java programming language. RapidMiner provides a GUI to design and execute analytical work flows. Those work flows are called ”Process” in RapidMiner and they consist of multiple ”operators”.

Each operator is performing a single task within the process and the output of each operator forms the input of the next one. Alternatively, the engine can be called from other programs or used as an API. Individual functions can be called from the command line. RapidMiner provides learning schemes and models and algorithms from WEKA and R scripts that can be used through extensions. RapidMiner provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process including results visualization, validation and optimization. RapidMiner is developed on a business source model. EachRapidMiner uses a modular concept for this, where each step of an analysis (e.g. a preprocessing step or a learning procedure) is illustrated by an operator in the analysis process. These operators have input and output ports via which they can communicate with the other operators in order to receive input data or pass the changed data and generated models over to the operators that follow. Thus a data flow is created through the entire analysis process, as can be seen by way of example in Figure 3.1.

Alongside data tables and models, there are numerous application-specific objects which can flow through the process.In the text analysis, whole documents are passed on, time series can be led through special transformation operators or preprocessing models are simply passed on to a storage operator (like a normalization) in order to reproduce the same transformation on other data later on. The most complex analysis situations and needs can be handled

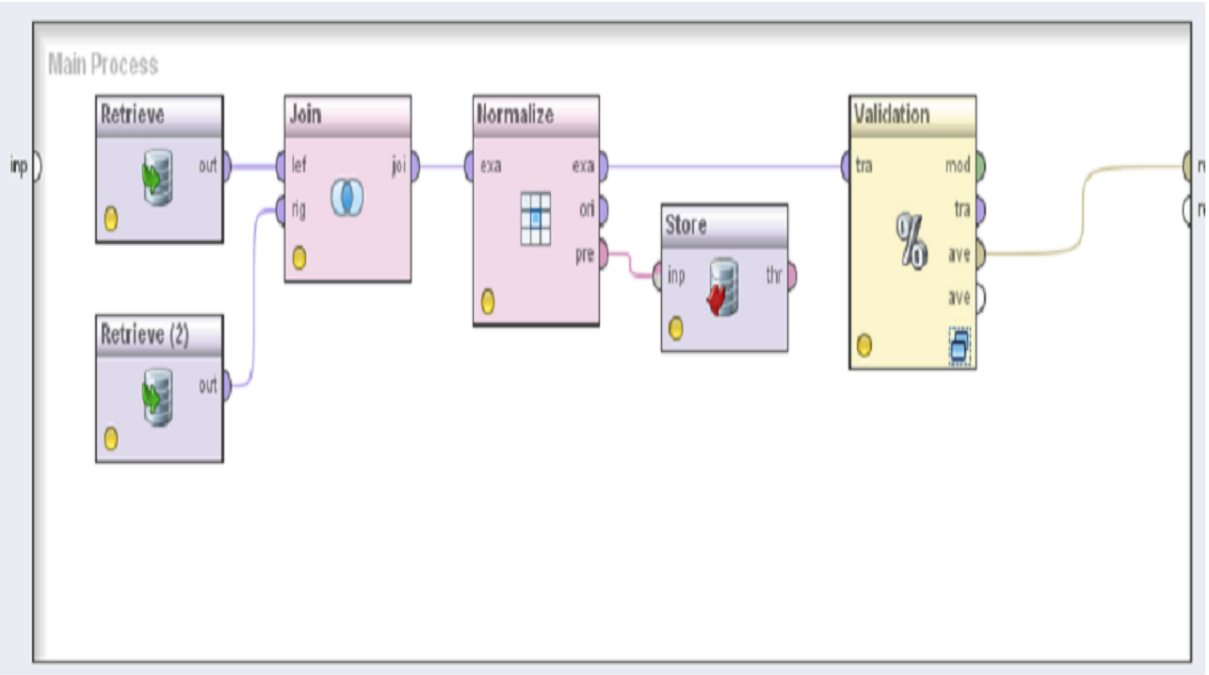


Figure 3.1: A simple process with examples of operators for loading, preprocessing and model production

by so-called super-operators, which in turn can contain a complete sub process. A well-known example is the cross-validation, which contains two sub processes. A sub process is responsible for producing a model from the respective training data while the second sub process is given this model and any other generated results in order to apply these to the test data and measure the quality of the model in each case.A typical application can be seen in Figure 3.2
, where a decision tree is generated on the training data, an operator applies the model in the test sub process and a further operator determines the quality based on the forecast and the true class.

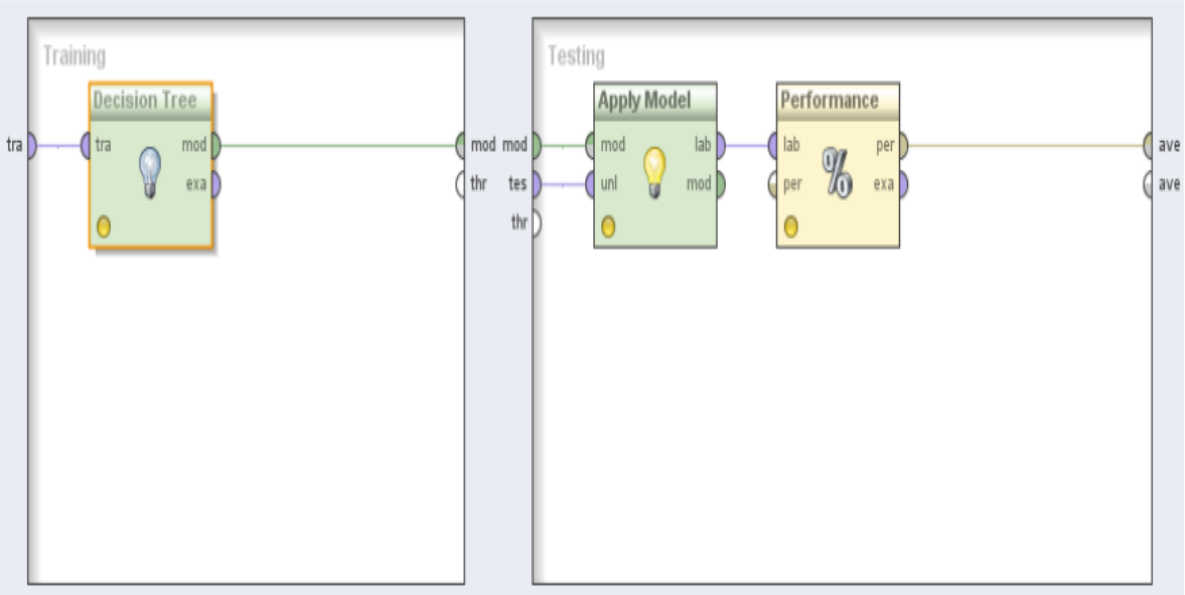


Figure 3.2: The internal sub processes of a cross-validation

Repositories enable the user to save analysis processes, data and results in a project-specific manner and at the same time always have them in view which is shown in Figure 3.3.

Thus a process that has already been created can be quickly reused for a similar problem, a model generated once can be loaded and applied or the obtained analysis results can simply be glanced at so as to find the method that promises the most success. The results can be dragged and dropped onto processes, where they are loaded again by special operators and provided to the process. In addition to the local repositories, which are stored in the file system of the computer, RapidAnalytics instances can also be used asa repository. Since the RapidAnalytics server has an extensive user rights management, processes and results can be shared or access for persons or groups of persons limited. The repositories provided by RapidAnalytics make a further function available, which makes executing analyses much

easier. The user can not only save the processes there, but also have them executed by the RapidAnalytics instance with the usual comfort. This means the analysis is completely implemented in the background and the user can find out about the analysis process via a status display. The user can continue working at the same time in the foreground, without his computer being slowed down by CPU and memory-intensive computations. All computations now take place on the server in the background, which is possibly much more efficient, as can be seen in Figure 3.4. This also means the hardware resources can be used more efficiently, since only a potent server used jointly by all analysts is needed to perform memory-intensive computations.

Alongside the core components of RapidMiner, there are numerous extensions which upgrade further functions, such as the processing of texts, time series or a connection to statistics package R or Weka . All these extensions make use of the extensive possibilities offered by RapidMiner and supplement these: They do not just supplement operators and new data objects, but also provide new views that can be freely integrated into the user interface, or even supplement entire perspectives in which they can bundle their views like the R extension in Figure 3.5.

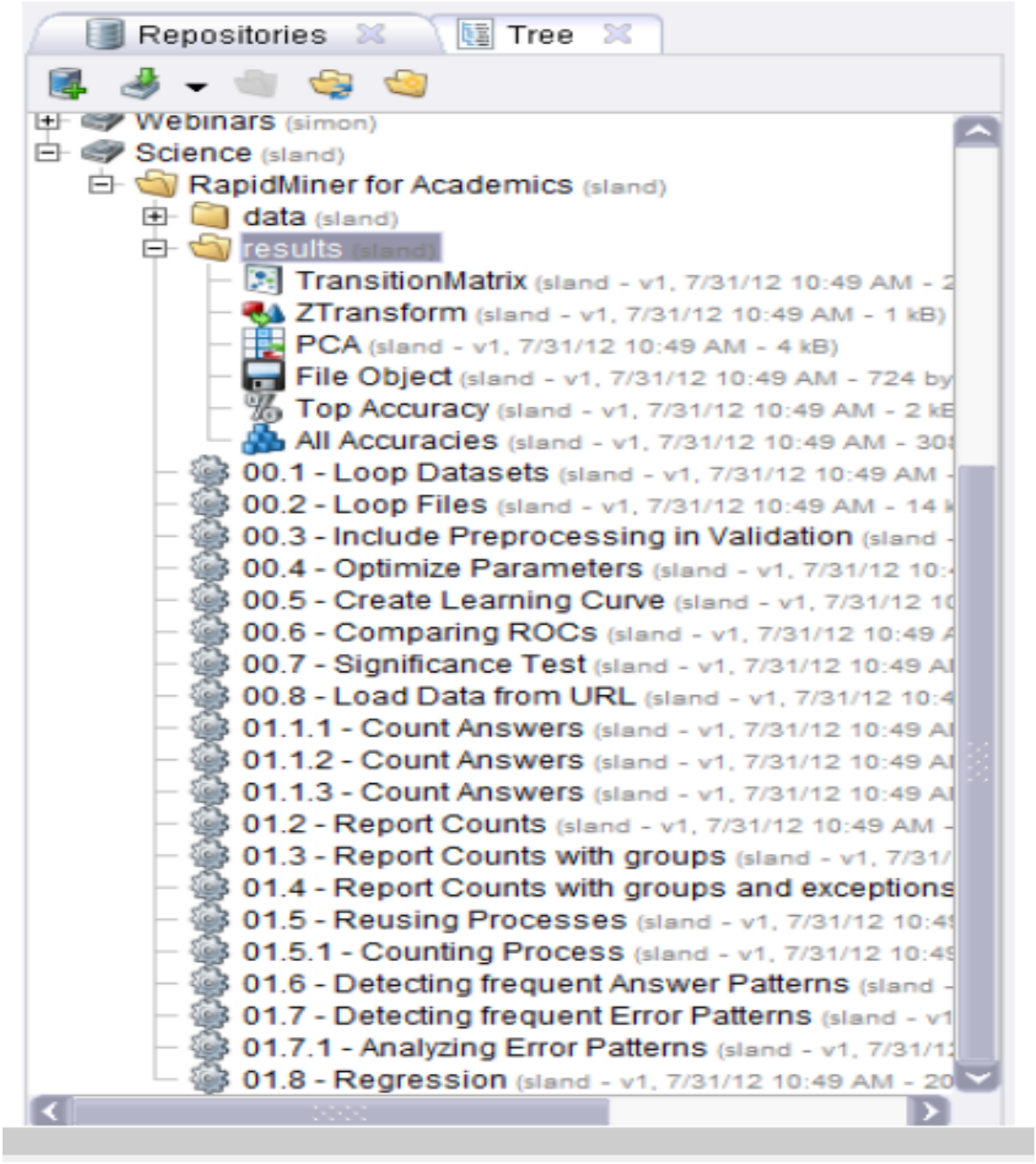


Figure 3.3: A well-filled repository

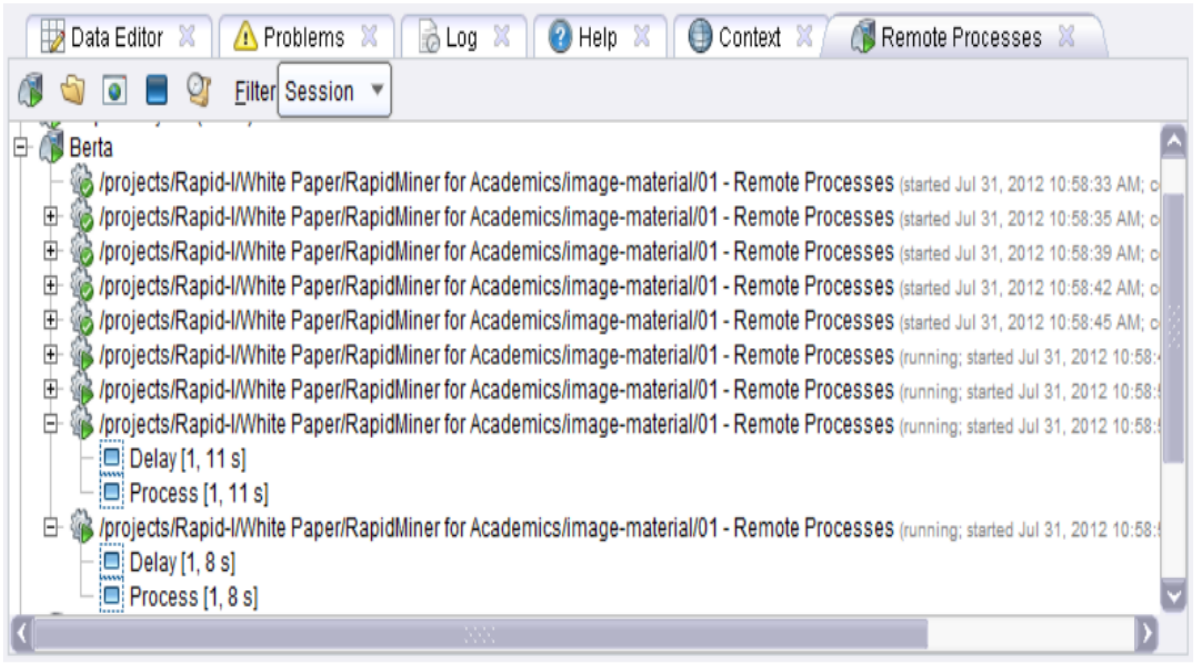


Figure 3.4: A process which was started several times on a RapidAnalytics instance. Some cycles are already complete and have produced results.

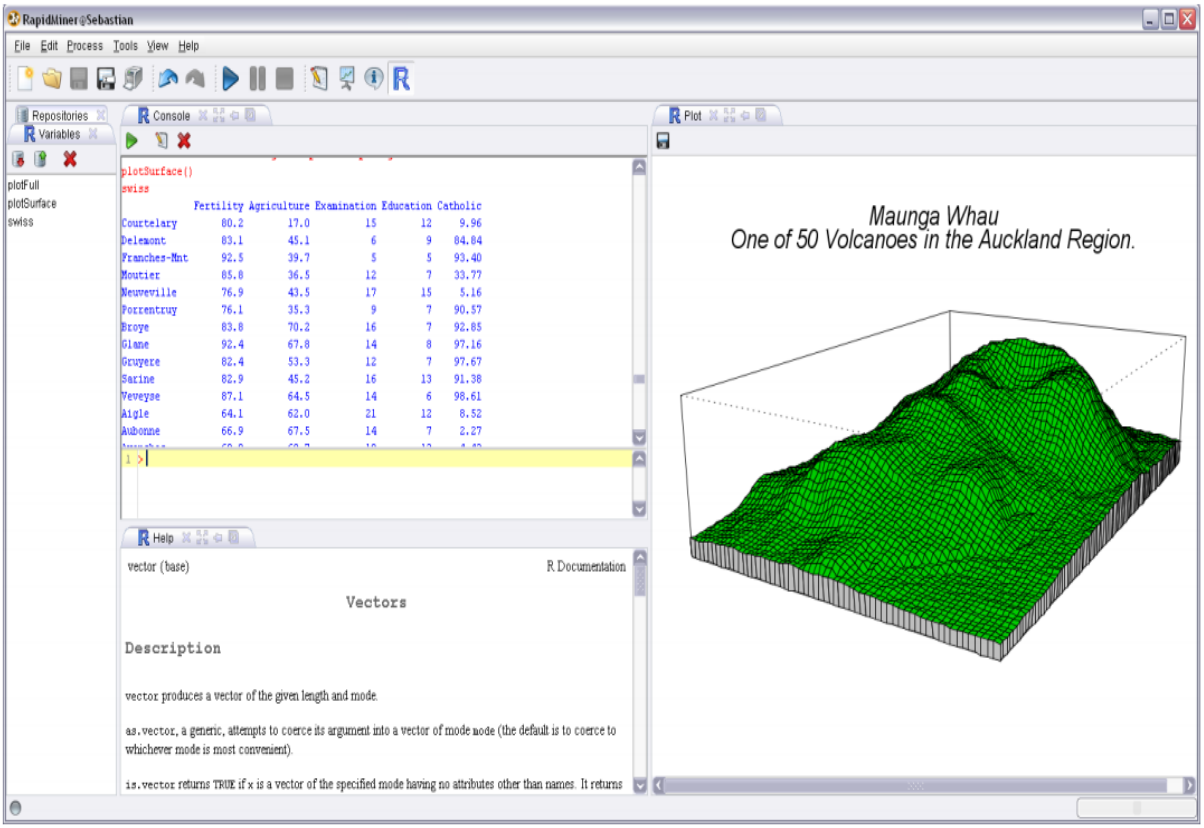


Figure 3.5: The R perspective in RapidMiner

CHAPTER 4
IMPLEMENTATION

4.1 Step of Implementation

The very first step of data mining process is the collection of the data. The second step is to develop an appropriate system through data mining. This data mining is performed by analyzing the condition of very large number of patient's data. The third step is to give the appropriate prediction through the result of data mining. The data are collected from the patient. The nurses noted down the data in a paper. Now they will use an android mobile application. In the android application there are all input parameter. The input data are stored in the server. We are using rapid miner for data mining. Rapid miner can import the data stored in the excel sheet. This excel sheet is taken as input. By applying various operator of rapid miner like cross validation, apply model, decision tree, performance etc., the data mining is performed. This mining result is then used as a system to predict the output parameter. For certain input parameter the output parameter changes its value. Now the system is developed. When any input will be given to the system then the appropriate value will be predicted.

4.2 Input Data Format

The input data are different component of blood. the component are cited in the tabular form.

The input data are directly taken from the patient. Every hour the condition of the patient are reported. due to different value of the input parameter the value of the out parameter change its value. The value of $FiCO_2$ changes when the value of PO_2 changes [12]. We increase $FiCO_2$ when PO_2 decreases. We decrease $FiCO_2$ when PO_2 increases. The value of PEEP changes when the value of PO_2 changes. We increase PEEP when PO_2 decreases. We decrease PEEP when PO_2 increases. When PCO_2 [13] increases we increase Vent Rate and TV. We decrease Vent Rate and TV when PCO_2 decreases. This changes are made from the previous experience. TV and PIP are proportional. When TV increases PIP is to be increased automatically. MAP is calculated from TV , PIP and PEEP. And PO_2 / $FiCO_2$ is the ratio of PO_2 and $FiCO_2$. Sometimes it needed to decide if the patient require certain

Table 4.1: Input Parameter

Sl	Parameter	Value
19	Ph	7.3
20	PCO_2	39
21	PO_2	54
22	HCO_2	21
23	BE	3.55
24	Hb%	17
25	Na	135
26	K	4.33
27	Cl	0.9
28	AnionGap	5
29	$AaDO_2$	5
30	Lact	0.77
31	Ca	0.8
32	Glu	102

Table 4.2: Output Parameter

Sl	10	11	12	13	14	15	16	17	18
parameter	Mode of vent	Vent Rate	TV	PIP	PEEP	MAP	Ti	$FiCO_2$	PO_2 / $FiCO_2$

medicine or not. All these are done by the close observation on the input data. Now we have to make the system which will be intelligent enough to gather the knowledge of the previous result and give exact output value. The output parameters corresponding to the given input parameters are given in Table 4.2

All the output values are related with the input data taken from the patient. Based on these output parameter we decide the condition of the patient. We can not take the decision only observing the values of output parameters. Since the values are different for patients. For example the blood pressure of baby is much less than an adult person. The blood pressure increases with the increase of age. Blood pressure is measured by measuring some parameter like PCO_2 and PO_2. the presence of different blood particle also vary with age.

4.3 Data Fitting with RapidMiner

Now we have to make our system . Let us consider that we have collected all the data for developing the system. We use RapidMiner for our system. Data mining with RapidMiner is very easy. For the first step of our work we import the input data. For the simplicity of our work we have collect our input data in an excel sheet. So we import the excel sheet in RapidMiner. This will serve as input. The process is illustrated in the Figure 4.1.
Now the data is to be saved in the respiratory. The respiratory is the place where the process and data are saved.Now we will use these imported data for our requirement. We can just drag and drop the excel file to the main process. Now we will apply the operator to the data to make our desired result. We can also apply our own algorithm. This is shown in the Figure 4.2.

Since we have to make decision tree so we use split validation operator. we can also use the cross validation operator. The Split Validation have two parts, One is Training another is Testing. In the training part we use the operator Decision tree. In the testing part we use the appropriate model. In some case we need to measure the performance, then we use the performance operator. In this way we create our system.

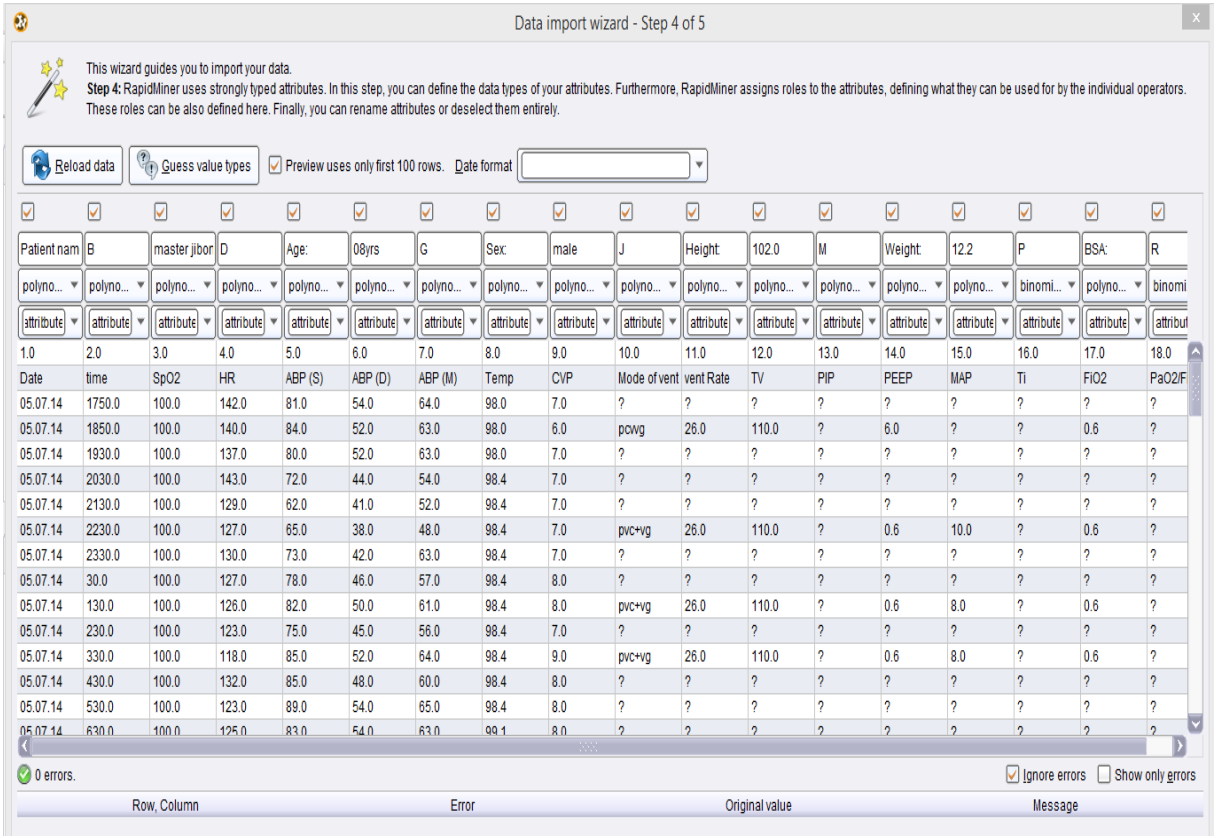


Figure 4.1: Importing Excel sheet to RapidMiner

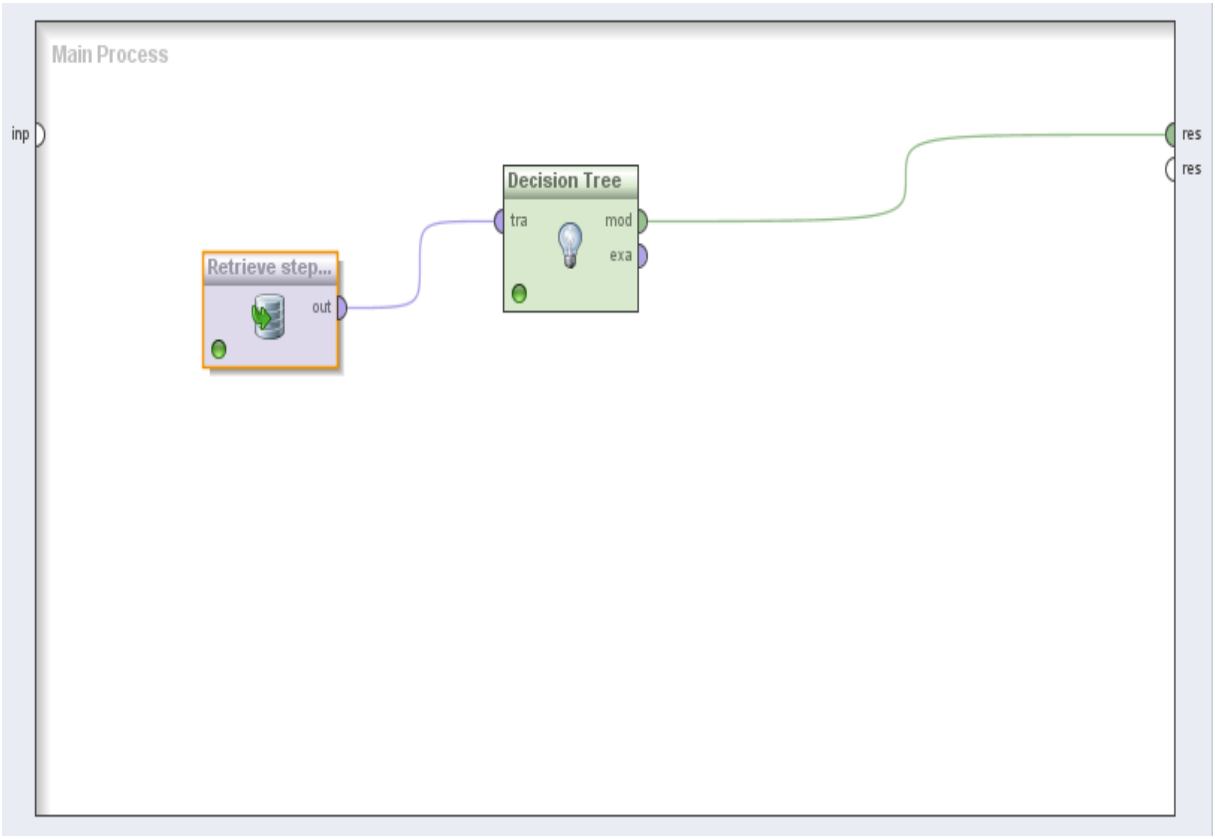


Figure 4.2: Application of operator

CHAPTER 5
RESULT AND DISCUSSION

In the describe figure, decision trees have been used to predict attributes for chances of a patient getting heart disease. The data is analyzed and implemented in Rapidminer tool. It is open source software which consists of a collection of machine learning algorithms for data mining tasks. Data mining finds out the valuable information hidden in huge volumes of data. RapidMiner tool is a collection of machine learning algorithms for data mining techniques. We mainly want to see that for which individual attribute, the condition of heart disease patient is affected.

5.1 Decision Tree

The decision trees created in our work are given in Figure 5.1 and Figure 5.2.

Here is our available data. From this data, after a huge amount of analysis and research and obviously with the help of the doctors, we just made a pattern of the heart disease patient's condition. There are various types of parameter. The outputs we get from inputs tell us the condition of the patient. Sometimes we get the actual result and some we don't get. Then we need to predict the condition of the patient. From a huge list of patient's data records, we compare all the data and able to find a pattern from which we can predict something. In our pattern, we divide the conditions of the patient into three categories. These are: good, average and bad. It can be classified into more fields including these, but because of the shortage of time, we can't go further. This is shown in the Figure 5.3. Sometimes, the prediction is not so good, like confidence level is below average. The reason behind it is the insufficient patient's data. To get a good result, we still need a lot of data with condition. In our example, we use 97 patient's conditional patterns and made decision tree from these data pattern. After that we take some random data and play it on our decision tree. This is shown in the Figure 5.4. Here, we take 6 random data and want to see the prediction. So first of all, we collect data pattern in excel sheet. The unknown conditional data is kept also here at the bottom. The difference is that for this 6 data, condition is empty. Because we want to see the prediction of our machine. Next, we just create our process and played it in RapidMiner. We see that the prediction is totally right. The result is shown in the Figure 5.5.

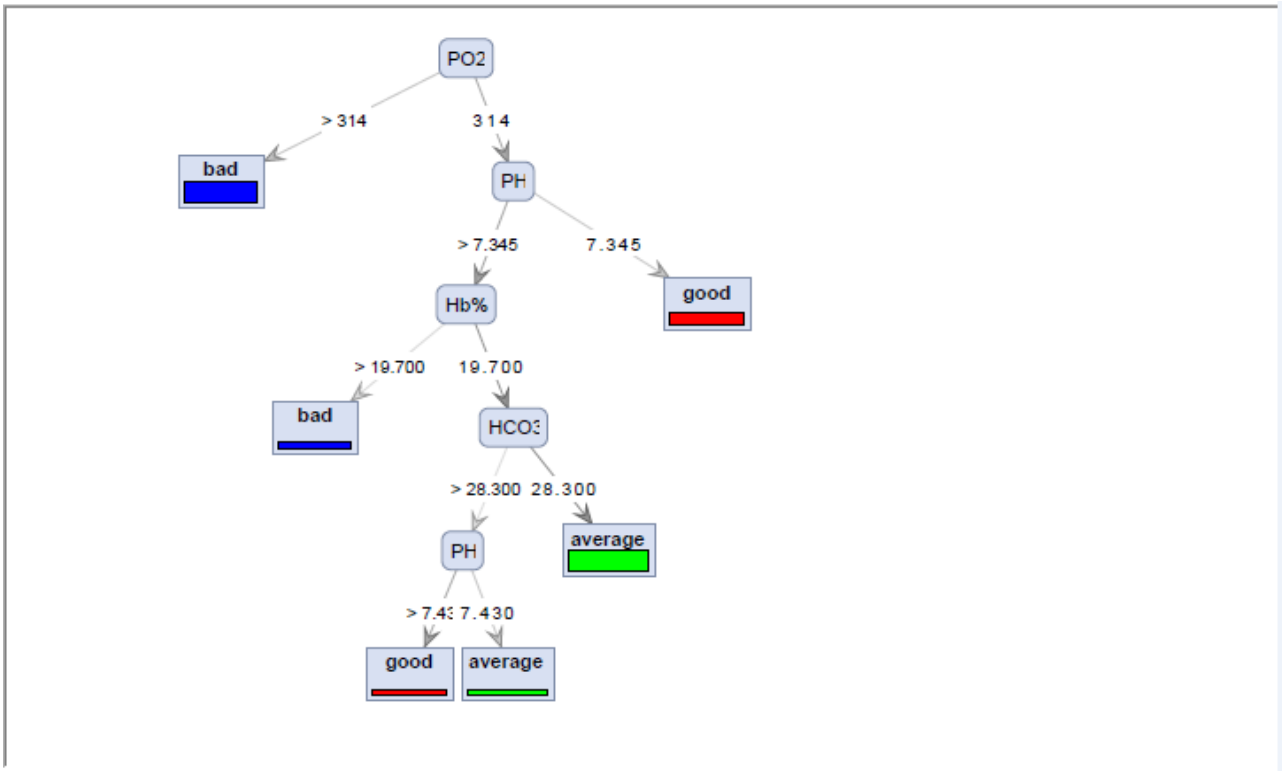


Figure 5.1: Output Tree

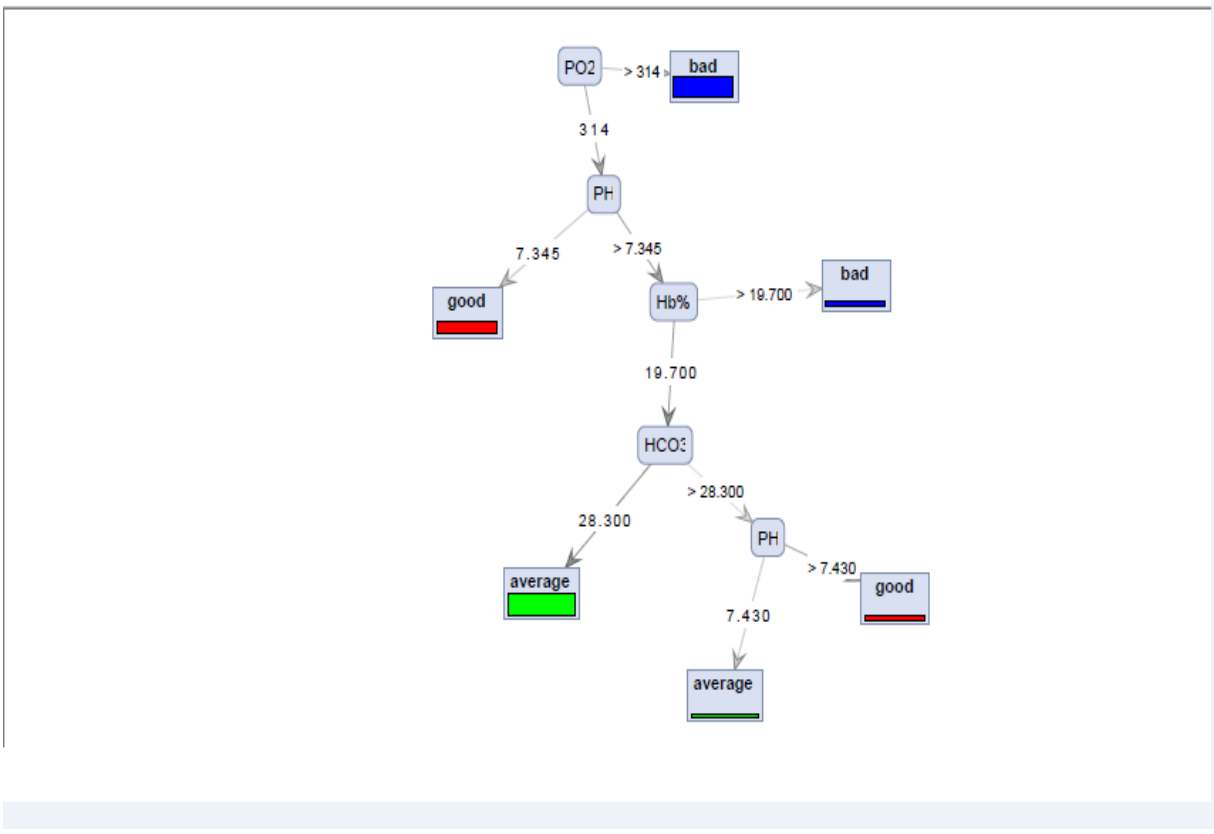


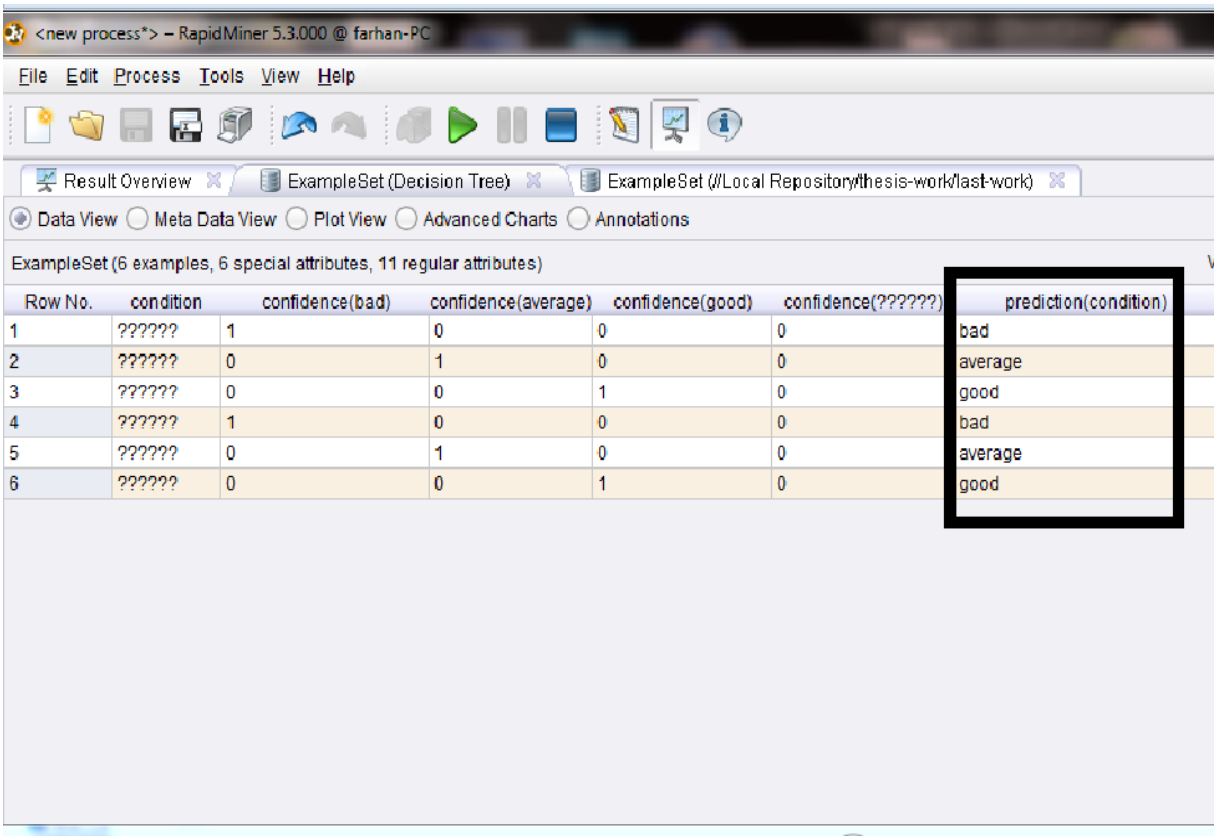
Figure 5.2: Output Tree

B	C	D	E	F	G	H	I	J	K	L	M
PH	PCO2	PO2	HCO3	BE	Hb%	Na	K	Cl	Lact	Ca	condition
7.41	28	389	18.1	3.8	20	139	3.8	111	6.7	0.91	bad
7.36	40	43	22.7	-2.1	20.5	141	3.6	110	2.4	0.83	bad
7.39	36	361	22.2	-1.4	19.5	144	3.7	105	6.7	0.8	bad
7.36	31	362	18	-5.2	17	138	3.1	109	2.4	0.8	bad
7.38	35	357	21	-2.6	17.9	141.4	4.7	106	8.7	0.81	bad
7.42	37	357	24	1	17.3	139.7	4.5	112	2.8	0.8	bad
7.46	29	311	21	0	17.4	141	3.2	107	6.7	0.58	average
7.46	37	274	26	3.8	17.4	141	3.9	107	2.4	0.76	average
7.46	29	350	21	0	15.1	142	2.8	106	8.7	0.5	bad
7.46	33	400	24	2.2	15	138	3.9	107	2.8	0.63	bad
7.45	36	126	25	2.8	16	131	3.3	108	6.7	0.96	average
7.45	36	192	25	2.4	15	139	3.1	100	2.4	0.5	average
7.47	30	321	22	1.1	15.8	137	3.1	98	8.7	0.5	bad

Figure 5.3: Input data format

B	C	D	E	F	G	H	I	J	K	L	M
7.51	30	34	24	3.5	18.5	142	2.3	99	6.7		average
7.51	28	38	23	2.8	19.1	143	3.1	103	2.4	1.72	average
7.49	27	53	21	1.2	18.7	144	3		8.7	0.8	average
7.46	29	32	21	0.4	16.7	140	4.1	101	2.8		average
7.35	37	47	21	3.4	19.3	134	5.2	100	6.7	1.4	average
7.41	33	58	21	1.4	20.1	137	5	98	2.4	1.6	bad
7.41	34	329	21	1.3	18.1	140	4.6	101	8.7	1.12	bad
7.38	35	255	21	2.3	20.7	145	5.2	102	2.8	1.9	bad
7.36	38	248	22	2.4	21.4	145	4.9	99	6.7	1.8	bad
7.39	30	249	18	4.3	17.8	141	3.1	0.5	2.4		average
7.38	33	350	20	3.1	18.9	143	4.9		8.7	1.72	bad
7.37	30	46	17	5.3	17.8	146	3.9		2.8	0.8	average
7.35	32	40	18	5.6	17.7	143	3.9	0.7	1.2		average
7.39	36	361	22.2	-1.4	19.5	144	3.7	105	6.7	0.8	??????
7.46	29	311	21	0	17.4	141	3.2	107	6.7	0.58	??????
7.29	40	262	19	6.4	14.1	144	3.7	98	2.3	0.77	??????
7.4	43	557	27	3.7	11.8	136	3.9	100	8.7	1	??????
7.38	35	248	21	2.5	11.1	141	4.2	100	2.4	0.79	??????
7.34	46	124	25	0.2	12.5	141	3.5	99		0.86	??????

Figure 5.4: Input Data for prediction



Row No.	condition	confidence(bad)	confidence(average)	confidence(good)	confidence(??????)	prediction(condition)
1	??????	1	0	0	0	bad
2	??????	0	1	0	0	average
3	??????	0	0	1	0	good
4	??????	1	0	0	0	bad
5	??????	0	1	0	0	average
6	??????	0	0	1	0	good

Figure 5.5: Result of the predicted data

CHAPTER 6
CONCLUSION

The objective of our work is to provide a study of data mining technique that can be employed in automated heart disease Condition. Various techniques and data mining classifiers are defined in this work which has emerged in recent years for efficient and effective heart disease diagnosis. The analysis shows that different technologies are used in all the papers with taking different number of attributes. So, different technologies used shown the different accuracy to each other. In some papers it is shown that neural networks given the more accuracy in prediction of heart disease. On the other hand, this is also given that Decision Tree has also performed well with accuracy . So, different technologies used shown the different accuracy depends upon number of attributes taken and tool used for implementation. Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amounts of data, researchers are using data mining techniques in the diagnosis of heart disease. Although applying data mining techniques to help health care professionals in the diagnosis of heart disease is having some success, the use of data mining techniques to identify a suitable treatment for heart disease patients has received less attention.

REFERENCES

[1] C. Ordonez, "Association rule discover with the train and test approach for the heart disease prediction," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 2, 2006.

[2] J. Nahar and T. Imam, "Association rule mining to detect factors which contribute to heart disease in males and females," *Expert Systems with Applications*, vol. 36, 2009.

[3] R. Das and I. Turkoglu, "Effective diagnosis of heart disease through neural networks ensembles," *Journal of expert system with applications*, vol. 36, 2009.

[4] H. Yan and J. Zheng, "Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm," *Applied Soft Computing*, vol. 8, 2008.

[5] Y.-J. Park and S.-H. Chun, "Cost-sensitive case-based reasoning using a genetic algorithm: Application to medical diagnosis," *Artificial Intelligence in Medicine*, vol. 51, 2011.

[6] J. Nahar and T. Imam, "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach," *Expert System with Application*, vol. 40, 2013.

[7] K. Polat and S. G. nes, "A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and airs," *Journal of Computer Methods and Programs in Biomedicine*, vol. 88, p. 164, 2007.

[8] K. Polat and S. S. ahan et al, "Automatic detection of heart disease using an artificial immune recognition system (airs) with fuzzy resource allocation mechanism and k-nn (nearest neighbor) based weighting preprocessing," *Journal of expert system with applications*, vol. 32, p. 625, 2007.

[9] P. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," *Journal of Computer and Information Sciences*, vol. 24, no. 27, 2012.

[10] N. M. Nawi and R. G. et al, "The development of improved back-propagation neural networks algorithm for predicting patients with heart disease," *In proceedings of the first international conference ICICA*, vol. 6377, p. 317, 2010.

[11] P. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules and decision tree rules," *Journal of Computer Sciences*, vol. 27, 2012.

[12] M. C. . R. K. S. Dilip Roy Chowdhury, "An artificial neural network model for neonatal disease diagnosis," *International Journal of Artificial Intelligence and Expert Systems*, vol. 2, no. 3, 2011.

[13] "Dictionary for medical science." http://medical-dictionary.thefreedictionary.com.