*International Journals of Advanced Research in Computer Science and Software Engineering*
*ISSN: 2277-128X (Volume-9, Issue-4)*

**Research Article**

**April 2019**

# A Review Paper on the Comparative Study of Intelligent Heart Disease Prediction System

**Monika Singh Patel, Komal Bhatt, Mohit Pathak, Dilkeshwar Pandey**
*CSE and AKTU University, Uttar Pradesh, India*
monikapatel1998@gmail.com, komalb921@gmail.com, mp1941997@gmail.com, dk.pandey@abesit.in

***Abstract:** In order to remove useful information, large amounts of medical data require intelligent equipment. Techniques have been implemented in a number of different areas, including bioinformatics, business, industry, computer vision. Many researchers have been done with the help of such techniques. There is a lack of effective analysis tools to find hidden relationships and trends in medical data from clinical records. Heart disease is considered the leading cause of death worldwide in the last 15 years. Medical data is still rich in information but knowledge is poor. Researchers have used several statistical analyses and various health care techniques or tools to improve the diagnosis accuracy in the medical healthcare service.*

***In this paper, it was majorly discussed all the research work being carried out using the data mining techniques to enhance heart disease diagnosis and prediction including decision trees, Naive Bayes classifiers, K-nearest neighbor classification (KNN), support vector machine (SVM), decision tree and PCA. Results show that SVM and knn perform positively high to predict the presence of coronary heart diseases (CHD). The use of a decision tree is considered as the best-recommended classifier to diagnose cardiovascular disease (CVD). Still, the performance of data mining techniques to detect coronary arteries diseases (CAD) is not encouraging (between 70—80%) and further improvements should be pursued.***

***Keywords**— Bioinformatics, computer vision, decision trees, Naive Bayes classifiers, K-nearest neighbor classification (KNN), support vector machine (SVM).*

## I. INTRODUCTION

Computational intelligence plays an important role in diagnosing and making intelligent decisions. There are a large number of medical applications and diagnostic procedures that can be classified using intelligent computational classification functions. Heart disease also called canary artery disease (CAD), is a generalized term that can be used to relate to any of the symptoms related to the heart. The diagnosis of diseases is a difficult but important task in medicine. "Detection of cardiovascular disease by various factors or symptoms is a multi-layered issue, which is often not freed from false assumptions with unpredictable effects". CAD is the most common form of cardiovascular disease where coronary arteries are narrow and sometimes wide, resulting in a heart attack. Death rates due to heart disease are increasing rapidly around the world, and approximately 500,000 women die every year due to heart attacks. Data mining technology provides users with ways to find new and underlying patterns from large-scale data. [3]In the healthcare domain, the knowledge discovered can be used to improve the accuracy of diagnosis by healthcare administrators and medical physicians, to enhance and reduce the goodness of surgical operation. “The extraction of hidden previously unknown and potentially useful information about data" is termed as Knowledge discovery in data [1].

The purpose of predictions in data mining is to help discover trends in patient data in order to improve their health [4]. Due to a change in lifestyles in developing countries, like South Africa, Cardio Vascular Disease (CVD) has become a leading cause of deaths [5]. CVD is projected to be a single largest killer worldwide accounting for all deaths [6]. An endeavor to exploit knowledge, experience and clinical screening of patients to diagnose or recognize heart attacks is regarded as a treasured opportunity [2]. Data mining plays an important role to predict diseases in the health sectors. The predictive end of the research is a data mining model. In this paper, the results of techniques were discussed and conclusions about future research were given. According to the World Health Organization (WHO) statistical data, CAD can result in abnormalities or permanent disability in many men and women; and almost one-third of the world's deaths occurred in developing countries until 2010. The prediction of disease made by a health practitioner is not 100% accurate.

## II. OVERVIEW OF THE PAPER

The third chapter is a literature survey on a medical decision support system. The fourth chapter contains the introduction of heart disease. The fifth chapter contains dataset information. The sixth chapter contains data mining techniques. The seventh chapter contains data mining algorithms. The eighth chapter contains comparative study and experimental results. The ninth chapter contains a conclusion. The tenth chapter contains references.

## III. LITERATURE REVIEW

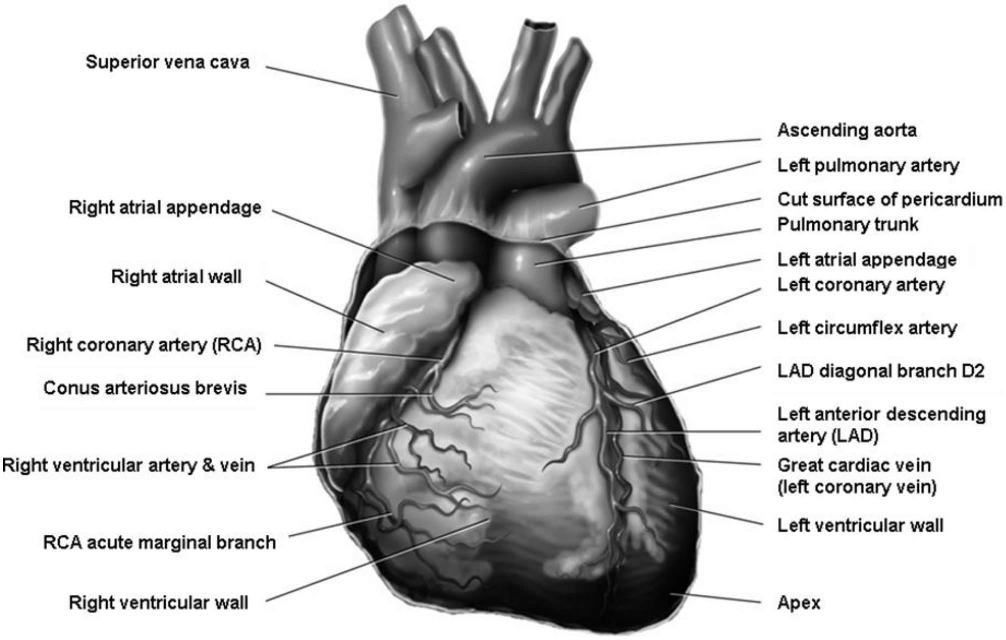
Many works in the literature related to heart disease diagnosis have led to this work by using data mining techniques. A brief literature survey is presented here. [4]A Model Intelligent Heart Disease Prediction System (IHDPS) with the help of data mining techniques such as DT, NB, and NN Sellappan was proposed by Palaniappan et al. The results explained the specificity of each method. In understanding the objectives of specified mining objectives. A novel technique to develop a multi-parametric feature with linear and non-linear features of HRV (heart rate Variability) was proposed by Hon Gue Lee et al. [9]IHDPS was able to answer questions which were conventional decision Support Systems were not enabled. Facilitated by the establishment of important knowledge between patterns, relationships, medical factors related to cardiovascular disease. IHDPS maintains well web-based, user-friendly, scalable, reliable and expandable.[11]Statistical and classification techniques were used to develop the multi-party specialty of HRV. In addition, they have assessed the linear and non-linear properties of HRV for three recumbent positions, to be precise, carefree, left background and right playback position. To develop the multi-parametric feature with linear and nonlinear characteristics of HRV (Heart Rate Variability) a novel technique was proposed by Heon Gyu Lee et al. [13]. Researchers have used several classifiers e.g. Bayesian Classifiers, CMAR (Classification based on Multiple Association Rules), C4.5 (Decision Tree) and SVM (Support Vector Machine).

- Franck Le Duff et al. [11] builds a decision tree with a database of the patient for a medical problem.
- Lathe Parthiban et al. [12] projected an approach on the basis of a coactive neuro-fuzzy inference system (CANFIS) for the prediction of heart disease. The CANFIS model uses neural network capabilities with the fuzzy logic and genetic algorithm.

It has been seen that the Comparison of traditional analysis and data mining analysis illustrated the contribution of the data mining method in the sorting of variables and concluded the significance or the effect of the data and variables on the condition of the study. [14]A major drawback of the process was knowledge acquisition and the need to collect adequate data to create an appropriate model. The researchers [7] used pattern recognition and data mining methods in predicting models in the domain of cardiovascular diagnoses. The experiments were carried out using classification algorithms Naïve Bayes, Decision Tree, K-NN and Neural Network and results prove that the Naïve Bayes technique outperformed other used techniques.

## IV. HEART DISEASE INTRODUCTION

The most common type of these diseases is coronary arterial disease (CAD) in which the coronary arteries are rigid and tight. The heart disease (CVD) refers to a wide range of conditions affecting the heart and blood vessels and the way the blood is pumped and circulated throughout the body. There are many comments and symptoms used by physicians to diagnose heart diseases. CVD is most likely to cause serious illness, disability, and death.



**Figures 1: Diagrammatic representation of heart**

The narrowing of the coronary arteries decreases the oxygen supplied to the heart and becomes the so-called coronary heart disease (CHD). Age, gender, chest pain type, blood pressure, cholesterol, fasting blood sugar, maximum heart rate, and hereditary meaningful symptoms. Apart from this, other habits, including stress, overweight, smoking, alcohol intake, and low exercise can be used. The sudden blockage of the coronary artery is usually caused by a blood clot that can cause a heart attack. Chest pain occurs when the blood received from the heart muscles is insufficient and unrelated.

## V. DATASET INFORMATION

Standard heart disease dataset from UCI repository [10] has been used for training and testing purpose. The database contains 76 attributes, but we have used only 14 of them in order to obtain the accurate results using less number of feature space. Dataset contains 303 samples and 13 input features as well as 1 output feature. A list of all those features is given in Table 1. Some important calculations are given in the table.

| Feature No. | Feature Name | Description |
|---|---|---|
| 1 | Age | Age in Years |
| 2 | Sex | 1=male<br>0=female |
| 3 | Cp | Chest Pain Type:<br>1=typical angina<br>2=atypical angina<br>3=non-angina pain<br>4=asymptomatic |
| 4 | Trestbps | Resting blood pressure (in mm Hg) |
| 5 | Chol | Serum cholesterol in mg/dl |
| 6 | Fbs | Fasting Blood Sugar > 120 mg/dl:<br>1=true<br>0=false |
| 7 | Resteg | Resting electrocardiographic results:<br>0 = normal<br>1 = having ST-T wave abnormality<br>2 =showing probable or define left ventricular hypertrophy by Estes ‘criteria |
| 8 | Thalach | Maximum heart rate achieved |
| 9 | Exang | Exercise induced angina:<br>1 = yes<br>0 = no |
| 10 | Oldpeak | Depression induced by exercise relative to rest |
| 11 | Slop | The slope of the peak exercise segment:<br>1 = up sloping<br>2 = flat<br>3= down sloping |
| 12 | Ca | Number of major vessels colored by fluoroscopy that ranged between 0 and 3. |
| 13 | Thal | 3 = normal<br>6= fixed defect<br>7= reversible defect |
| 14 | Num | Diagnosis classes:<br>0 = healthy<br>1= patient who is subject to possible heart disease |
| 15 | Pid | Patient id diagnosed accordingly |

## VI. DATA MINING TECHNIQUES IN HEALTH CARE

After exploring some types of heart arteries diseases and symptoms that when have certain values denote a heart disease, in this section we will explore different data mining techniques applied generally to healthcare. The various features like Symptoms and patient records and such huge amounts of data can be used for knowledge discovery in the health care domain. A general framework proposed by for medical data mining. The framework starts with a specific medical problem wherein a dataset should be pre-processed and cleaned before mining the data using one of the available data mining tools. Knowledge evaluation comes at the last and expertise from the medical domain should involve.

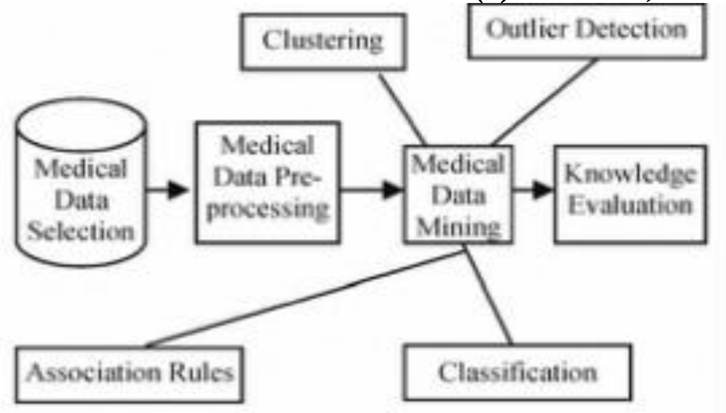


**Figure 2: knowledge discovery in health care**

## VII. DATA MINING ALGORITHMS

Research on data mining has led to the formulation of several data mining algorithms. The various algorithms can be directly used on a dataset for creating some models or to draw vital conclusions and inferences from that dataset. Some popular data mining algorithms are KNN and SVM etc. They are discussed in the follows section.

### K-Nearest Neighbour: -

This algorithm is arguably the simplest machine learning algorithm For a new data point, the algorithm finds the closest data points in the training dataset, known a"nearest neighbors".▪ KNN is used in many applications such as

- Classification and interpretation
- problem-solving
- Function learning and teaching & training.

KNN is one of the most simple and straight forward data mining techniques. It is called Memory-Based Classification as the training examples need to be in the memory at run-time [11]. When dealing with continuous attributes the difference between the attributes is calculated using the Euclidean distance i.e. dist(p,q) = sqrt( (p1-q1)2 + p2-q2)2 + ….+ pn-qn)2 ) .

### SVM Classifier

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space, this hyperplane is a line dividing a plane into two parts wherein each class lay in either side.

Suppose you are given the plot of two label classes on the graph as shown in the image (A). Can you decide a separating line for the classes?

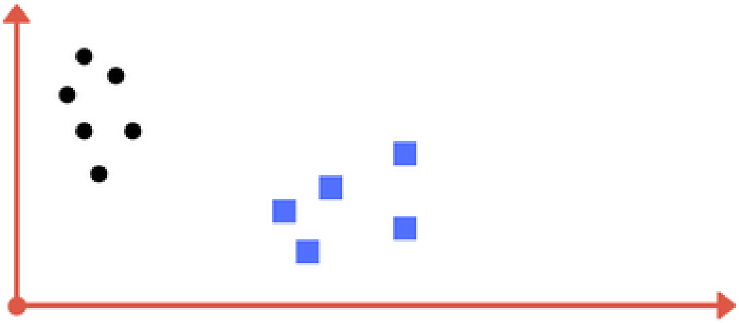


**Figure 3: Image A: Draw a line that separates black circles and blue squares.**

You might have come up with something similar to the following image (image B). It fairly separates the two classes. Any point that is left of the line falls into the black circle class and on the right falls into the blue square class. Separation of classes. That's what SVM does. It finds out a line/ hyperplane (in multidimensional space that separate outs classes). Shortly, we shall discuss why I wrote multidimensional space.

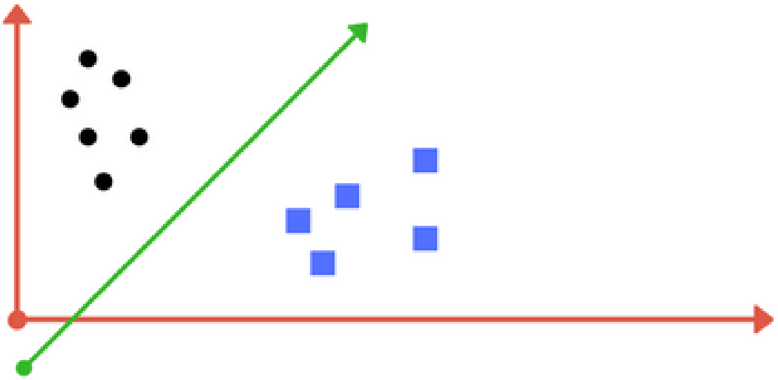


**Figure4: Image B: Sample cut to divide into two classes.**

**Naive Bayes**

Naive Bayes classifier is based on Bayes theorem. This classifier algorithm uses conditional independence, which means it assumes that an attribute value on a given class is independent of the values of other attributes. [12] The Bayes theorem is as follows: Let X={x1, x2, ....., xn} be a set of n attributes. In Bayesian, X is considered as evidence and H be some hypothesis means, the data of X belongs to specific class C. We have to determine P (H|X), the probability that the hypothesis H holds given evidence i.e. data sample X.

According to Bayes theorem, the P (H|X) is expressed as

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

**Decision Tree Classification**

Decision Tree Classifier repetitively divides the working area(plot) into subpart by identifying lines. (repetitively because there may be two distant regions of same class divided by other as shown in the image below)

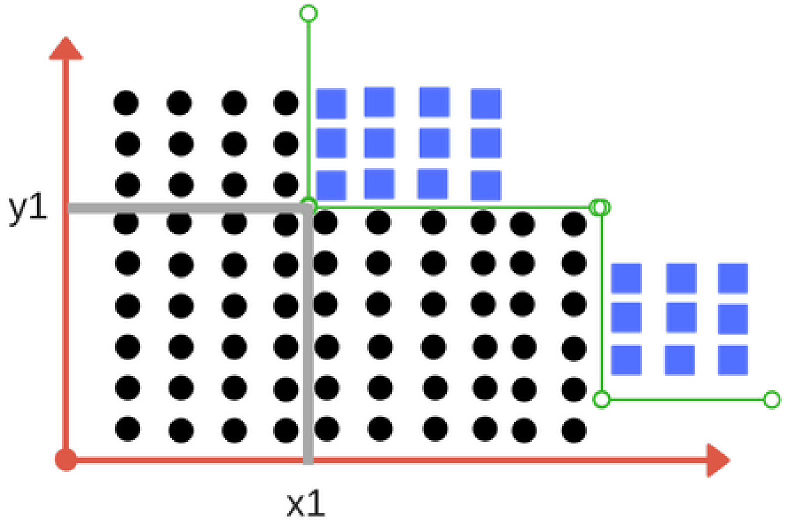


**Figure 5: Decision Tree Classifier, repetitively divides the working area(plot) into sub part by identifying lines**

The algorithm for the decision tree is given below:
Step 1: Identify the information gain for the attributes in the dataset.
Step 2: Sort the information gain for the heart disease datasets in descending order.
Step 3: After the identification of the information gain assign the best attribute of the dataset at the root of the tree.
Step 4: Then calculate the information gain using the same formula.
Step 5: Split the nodes based on the highest information gain value.
Step 6: Repeat the process until each attributes are set as leaf nodes in all the branches of the tree

For example [4]The J48 algorithm grows an initial tree using the divide and conquers technique. Fig 1 shows the visualization of the tree from modeling the dataset using the J48 algorithm. The tree is pruned to evade over fitting. The tree-construction in J48 differs with the tree-construction in several respects from REPTREE structure.

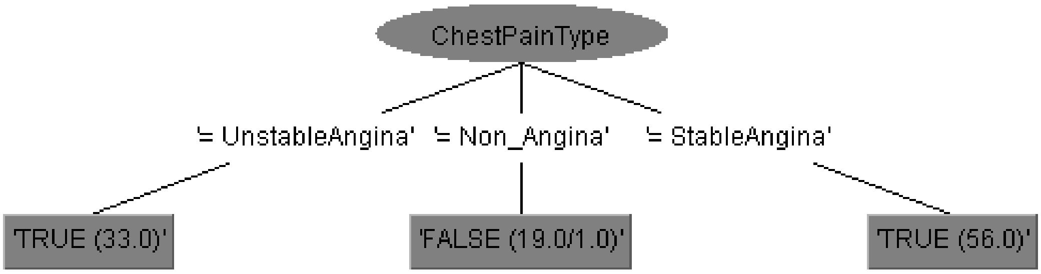


**Figure 6. J48 Pruned Tree**

## VIII. COMPARATIVE STUDY AND EXPERIMENTAL RESULTS

After exploring some types of Research on data mining has led to the formulation of several data mining algorithms. These algorithms can be directly used on a dataset for creating some models or to draw vital conclusions experimental results. A list of all those Comparative study and experimental results is given in Table 2.

| **Data Mining Techniques** | **Purpose of study** | ***Maximum Accuracy*** | **Ref.** |
|---|---|---|---|
| Naïve Bayes Classifier<br>*+GA Feature Reduction* | To diagnose the presence of CVD. | **96.5%** | Anbarsi *at el.* [19] |
| Decision Tree<br>*+GA Feature Reduction* | To diagnose the presence of CVD. | **99.2%** | Anbarsi *at el.* [19] |
| SVM | To diagnose the presence of CHD. | **92.1%** | AbuKhousa and Campbell [12] |
| KNN | To diagnose CAD | **61.39%** | Alizadehsani *et al.* [9] |
| K-Nearest Neighbor(KNN) + | To diagnose CAD | **85.44%** | Proposed approach |
| Principle Component | | | |
| Analysis(PCA) + Decision Tree | | | |
| Support Vector Machine(SVM) | To diagnose the presence of CHD | **74.77%** | Proposed approach |

## IX. CONCLUSION

The prediction of cardiovascular disease is useful for cardiovascular patients, in which the patient has a record. Paper proposes a method to examine the performance of two different classification algorithms such as Decision Tree and Heart Disease Supporting the Vector Machine on Dataset. Uses an application on research various data mining algorithms for predicting heart attacks and to compare the best method of prediction. Search the result does not present a dramatic difference prediction while using different classification algorithms data mining. Data mining techniques applied for CVD and CHD are promising.[1] Results showed that the optimization and feature reduction utilizing GA or principal component analysis (PCA) for a certain disease may strongly increase the accuracy of a classifier. [3] It is found that decision trees and Naïve Bayes classifiers are recommended for CVD diagnosis with an accuracy reaching more than 95%.2. The model from the classification will be able to answer more complex queries in the prediction of heart attack diseases The overall objective of our work is to predict more accurately the presence of heart disease.[11] In this paper, two more input attributes obesity and smoking are used to get more accurate results The dataset has a large volume of data which consumes more time for classification. Hence in this reduction of the dimensionality of data using the attribute selection methods. [15] Then the reduced data is classified using various classification algorithms. We found that the NB classifier gives better accuracy for heart disease prediction after applying the CFS attribute selection method. Then the reducing data is classified using different classification algorithms. We found that the support vector machine with the decision tree along with knn gives better accuracy than other algorithms for the cardiovascular disease prediction system.

## REFERENCES

[1] M. Bramer, *Principles of Data Mining*: Springer-Verlag, 2007.

[2] Z. Jitao and W. Ting, "A general framework for medical data mining," presented at the International Conference on Future Information Technology and Management Engineering (FITME), 2010.

[3] S. Oyyathevan and A. Askarunisa, "An expert system for heart disease prediction using data mining technique: Neural network," *Lecture Notes on Information Theory Vol. 2, No. 4, December 2014*

[4] S. . Ishtake and S. . Sanap, "' Intelligent Heart Disease Prediction System Using Data Mining Techniques '," *International Journal of healthcare & biomedical Research*, vol. 1, no. 3, pp. 94–101, 2013.

[5] V. Chaurasia, "Early Prediction of Heart Diseases Using Data Mining," *Caribbean Journal of Science and Technology*, vol. 1, pp. 208–217, 2013.

[6] K. Sudhakar, "Study of Heart Disease Prediction using Data Mining," vol. 4, no. 1, pp. 1157–1160, 2014.

[7] T. J. Peter and K. Somasundaram, "AN EMPIRICAL STUDY ON PREDICTION OF HEART DISEASE USING CLASSIFICATION DATA MINING TECHNIQUES," 2012

[8] Rajkumar, A. and Reena, G.S.: Diagnosis of Heart Disease Using Datamining Algorithm. In: Global Journal of Computer Science and Technology, Vol. 10 (2010).

[9] Palaniappan, S., Awang, R.: Intelligent Heart Disease Prediction System Using Data Mining Techniques. 978-1-4244-1968-5/08/ ©IEEE (2008)

[10] M. J. Abdi and D. Giveki, "Automatic detection of erythematosquamous diseases using PSO–SVM based on association rules",Engineering Applications of Artificial Intelligence, vol. 26, (2013), pp.603-608.

[11] Franck Le Duff, Cristian Munteanb, Marc Cuggiaa, Philippe Mabob, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", Studies in health technology and informatics, Vol. 107, No. Pt 2, pp. 1256-9, 2004.

[12] Latha Parthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences, Vol. 3, No. 3, 2008.

[13] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2007

[14] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008

[15] Andreeva P., M. Dimitrova and A. Gegov, Information Representation in Cardiological Knowledge Based System, SAER'06, pp: 23-25 Sept, 2006.

[16] Latha Parthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFlS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences 3; 3, 2008