Efficient heart disease prediction system using data mining techniques and neural networks

[1]Shaikh Mohd.Shafique,[2]DharmendraChoukse

[1]Student ,[2]Associate professor
Department of computer science and engineering,
Swami Vivekanand college of engineering
Indore, India.

***Abstract*- Data mining is gaining importance in every field such as health industry, e-commerce, marketing, IT sector, Artificial Intelligence, Machine learning and Data science, Business and Finance,Govt organizations etc .In modern day to day life and unnecessary tensions and thinking human beings are suffering from serious health problems such as diabetes and heart disease. Due to busy life schedule and no timely food and rest leads to acidity, diabetes, blood pressure and risk for heart disease. The health care industry is rich in information but poor in knowledge. Various data mining techniques are used to predict the heart disease from the data sets. Data mining is a component of a wider process called knowledge discovery from databases. It involves scientists from a wide range of disciplines such as computer scientists, mathematicians and statisticians,as well as those working in the fields such as machine learning, data science and artificial intelligence, information retrieval and pattern recognition. Before a data set can be mined, it first has to be cleaned which removes errors and ensures consistency and takes mining techniques time is reduced in detecting missing values into account. By using the developed system, one can predict the disease with best accuracy.**

***IndexTerms*- Data mining,KDD, ANN, cholesterol, obesity, diabetes, blood pressure.**

I. INTRODUCTION

Heart is a vital organ of human body. A healthy heart is necessary for good health. Now a days due to bad food habits and junk food such as pizzas, burgers etc. and lack of physical exercise human beings are suffering from cholesterol ,blockages and various other diseases such asdiabetes which leads to improper functioning of the heart and other organs of the body. In this review paper we discuss the various problems caused by heart disease and efficient prediction and detection of the same so that an individual can take preventive measures time to time and get diagnosis of the disease in time. There are a number of factors which increase the risk of having a heart disease.

The World Health Organization (WHO) has surveyed that 12 million deaths that occur throughout the world are due to heart disease. Approximately 17.3 million people died throughout the world due to heart disease in the year 2008.WHO estimated that by the year 2030, 80% of the deaths worldwide will be due to heart disease. By using various data mining techniques we can predict the disease accurately .An efficient heart disease prediction system can discover and extract hidden knowledge related with heart disease from a historical heart disease database .It can answer complex queries for diagnosing heart disease and thus help the medical practitioners to take accurate decisions and diagnose the disease efficiently.

II. LITERATURE REVIEW

Data mining is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. With the widespread use of database and the explosive growth in their sizes, organizations are faced with the problem of information overload. Data mining techniques support automatic exploration of data.

Data mining attempts to source out patterns and trends in the data and infers rules from these patterns. With these rules the user will be able to support, review and examine decisions in some related business or scientific area. Data mining or knowledge discovery in databases is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. This includes a number of technical approaches such as clustering, data summarization, classification, finding dependency networks, analyzing changes and detecting anomalies.

Symptoms of heart attack can include:

- Uncomforted pain in the chest, arm or below the breast bone. Fullness, indigestion, heart burn or burning sensation in the chest.
- Sweating, vomiting, nausea.
- Difficulty in breathing, uneasiness, rapid or irregular heartbeats.

Stages of KDD:-

The stages of KDD, starting with the raw data and finishing with the extracted knowledge are as follows.

*Selection:-*This stage is concerned with selecting or segmenting the data that are relevant to some criteria.

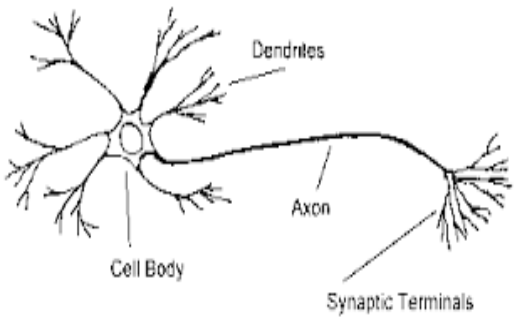
Preprocessing:-It is the data cleaning stage where unnecessary information is removed. When the data is drawn from several sources it is possible that the same information is represented in different sources in different formats. This stage reconfigures the data to ensure a consistent format, as there is a possibility of inconsistent formats.

Transformation: - The data is not merely transferred across, but transformed in order to be suitable for the task of data mining. In this stage the data is made usable and navigable.

Data mining:-This stage is concerned with the extraction of patterns from the data.

Interpretation and evaluation:-The patterns obtained in the data mining stage are converted into knowledge, which in turn, is used to support decision making.

Data visualization: - It makes it possible for the analyst to gain a deeper, more intuitive understanding of the data. Data visualization helps users to examine large volumes of data and detect the patterns visually. Visual displays of data such as maps, charts and other graphical representations allow data to be presented compactly to the users.
Artificial Neural Network: - Artificial neural networks models have been studied for many years in the hope of achieving human like performance in several fields. In Neural Networks, basic elements are neurons or nodes. These neurons are interconnected and within the network they worked together in parallel in order to produce the output functions. From existing observations they are capable to produce new observations even in those situations where some neurons or nodes within the network fails or go down due to their capability of working in parallel. An activation number is associated to each neuron and a weight is assigned to each edge within a neural network. In order to perform the tasks of classification and pattern recognition neural network is mainly used. ANN is based on the biological neural networks in the human brain and described as a connectionist model Fig -1: A Sketch of a Neuron in the Human Brain



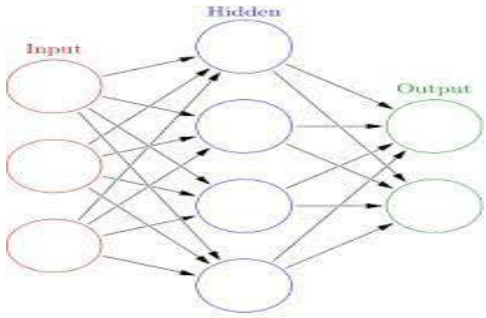


Fig1. A Sketch of a Neuron in the Human Brain Fig 2 Artificial Neural Networks

It is based on the neuron, a cell that processes information in the human brain. The neuron cell body contains the nucleus, and has two types of branches, the axon and the dendrites. The axon transmits signals or impulses to other neurons while the dendrites receive incoming signals or impulses from other neurons. Every neuron is connected and communicates through the short trains of pulses. The nodes are the artificial neuron and the directed edges represented the connection between output neurons and the input neurons. In training phase, the internal weights of the neural network are adjusted according to the transactions used in the learning process. For each training transaction the neural network receives in addition the expected output. This allows modification of weight.

Factors causing heart disease include
- ❖ Family history of heart disease
- ❖ Obesity
- ❖ Cholesterol
- ❖ Physical inactivity
- ❖ Hypertension
- ❖ Poor diet
- ❖ High blood pressure
- ❖ High blood cholesterol
- ❖ Smoking

III METHODOLOGY
Data mining techniques like clustering, association rule mining, classification algorithms such as decision tree, C4.5 algorithm, Naïve Bayes are used to explore the different kinds of heart based problems. Data mining methods such as k-means clustering and c4.5 algorithm are used for validating the accuracy of medicinal data. These algorithms can be used to enhance the data storage for practical and legal purposes.
Various data mining techniques and algorithms are used such as association, clustering, decision trees for the efficient prediction and detection of cardiovascular heart attacks which leads to sudden deaths. When the cholesterol level of bad cholesterol increases

in the blood it narrows the passage of blood flow of arteries and veins which leads to blockage of blood in the blood vessels and this results in heart attacks.

DATA MINING TOOLS

There are various data mining tools used for data mining purpose. These are WEKA, TANAGRA, MATLAB and .NET FRAMEWORK.

.*NET FRAMEWORK:* It is a software framework developed by Microsoft which runs primarily on Microsoft windows. It provides secure communication and consistent applications. It provides language interoperability (each language can code written in other languages) across several programming languages.

WEKA: It is a data mining tool which was developed in New Zealand by the University of Waikato that implements data mining algorithms using JAVA language. WEKA is a collection of machine learning algorithms and their application to the data mining problems. These algorithms are directly applied to the dataset. WEKA supports data file in ARFF format. WEKA is open source software and hence, it is not dependent on any platform. It includes algorithms for data processing, classification, regression, clustering, association and also visualization tools.

TANAGRA: It is open source software as researchers can access to the source code and add their own algorithms and compare their performances, if it conforms to the software distribution license. It includes several data mining algorithms from statistical learning, machine learning, and data analysis and database area.

MATLAB: It is a data mining tool built in high level language. It provides interactive environment for visualization, numerical computation and programming. The built in math functions, language and tool explore various approaches and helps to reach a solution faster than with the spreadsheet of traditional programming languages like C,C++

J48 TECHNIQUE

Decision tree is a kind of classifying and predicting data mining technology, belonging to inductive learning and supervised knowledge mining technology. As decision tree is advantageous in fast construction and generating easy-to-interpret If-Then decision rule, it has become the most widely applied technique among numerous classification methods. Decision tree algorithm has been applied in many medical tasks, for examples, in increasing quality of dermatologic diagnosis, predicting essential hypertension, and predicting cardiovascular disease. Decision tree is one of the most popular tools for classification and prediction. Production of a decision tree is an efficient method for classification of data. This tree using a top-down strategy to build a test on each node. J48 decision tree method is the implementation of c4.5 decision tree in weka data mining tool. J48 decision tree supports continuous and discrete features. It can also manage features with missing value.

NAÏVE BAYES ALGORITHM

Naïve Bayes Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. A naive Bayes classifier is a term dealing with a simple probabilistic classification based on applying Bayes theorem. In simple terms, a naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. It learns from the “evidence” by calculating the correlation between the target (i.e., dent) and other (i.e., independent) variables. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting Naive Bayes classifiers often work much better in many complex real-world situations than one might expect. Here independent variables are considered for the purpose of prediction or occurrence of the event. The algorithm is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. For example, a patient may bed to have certain symptoms. Based on the observation, Bayes' theorem can be used to compute the probability that a proposed diagnosis is correct. Bayes Theorem finds the probability of an even occurring given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Bayes theorem can be stated as follows, P (B given A) = Prob (A and B) / Prob (A) To calculate the probability of B given A, the algorithm counts the number of cases where A and B occur P (Ci/X)>P (Cj/X) for all 1< = j< = m and j! = i.

NEURAL NETWORK

Artificial Neural Network is a data processing algorithm, originated from human brain. The system includes a large number of tiny processors to handle data processing. The processors act in the form of an interconnected network parallel to each other to solve a problem. Using programming knowledge, in this networks a data structure is designed that can act as neurons. This data structure is called the neuron.

Neural network is a parallel, distributed information processing structure consisting of numerous quantities of processing elements called node, they are interconnected via unidirectional signal channels called connections. Each processing element has a single output connection that branches into many connections and each conveys the equivalent signal.

The NN can be classified in two main groups according to the way they learn. They are supervised learning and unsupervised learning. In supervised learning the network compute a response to each input and then compares it with the target value. If the computed response differs from the target value, the weights of the network are adapted according to a learning rule. Examples of

supervised learning are Single layer perceptron and Multilayer perceptron. In unsupervised learning the networks learn by identifying special features in the problems they are exposed to. Example for unsupervised learning is self organizing feature maps.

Table I. Comparison of data mining tools with techniques and accuracy

Data Mining tools	Techniques	Accuracy
Weka 3.6.4	J48 Technique	95.56%
Weka 3.6.4	Naive Bayes	92.42%.
Weka	Neural Network	79.19%
TANAGRA	Fuzzy Logic	83.85%
Weka 3.6.6	Naive Bayes	99.52%
TANAGRA	Decision Trees	52.33%
.NET data mining tool	Neural Network	96.5%

IV. RESULTS AND DISCUSSION

In medical field, Data mining provides various techniques and has been widely used in clinical decision support systems that are useful for predicting and diagnosis of various diseases. These data mining techniques used in heart diseases takes less time and make process fast for the prediction system to predict heart diseases with good accuracy in order to improve their health. In this work, the dataset used in the system as visualized in weka tool is as shown in figure below:

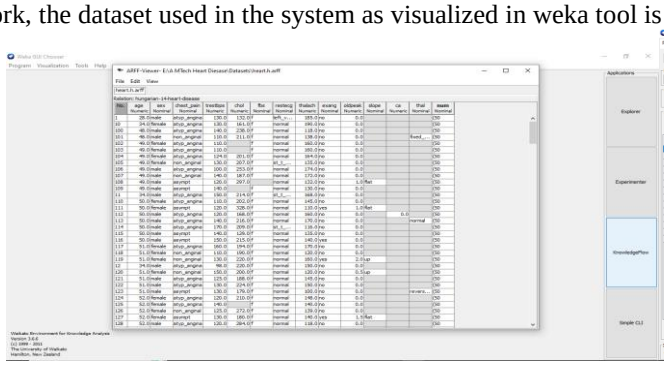


Fig3. Dataset used for training

Fig4. Dataset Attributes

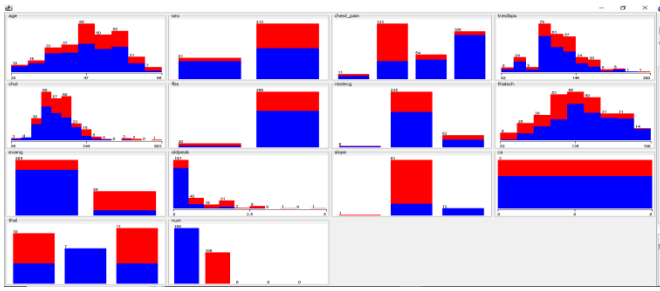


Fig5 Dataset Visualization

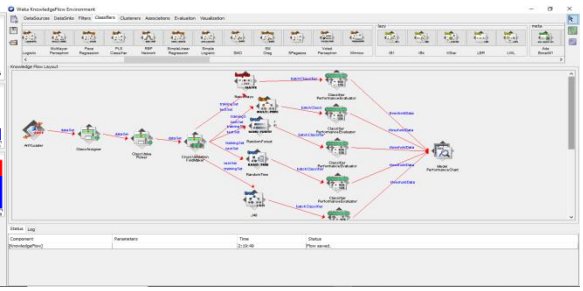


Fig6 Modal Evaluation

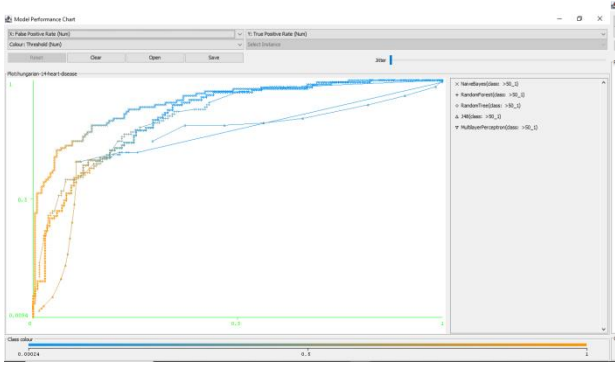


Fig7 Classifiers Performance 1

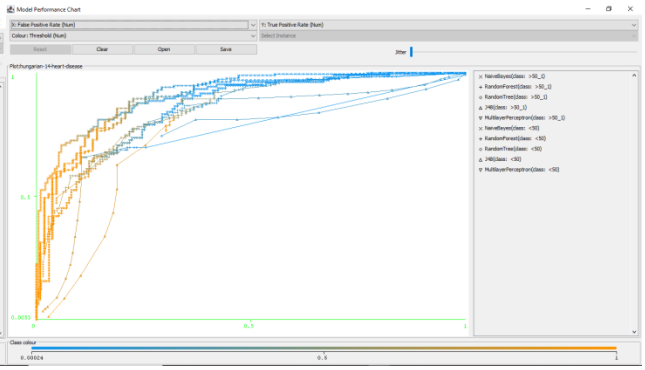


Fig8 Classifier Performance 2

Classifier	Correctly Classified Instances
Naive Bayes	85.03%
Decision table	83.33%
Random Forest	83.33%
Random Tree	100%
J48 (Tree)	84.01%
MultilayerPerceptron	96.94%

Fig9. Classifier Performance 3

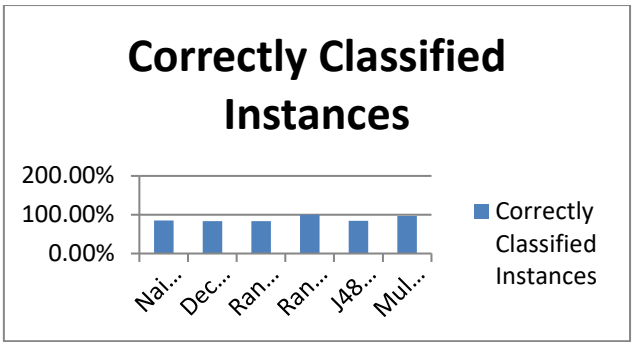


Fig10. Classifier Performance 4

Classifier	ROC Area (<50)	ROC Area (>50_1)
Naive Bayes	0.919	0.919
Decision table	0.826	0.838
Random Forest	0.826	0.838
Random Tree	1	1
J48 (Tree)	0.825	0.825
Multilayer Perceptron	0.987	0.986

Fig11. Classifier Performance 5

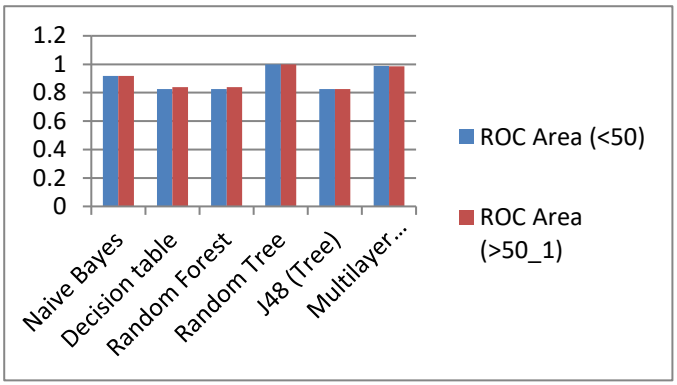


Fig12. Classifier Performance 6

The thesis will be of help to the doctor's to efficiently and effectively detect the cardiovascular disease based on the training algorithm and input parameters supplied to the neural network for training the algorithm. The use of decision trees, association rules, *WEKA* tool and various data mining techniques will be useful to the doctors to detect heart attack symptoms early so that they can diagnose the patient with appropriate skill and in this way they can attempt to save the precious lives of human beings which are dying due to heart attacks.
As there are deaths due to heart attack at very rapid rate due to tensions, incorrect food habits, lack of physical movement and regular exercise, diabetes. The doctor can advise the patients early so that he can prevent from being dead early due to sudden attack.

Our research will provide enough data to the medical field to predict cardiovascular heart attacks so that a patient can take precautions to avoid sudden death due to heart attack. It will help the doctors to advise the patient that such types of attacks are going to occur in the near future so that the patient can get checkup of blood pressure, bad cholesterol and lipid profile. A person suffering from diabetes is also at risk of heart attacks.

REFERENCES:

1. A. K. Sen, S. B. Patel, and D. P. Shukla, "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level," *International Journal of Engineering and Computer Science*, vol. 2, no. 9, pp. 1663–1671, 2013.
2. 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.
3. V. Chaurasia, "Early Prediction of Heart Diseases Using Data Mining," *Caribbean Journal of Science and Technology*, vol. 1, pp. 208–217, 2013.
4. D. S. Chaitrali and A. S. Sulabha, "A Data Mining Approach for Prediction of Heart Disease Using Neural Networks," *International Journal of ComputerEngineering& Technology (IJCET)*, vol. 3, no. 3, pp. 30–40, 2012.
5. M. Jabbar, P. Chandra, and B. Deekshatulu, "CLUSTER BASED ASSOCIATION RULE MINING FOR," *Journal of Theoretical & Applied Information Technology*, vol. 32, no. 2, pp. 196–201, 2011.
6. R. Rao, "SURVEY ON PREDICTION OF HEART MORBIDITY USING DATA MINING TECHNIQUES," *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol. 1, no. 3, pp. 14–34, 2011.
7. S. Vijiyarani and S. Sudha, "Disease Prediction in Data Mining Technique – A Survey," *International Journal of Computer Applications & Information Technology*, vol. II, no.I, pp. 17–21, 2013.

8. T. J. Peter and K. Somasundaram, "AN EMPIRICAL STUDY ON PREDICTION OF HEART DISEASE USING CLASSIFICATION DATA MINING TECHNIQUES," 2012.
9. S. B. Patil and Y. S. Kumaraswamy, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 9, no. 2, pp. 228–235, 2009.
10. K. Sudhakar, "Study of Heart Disease Prediction using Data Mining," vol. 4, no. 1, pp. 1157–1160, 2014.
11. R. Chitra and V. Seenivasagam, "REVIEW OF HEART DISEASE PREDICTION SYSTEM USING DATA MINING AND HYBRID INTELLIGENT TECHNIQUES," *Journal on Soft Computing (ICTACT)*, vol. 3, no. 4, pp. 605–609, 2013.
12. N. A. Sundar, P. P. Latha, and M. R. Chandra, "PERFORMANCE ANALYSIS OF CLASSIFICATION DATA MINING TECHNIQUES OVER HEART DISEASE DATA BASE," *International Journal of Engineering Science & Advanced Technology*, vol. 2, no. 3, pp. 470–478, 2012.
13. S. A. Pattekari and A. Parveen, "PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES," *International journal of Advanced Computer and Mathematical Sciences*, vol. 3, no. 3, pp. 290–294, 2012.
14. C. Ordonez, "Association rule discovery with the train and test approach for heart disease prediction.," *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 10, no. 2, pp. 334–43, Apr. 2006.
15. Y. Xing, J. Wang, Z. Zhao, and A. Gao, "Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease," in *2007 International Conference on Convergence Information Technology (ICCIT 2007)*, 2007, pp. 868–872.
16. K. Srinivas, K. Raghavendra Kao, and A. Govardham, Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," in *The 5th International Conference on Computer Science & Education*, 2010, pp. 1344–1349.
17. J. Liu, Y.-T.HSU, and C.-L. Hung, "Development of Evolutionary Data Mining Algorithms and their Applications to Cardiac Disease Diagnosis," in *WCCI 2012 IEEE World Congress on Computational Intelligence*, 2012, pp. 10–15.
18. P. Chandra, M. .Jabbar, and B. .Deekshatulu, "Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection," in *12th International Conference on Intelligent Systems Design and Applications (ISDA)*, 2012, pp. 628–634.
19. S. U. Amin, K. Agarwal, and R. Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors," in *Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013)*, 2013, no. Ict, pp. 1227–1231.
20. A. AZIZ, N. ISMAIL, and F. AHMAD, "MINING STUDENTS'ACADEMIC PERFORMANCE.," *Journal of Theoretical & Applied Information Technology*, vol. 53, no. 3, 2013.
21. B.Venkatalakshmi, M.V.Shivashankar, "Heart Disease Diagnosis Using Predictive Data Mining", IJIRSET, Vol 3, March 2014, ISSN : 2319-8753.
22. Hlaudi Daniel Masethe, Mosima Anna Masethe, "Prediction of Heart Disease using Classification Algorithms", Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, 2014, San Francisco, USA.
23. Ms.RupaliR.Patil, "Heart Disease Prediction System using Naïve Bayes and Jelinek-mercer smoothing",
24. Data Mining Techniques ByArun K Pujari universities press Third edition
25. Soft computing by S.N.sivanandan, S.N.Deepa Wiley publications