e-ISSN (O): 2348-4470

Scientific Journal of Impact Factor (SJIF): 4.72

International Journal of Advance Engineering and Research Development

Emerging Trends and Innovations in Electronics and Communication Engineering - ETIECE-2017

Volume 5, Special Issue 01, Jan.-2018 (UGC Approved)

Prediction of Heart Diseases Using Decision Tree and Neural Network Data Mining Techniques-A Review.

[1]Mujtaba Ashraf Qureshi , [2]Dr, Irshad Ahmad Mir , [3]Tawseef Ahmad,[4]Mohd Iqbal Sheikh.

[1]Ph.D Scholar C.S.Deptt.Mewar University , Rajasthan India.
[2]Asst. Professor ,C.S.Deptt. Collage for Women ,Srinagar J&K India.
[3]Asst. Professor ,C.S. Deptt. Kupwara Degree Collage, J&K India.
[4]Ph.D Scholar C.S.Deptt.Mewar University , Rajasthan India.

ABSTRACT-*Data mining is the automated process to analyze huge medical databases and then extracting the useful patterns. A data mining tool gives answer to complex queries in an accurate manner and takes less time to resolve. The huge amounts of medical data related to heart diseases are stored in healthcare databases are too complex and large to be handled by conventional methods. With the help of data mining techniques it is possible to transform these mounds of data into functional information so that it can be used for heart disease diagnosis. Data mining techniques plays a vital role for the prediction of heart diseases in an efficient and accurate manner with lesser amount of time. In this paper we present survey of different data mining techniques and in particular Decision tree and neural networks used for the prediction of heart diseases. Results are shown how Decision Tree and Neural Networks using various parameters perform.*

Keywords - *Heart disease, Data mining, Decision tree, Neural Networks.*

I. INTRODUCTION.

Data mining mainly focus on the extraction of valuable information from huge databases in the most accurate and efficient ways. A widely accepted formal definition of data mining is given subsequently, "Data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data" [18][1]. In short, it is a process of analyzing data from different perspective and gathering the knowledge from it. In healthcare industry the available raw medical data is mixed, huge and distributed over various sites but lacks technology which can be used for mining and utilize useful patterns of data for diseases prediction. Over the years various data mining techniques have been developed for mining useful data from huge medical databases and thus to predict diseases efficiently. In this study we present an overview and comparison of various data mining techniques developed to predict heart diseases. As per our study it is observed that Decision Tree and Neural Network show promising accuracy to diagnose heart defects over other techniques.

The term "heart disease" includes a broad range of conditions that affect the heart and the blood vessels and the method in which blood is pumped and disseminated through the body. About 80% deaths occur worldwide because of heart diseases. WHO estimated by 2030, almost 23.6 million people will die due to heart disease as written in *[17]*[2]. Various risk factors responsible for heart diseases are smoking, high cholesterol, poor diet, high blood pressure, obesity, physical inactivity and so on. There are various types of heart diseases that include coronary heart disease, angina pectoris, cardiomyopathy, arrhythmias, myocarditis, and congestive heart failure.
Section II presents the literature survey of different research works used to predict heart diseases .In section III we present neural network and its operation in the heart disease prediction system. In section IV we discuss the working of the decision trees. Then we identify the most used algorithms for heart diseases diagnosis in section V and finally we show the conclusion in section VI and references of our work in section VII.

II. REVIEW OF LITERATURE.

Over the years, various research work have been done related to heart disease prediction system using diverse data mining techniques by different authors. Most of the researchers frequently used Decision Tree and Neural Network data mining techniques to predict cardiovascular diseases and shows promising results. So this paper aims at analyzing

[1] Frawley and Piatetsky-Shapiro, *"Knowledge Discovery in Databases: An Overview."*, The AAAI/MIT Press, Menlo Park, C.A, 1996.

[2] Nidhi Bhatla etl *"An Analysis of Heart Disease Prediction using Different Data Mining Techniques"*, IJERT, ISSN: 2278-0181,Vol. 1 Issue 8, October – 2012.

Decision Tree and Neural Network data mining techniques that have been introduced in recent years for heart disease prediction system by different researchers.

In [20], an Intelligent web-based, user-friendly and reliable Heart Disease Prediction System is (IHDPS) built with the aid of data mining techniques like Decision Trees, Naïve Bayes and Neural Network. IHDPS gives answer to complex queries which traditional decision support systems cannot. It predicts efficiently whether there are chances of heart attack or not by using attributes such as age, sex, blood pressure and blood sugar as input data. It is implemented on the .NET platform and uses 15 attributes to perform experimental work. This research work shows Decision Trees, Naïve Bayes and Neural Network with accuracies 94.93%, 95% and 93.54% respectively. The results illustrated the peculiar strength of each of the methodologies in comprehending the objectives of the specified mining objectives.

In [15], comparison of different data mining techniques is calculated using 13 attributes for the prediction of heart diseases. Models developed and validated by using five algorithms including C5.0 decision tree, Neural Network, Support Vector Machine (SVM), K-Nearest Neighborhood (KNN) and Logistic Regression. The accuracy of models developed by C5.0 Decision Tree, Neural Network, Support Vector Machine (SVM), K-Nearest Neighborhood (KNN) is 93.02%, 80.23%, 86.05%, 88.37% respectively. The results shown by Decision Tree are simple to interpret and easy to use by medical professionals.

In [11] four data mining algorithms namely j48 decision tree, Naive Bayes, KNN and SMO are analyzed and compared on heart disease dataset performed on weka simulated platform. After comparison of j48 decision tree, Naive Bayes, KNN and SMO accuracy achieved is 83.73%, 81.81%, 82.775% and82.775% respectively.

In [19] researchers reduced number of attributes to only six attributes for heart diseases prediction by using Genetic algorithm. By means of reduction in the number of medical attributes more accuracy is achieved in this work to predict heart diseases. Accuracy achieved by Decision Tree, Naïve Bayes and Classification Clustering is 99.2%, 96.5% and 88.3% respectively.

Chitra R.et al. [1] developed the computer based heart disease analysis system that helps the medical professionals as a tool to predict heart diseases. It is analyzed that neural network with offline training is good for disease prediction in premature period and good presentation can be obtained by pre-processed and normalized dataset.

Nidhi Bhatla et al. [2] performed work to study heart disease prediction using data mining techniques for developing cardiovascular prediction system. Our examination presents that neural network shows highest accuracy using 15 medical parameters than other techniques.

Shadab et al. [3] used 15 attributes and applied Naive Bayes technique to predict cardiovascular diseases.

In [4] more two additional attributes like obesity and smoking are used other than frequently used 13 attributes such as sex, blood pressure, cholesterol and so on for the prediction of cardiovascular diseases. This work is simulated on WEKA 3.6.6 tool. Decision trees, Naïve Bayes and neural Networks techniques are analyzed for heart disease prediction. Results of applied techniques are compared and achieved accuracy of Neural Networks, Decision Trees, and Naive Bayes is 100%, 99.62%, and 90.74% respectively. From analysis it shows Neural Networks outperformed other two techniques in heart disease prediction accuracy.

In [5], performed work, "A Novel Approach for Heart Disease Diagnosis using Data Mining and Fuzzy Logic". In this research mostly used number of attributes of heart diseases is reduced to 4 to decrease number of clinical tests to be performed by a patient. The effectiveness of the projected system is also developed to predict coronary diseases more accurately. Accuracy of Decision Tree and Naive Bayes achieved is 100% and 100% respectively using 4 attributes each. This work shows that Decision Tree and Naive Bayes using fuzzy logic have presented good results over other techniques.

Vanisree K et al. [6], proposed a Decision Support System for diagnosis of Congenital Heart Disease. The proposed system is designed and developed by using MATLAB's GUI feature with the execution of Backpropagation Neural Network. The Backpropagation Neural Network used in this study is a multi layered Feed Forward Neural Network, which is trained by a supervised Delta Learning Rule. The dataset used in this study are the signs, indications and the results of physical evaluation of a patient. The proposed system achieved an accuracy of 90%.

The researchers in [7] used decision trees, naïve bayes, neural networks, association classification and genetic algorithm data mining techniques for diagnosing and analyzing heart disease from the datasets. An experiment performed by *[8]* the researchers on a dataset produced a model using neural networks and hybrid intelligent algorithm, and the results shows that the hybrid intelligent technique enhanced accuracy of the prediction.

Latha Parthiban et al. [9] projected an approach on the basis of coactive neuro-fuzzy inference system (CANFIS) for prediction of heart disease. The CANFIS model uses neural network capabilities with the fuzzy logic and genetic algorithm.

In [10] six parameters are used to predict heart diseases and shows more accurate and efficient results. In this work three classifiers are used like Naive Bayes, Classification by clustering and Decision Tree to diagnosis of heart patients and achieved accuracy is 96.5%, 88.3% and 99.2% respectively. Decision tree shows promising results than others.This work is simulated on weka 3.6.0 tool.

In [12], a methodology is introduced which uses SAS base software 9.1.3 and 13 attributes were used for diagnosing of the heart disease. SAS base software is an intelligent integrated platform allows the user to estimate their system performance from many different points of views. A neural networks ensemble model is developed by combining three independent neural networks models. The number of neural networks node in the ensemble model was also increased but

no performance improvement was obtained. The experimental results gained 89.01% classification accuracy, 80.95% sensitivity and 95.91% specificity values for heart disease diagnosis.

III. NEURAL NETWORK AND ITS OPERATION.

An artificial neural network is a mathematical model or computational model also called as neural network. This network is based on biological neural networks. Artificial neural network is based on observation of a human brain [14][3].Brain of human beings consists of complex network of neurons.Fig.3a shows the basic structure of biological neuron.

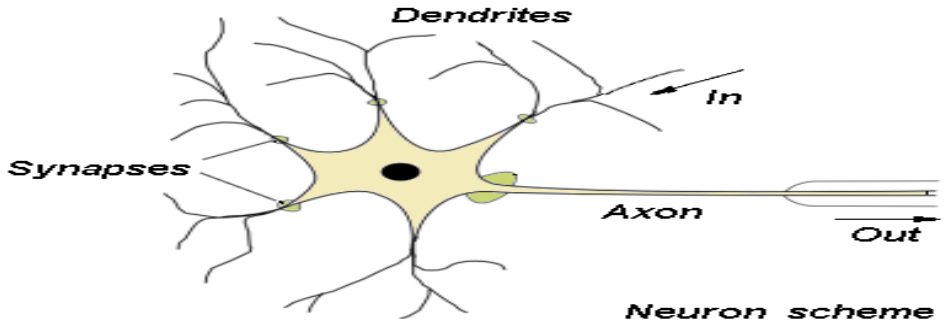


Fig.3a- Biological neuron

Analogically ANN consists of three layers namely input, hidden and output layer. The parameters that are entered as input to the next form a first layer. In medical diagnosis patients risk factors are treated as input to the artificial neural network. So to predict heart diseases risk factors such as high blood pressure, high cholesterol level, gender, chest pain type and so on of patients acts as input data to neural networks. As shown in research work [4][4], [12][5], [16][6], [15][7] and [20][8] respectively how neural network produces promising results in the diagnosis of heart diseases as below mentioned number of parameters is entered in the system, mentioned in Table.3b.

(Table.3b)

Technique	***No. of Attributes***	***Accuracy***
Neural Network	13	100%
	13	97.47%
	19	73.9%
	13	93.02%
	15	93.54%

[3] Nang Y,"The Hand book of data mining", Lawrence Erlbaum associates (2003).

[4] Chaitrali S. Dangare etl "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", IJCA, Vol. 47, June 2012 (pp 44-48).

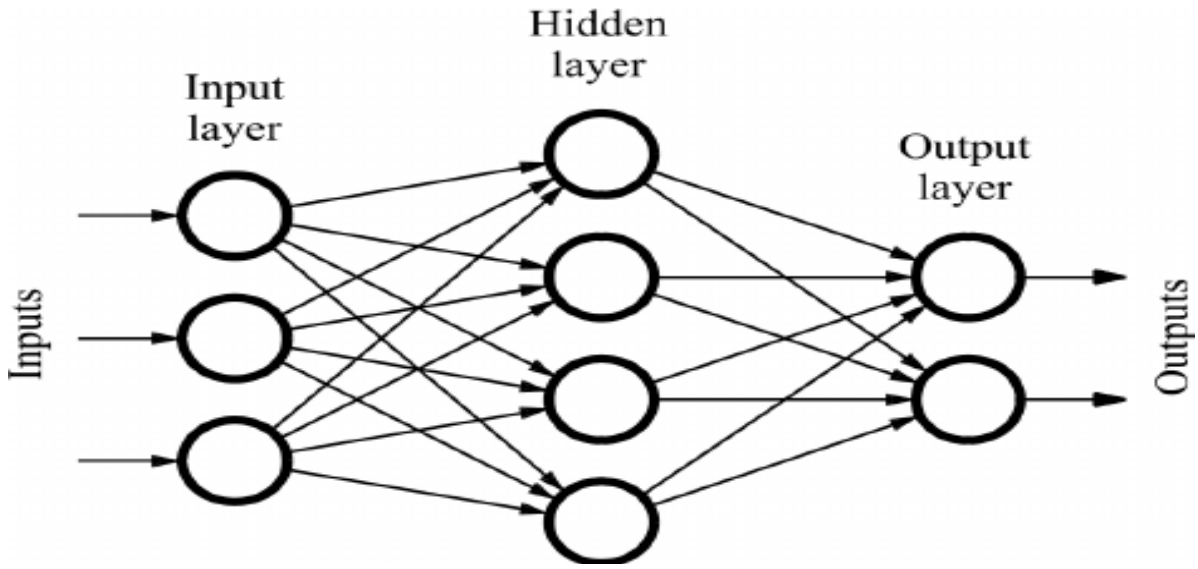
[5] Nidhi Bhatla etl"A Novel Approach for Heart Disease Diagnosis using Data Mining and Fuzzy Logic",IJCA, Volume 54– No.17, September 2012.

[6] M.G. Tsipouras etel,*"Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling",* IEEE transactions on information technology in biomedicine, July 2008.

[7] Moloud Abdar etl *"Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases."* Vol. 5, December 2015.

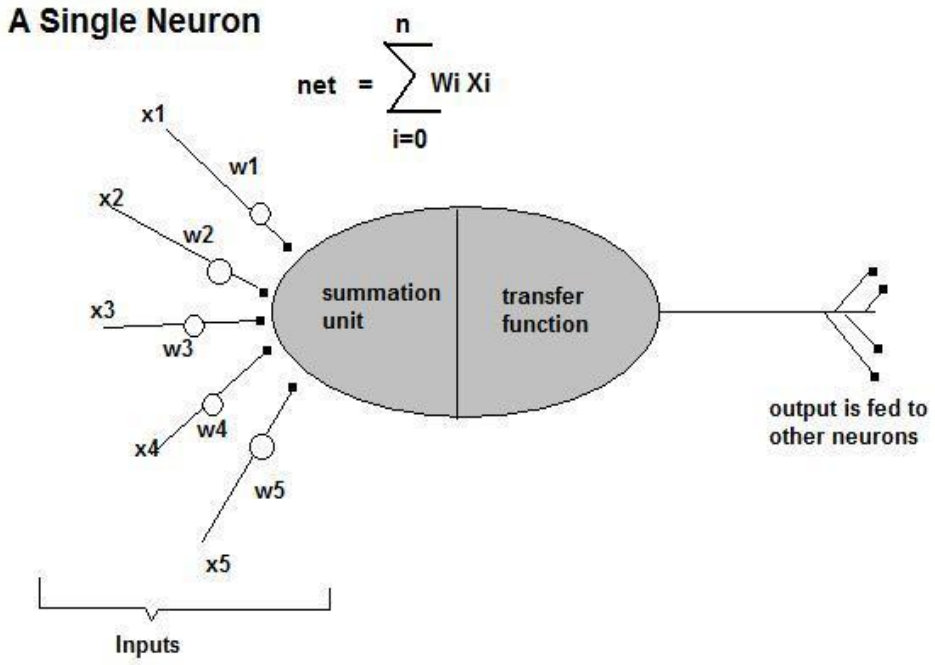
[8]Sellappan Palaniappan etal *"Intelligent Heart Disease Prediction System Using Data Mining Techniques*", IJCSNS Vol.8 No.8, August 2008.

Fig.3c depicts diagram of ANN and shows its three working layers.



(fig.3c)

Generally input layer of ANN collects information presented from the surroundings. In other words input signals are combined usually by simple summation and are passed to the next layer through joined paths existing between neurons. PEs in any one layers are joined with all PEs in the layer above. The output layer generates a response to a given output. The layer between input and output layers is called the hidden layer. Neural networks are trained by experience, when applied an unknown input to the network it can generalize from past experiences and product a new result. Artificial neuron model given by Hanbay.et al., 2008, is shown in fig.3d.



(fig.3d)

The output of the neuron net is given by the following equation:

$$net = \sum_{i=0}^{n} w_i x_i$$

where X_I = x1,x2,x3,x4 and x5 represent the input applied to the neuron, Wi = w1,w2,w3,w4 and w5 represent the weights for each input Xi. Summation of this input value is calculated by the multiplication of wi and xi and required value stores in variable net as given above. If obtained value of any input is below threshold value that time signal is stopped. However if obtained value is greater than threshold value signal is passed to next neuron.

IV. DECISION TREE AND ITS OPERATION

Decision trees are introduced in 1960's as an efficient data mining technique; they have been widely used in medical science particularly to predict cardiovascular diseases. Because this technique is easy to be used, free of ambiguity, and robust even in the presence of missing values. Decision tree method is a frequently used data mining process for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable in heart diseases diagnosis system. This technique classifies a population into branched segments that construct an upturned tree like structure with a root node, internal nodes, and leaf nodes. The algorithm is non-parametric and can competently deal with large and complex heart disease datasets. When the sample size is large enough, datasets can be divided into training and validation datasets. Using the training dataset to build a decision tree model and a validation dataset to decide on the appropriate tree size needed to achieve the optimal final model. Because of its robustness, simplicity and to tackle large datasets with great ease medicinal scientists started to use decision tree technique frequently to predict heart diseases and proved successful technique. Following table (Table.4a) depicts use of decision tree by the researchers as in [5][9], [10][10], [4][11], [15][12] and [20][13] respectively for the prediction of heart diseases and obtained higher accuracy and efficiency.

(Table.4a)

Technique	***No. of Attributes***	***Accuracy***
Decision Tree	04	100%
	06	99.2%
	15	99.62%
	13	93.02%
	15	94.93%

[9] Nidhi Bhatla etal *"A Novel Approach for Heart Disease Diagnosis using Data Mining and Fuzzy Logic",* International Journal of Computer Applications, Volume 54– No.17, September 2012.

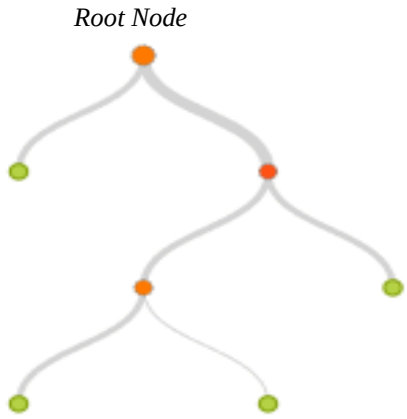
[10] M.Anbarasis etal *"Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm",* IJEST ,Vol. 2(10), 2010.

[11] Chaitrali S. Dangare etl "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", IJCA, Vol. 47, June 2012 (pp 44-48).

[12] Moloud Abdar etl *"Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases."* Vol. 5, December 2015.

[13] Sellappan Palaniappan etal *"Intelligent Heart Disease Prediction System Using Data Mining Techniques*", IJCSNS Vol.8 No.8, August 2008.

Common usages of decision tree models include selection of variables, obtain importance of variables, dealing of missing data values, diagnose and manipulation of data.
Decision tree technique consists of nodes and branches. Nodes are further divided into root node or decision node, internal nodes and leaf nodes. Root node, provide alternative options that will result in the subdivision of all records into two or more independent branches of parent dataset which are mutually exclusive. Internal nodes or chance nodes, represent one of the possible choices available at that point in the tree structure; the top edge of the node is linked to its parent node and the base edge is connected to its child nodes or leaf nodes. Leaf nodes, also called end nodes, presents the final result.Fig.4b shows diagram of decision tree.



(fig.4b)

The most important steps in building a model are splitting, stopping, and pruning. *Splitting* is the first step in decision tree technique. In this step input variables associated to the target variable are used to split parent nodes into pure child nodes of the target variable. Here both discrete as well as continuous medical variables are allowed. During model formation important input variables are identified and then split records at the root node and at following internal nodes into two or more classes based on the rank of these variables. Until pre-determined homogeneity or stopping criteria aren't met, the splitting remains in progress. To prevent decision tree model from over complex, stopping rules are applied. Common parameters used in stopping rules include the minimum number of records in a leaf, the minimum number of records in a node prior to splitting and the number of steps. *Stopping* rules must be applied when building a decision tree to prevent the model from becoming overly complex. *Pruning*. In some conditions it is allowed to increase the size of decision tree to maximum. After that optimal size is obtained by applying pruning to large decision tree by removing less information nodes. Some common methods to select the best possible sub-tree from several candidates is to consider the proportion of records with error prediction or to select best one branch is to use a validation dataset or cross validation ,for small samples is used.

V. PERFORMANCE STUDY.

The table.5a given below present the results for the prediction of heart diseases obtained as per survey of various research works by applying decision tree and neural network. Table 5a part-1 shows number of parameters used and accuracy achieved by applying decision tree data mining technique in [5], [10], [4], [15] and [20] respectively to predict cardiovascular diseases. Table 5a part-2 also presents number of parameters used and accuracy achieved by applying neural network data mining technique in [4], [12], [16], [15] and [20] respectively to predict cardiovascular diseases. Comparison is made among these classification algorithms i.e., decision tree and neural network out of which the decision tree algorithm is considered as the better performance algorithm. It is also observed that Decision tree algorithm

takes lesser number of attributes and presents higher accuracy mostly as compared to neural networks and other techniques.

(Table.5a)

Part No.	***Technique***	***No. of Attributes***	***Accuracy***
1	Decision Tree	04	100%
		06	99.2%
		15	99.62%
		13	93.02%
		15	94.93%
2	Neural Network	13	100%
		13	97.47%
		19	73.9%
		13	93.02%
		15	93.54%

VI. CONCLUSION.

The objective of this paper was to study various data mining techniques predominantly decision tree and neural network that can be employed in computerized cardiovascular prediction systems. Working mechanism of decision tree and neural network are given in detail as medical attributes are entered to predict heart diseases. The analysis shows that various data mining techniques have been used by various researchers. Two techniques were studied and compared in this work and the analysis shows that decision tree shows more accuracy and uses fewer numbers of attributes than neural network. It is also analyzed that to employ decision tree for prediction is very easy as compared to neural network.

VII. REFERENCES.

[1] Chitra R and Seenivasagam V, *"Review of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques",* ISSN: 2229-6956(online) ICTACT Journal on Soft Computing, July 2013, volume: 03, ISSUE: 04, 2013.

[2] Nidhi Bhatla and Kiran Jyoti, *"An Analysis of Heart Disease Prediction using Different Data Mining Techniques",* International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181,Vol. 1 Issue 8, October – 2012.

[3] Shadab Adam Pattekari and Asma Parveen, *"prediction system for heart disease using naïve bayes"*, International Journal of Advanced Computer and Mathematical Sciences, 2012.

[4] Chaitrali S. Dangare, Sulabha S. Apte , *"Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques",* International Journal of Computer Applications (IJCA) (0975 – 8887), Vol. 47, No. 10, June 2012 (pp 44-48).

[5] Nidhi Bhatla, Kiran Jyoti, *"A Novel Approach for Heart Disease Diagnosis using Data Mining and Fuzzy Logic",* International Journal of Computer Applications, Volume 54– No.17, (pp 16-21), September 2012, ISSN 0975 – 8887.

[6]Vanisree K, Jyothi Singaraju, *"Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks*", International Journal of Computer Applications (0975 – 8887) Volume 19– No.6, April 2011.

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

[8] 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.

[9] 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.

[10] M.Anbarasis,E.Anupriya,N.CH.S.N.Iyengar**,***"Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm",* International Journal of Engineering Science and Technology ,Vol. 2(10), 2010, 5370-5376.

[11]Boshra Bahrami, Mirsaeid Hosseini Shirvani,*"Prediction and Diagnosis of Heart Disease by Data Mining Techniques",* Journal of Multidisciplinary Engineering Science and Technology (JMEST) ISSN: 3159-0040 Vol. 2 Issue 2, February – 2015.

[12] Resul Das, Ibrahim Turkoglu , Abdulkadir Sengur,*"Effective diagnosis of heart disease through neural networks ensembles."* Expert Systems with Applications 36 (2009) 7675–7680.

[13] Shamsher Bahadur Patel , Pramod Kumar Yadav , Dr. D. P.Shukla, *"Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques",* IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS) e-ISSN: 2319-2380, p-ISSN: 2319-2372.Volume 4, Issue 2 (Jul. - Aug. 2013), PP 61-64.

[14] Nang Y,*"The Hand book of data mining",* Lawrence Erlbaum associates (2003).

[15] Moloud Abdar, Sharareh R. Niakan Kalhori, Tole Sutikno, Imam Much Ibnu Subroto, Goli Arji, *"Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases."* Vol. 5, No. 6, December 2015.

[16] Markos G. Tsipouras, Themis P. Exarchos, Dimitrios I. Fotiadis, Anna P. Kotsia, Konstantinos V. Vakalis,Katerina K. Naka, and Lampros K. Michalis,*"Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling",* IEEE transactions on information technology in biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society ,July 2008.

[17] Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, *"A data mining approach for prediction of heart disease using neural networks*", international journal of computer engineering and technology, 2012.

[18] Frawley and Piatetsky-Shapiro*,"Knowledge Discovery in Databases: An Overview*.", The AAAI/MIT Press, Menlo Park, C.A, 1996.

[19] Shamsher Bahadur Patel , Pramod Kumar Yadav , Dr. D. P.Shukla, "*Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques",* IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS) e-ISSN: 2319-2380, p-ISSN: 2319-2372.Volume 4, Issue 2 (Jul. - Aug. 2013), PP 61-64.

[20] 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.

.

