National Conference on Innovations in Computing and Networks (NCICN) - 2019
© 2019 IJSRSET | Volume 5 | Issue 9 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099

Prediction of Heart Disease and Breast Cancer Using Random Forest (RF) and Multi-Layer Perceptron Neural Network (MLP) Approaches

Dr. Nijil Raj. N*[1], Shabana S*[2]

*[1]Professor & Head, Department of Computer Science and Engineering, Younus College of Engineering and Technology, Kollam, Kerala, India

*[2]B.Tech student, Department of Computer Science and Engineering, Younus College of Engineering and Technology, Kollam, Kerala, , India

ABSTRACT

Disease diagnosis is one of most important application of data mining to proving successful results. Breast Cancer Diagnosis a re two medical applications which became a big challenge to the researchers. The use of machine learning and data mining techniques has changed the whole process of breast cancer Diagnosis. Most data mining methods which are commonly used in this domain are considered as classification category and applied prediction techniques assign patients to either a" benign" group that is non- cancerous or a" malignant" group that is cancerous he project focuses on the prediction of various diseases like heart disease and breast cancer that can assist medical professionals in predicting disease status based on the clinical data of patients. In existing method reveals that 91% accuracy by using random forest approach in breast cancer and heart disease datasets. In our proposed method reveal that 94.4% and 85% accuracy in random forest, and multi-layer perceptron approach reveals that 95.8% and 86.7% respectively in breast cancer and heart disease datasets .Its seems to be that our proposed methods are better than the existing method.

Keywords : Random forest, Multi-layer perceptron, Machine Learning

I. INTRODUCTION

Heart disease also called as coronary artery disease is a condition that affects the heart. Heart disease is a leading cause of death worldwide. Physicians generally make decisions by evaluating current test results of the patients. Previous decisions taken by other patients with the same conditions are also examined. So diagnosing heart disease requires experience and highly skilled physicians. Heart disease will become a leading cause of death by 2020. Heart disease diagnosis is an important yet complicated task. Today many hospitals collect patient data to manage health care of patients. This information is in different format like numbers, charts, text and images. But this database contains rich information but poorly used for clinical decision making.

The main cause of breast cancer is when a single cell or group of cells escapes from the usual controls, that regulate cellular growth and begins to multiply and spread. This activity may result in a mass, tumor or neoplasm. Many masses are benign that means the abnormal growth is mainly restricted to a circumscribed, single and expanding mass of cells. Medical data mining has great potential for exploring predictions in this research work. The predictions are made using the classification model that is built from the classification algorithms when the heart disease dataset is used for training. This

final model can be used for prediction of any types of heart diseases and breast cancer.

II. RELATED WORKS

Smaranda Belciug [2] proposed a two stage model containing severaldifferentneuralnetworks : multi-layerneural perceptron(MLP), radial basis function(RBF)and Probabilistic Neural Networks (PNN), and the effectiveness of this system on a real breast cancer database, to support the medical decision process The classification results were consistent with some of the highest results obtained by using sophisticated and expensive imaging medical techniques and finally get 85% accuracy.

Gouda I. Salama et.al [6] proposed a comparison among the different classifiers decisiontree (J48),Multi-Layer Perception (MLP), Naive Bayes (NB), Sequential Minimal Optimization (SMO), and Instance Based for K-Nearest neighbor (IBK) on three different databases of breast cancer (Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC)) by using classification accuracy 75% and confusion matrix based on 10-fold cross validation method

Sanjay Kumar Sen proposed[3] "Predicting and Diagnosing of Heart Disease Using Machine Learning" In this paper, they carried out an experiment to find the predictive performance of different classifiers. select four popular classifiers considering their qualitative performance for the experiment. We also choose one dataset from heart available at UCI machine learning repository.Naïve base classifier is the best in performance.Finally get an accuracy of 77%.

Mai shouman et.al [4] proposed a decision tree for diagnosing heart disease patients. Different types of decision trees are usedforclassification. Theresearchinvolvesdata discretization, decision tree selection and reduced error pruning. Their method outperforms bagging and j48 decision tree. Their approach achieved 79.1% accuracy.

P.K.Anooj [2] developed a clinical decision support system to predict heart disease using fuzzy weighted approach. The method consists of two phases. First phase consists of generation of weighted fuzzy rules, and in second phase fuzzy rule based decision support system is developed. Author used attribute selection and attribute weight method to generate fuzzy weighted rules. Experiments were carried out on UCI repository and obtained accuracy of 57.85%

Robert Detrano et.al [3] proposed probability algorithm for the diagnosis of coronary artery disease. The probabilities that resulted from the application of the Cleveland algorithm were compared with Bayesian algorithm. Their method obtained an accuracy of 77% .

Jabbar et.al [7] proposed a decision tree for early diagnosis of heart disease. Alternating decision tree is a new type of classification, which is a generalization of decision tree, voted decision trees and voted decision stumps. Principal component analysis is used as a feature selection measure and used to select best features .Heart disease data consists of 96 patient‟s records with 10 features. Their proposed approach achieved an accuracy of 91.66%.

My chau Tu et.al [5] proposed diagnosis of heart disease through bagging approach. Proposed bagging algorithm is used to identify warning signs of heart disease. They made a comparison with decision tree. Their approach claimed an accuracy of 81.4%.

Resul das et.al[1] proposed method creates a new model by combining posterior probabilities from multiple predecessor models. They implemented the method with SAS base software on Cleveland heart disease data set and obtained 89.01% accuracy

Table 1 : comparison of related work

Author	Algorithm	Accuracy	Year
Smaranda Belciug	Multi layer perceptron	85%	2015
Sanjay Kumar sen	1. Naive base 2.Decision tree 3. SVM	83.49% 77.55% 84.15%	2017
Gouda I salama	MLP	75%	2013

III. METHODS AND METHODOLOGY

A. DATA SETS

The Stat Log dataset from UCI machine learning repository is utilized for making heart disease and breast cancer prediction in this research work. Two sets of data"s are used for heart diseases and breast cancer.

S1 - Heart diseases
S2 – Breast cancer

1. Heart diseases (S1)

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous. These data need to be collected in an organized form. This collected data can be then integrated to form a hospital information system. Data mining technology provides a user oriented approach to novel and hidden patterns in the data. Theterm Heart disease encompasses the diverse diseases that affect the heart. Heart disease was the major cause of casualties in the different countries including India. Heart disease is a term covering any disorder of the heart. Unlike cardiovascular disease, which describes problems with the blood vessels and circulatory system as well as the heart, heart disease refers to issues and deformities in the heart itself .Heart disease kills one person every 34 seconds in the United States.

In this research work, the supervised machine learning concept is utilized for making the predictions. A comparative analysis of the three data mining classification algorithms namely Decision Tree and Logistic Regression are used to make predictions. The Stat Log dataset from UCI machine learning repository is utilized for making heart disease refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0). One file has been "processed", that one containing the Cleveland database. All four unprocessed files also exist in this directory.

1.1 Attribute information

Only 14 attributes used:
1. age
2. sex
3. Cp = chest pain type

4. trestbps - resting blood pressure
5. Chol- serum cholestoral in mg/dl
6. Fbs- (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg - resting electrocardiographic results
8. thalach - maximum heart rate achieved
9. Exang- exercise induced angina (1 = yes; 0 = no)
10. Oldpeak- ST depression induced by exercise relative to rest
11. Slope- the slope of the peak exercise ST segment
12. Ca- number of major vessels (0-3) colored by flourosopy
13. Thal- : 3 = normal; 6 = fixed defect; 7 = reversable defect
14. num- diagnosis of heart disease (angiographic disease status)

2. Breast Cancer (S2)

The data set consist of 15 types of attributes

1. Number of instances: 569
2. Number of attributes: 32 (ID, diagnosis, 30 real-valued input features
3. Attribute information
1) ID number
2) Diagnosis (M = malignant, B = benign)
3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the peri)
b) texture (standard deviation of gray-scale values)
c) perimeter d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour) h)concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

Several of the papers listed above contain detailed descriptions of how these features are computed. The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each

image, resulting in 30 features. For instance, field 3 is Mean
Radius, field 13 is Radius SE, and field 23 is Worst Radius. All feature values are recoded with four significant digits.
4. Missing attribute values: none
5. Class distribution: 357 benign, 212 malignant

B. ALGORITHMS USED

1. Classification using random forest

Random forest algorithm is one of the most effective ensemble classification approach. The RF algorithm has been used in prediction and probability estimation.RF consists of many decision trees .Each decision tree gives a vote that indicate the decision about class of the object. Random forest item was first proposed by Tin kam HO of bell labs in 1995.
RF method combines bagging and random selection of features. There are three important tuning parameters in random forest

1) No. of trees (n tree)
2) Minimum node size
3) No. of features employed in splitting each node
4) No. of features employed in splitting each node for each tree (m try).

2. Classification using multi layer perceptron

The most common neural network model is the multi layer perceptron (MLP). This type of neural network is known as a supervised network because it requires a desired output in order to learn. The goal of this type of network is to create a model that correctly maps the input to the output using

historical data so that the model can then be used to produce the output when the desired output is unknown. The MLP and many other neural networks learn using an algorithm called back propagation. With back propagation, the input data is repeatedly presented to the neural network. With each presentation the output of the neural network is compared to the desired output and an error is computed. This error is then fed back (back propagated) to the neural network and used to adjust the weights such that the error decreases with each iteration and the neural model gets closer and closer to producing the desired output. This process is known as training.

The supervised learning problem of the MLP can be solved with the back-propagation algorithm. The algorithm consists of two steps. In the forward pass, the predicted outputs are calculated corresponding to the given inputs. In the backward pass, partial derivatives of the cost function with respect to the different parameters are propagated back through the network. A typical multilayer perceptron (MLP) network consists of a set of source nodes forming the input layer, one or more hidden layers of computation nodes, and an output layer of nodes.

C. PERFORMANCE METRICS

1. Precision

Precision Recall graphs of heart diseases and breast cancer are shown below. In Precision is the part of significant instances between the retrieved instances. The Eq. of precision is given in Eq.(2)

Precision = TP/(TP+FP) (2)

2. Recall

Recall is the small part of appropriate instances that have been retrieved over the total quantity of relevant instances. The Eq. of recall is given in Eq.(3).

Recall = TP/(TP + FN) (3)

3.ROC area

Roc Curves are commonly used to show in a graphical way the connection/ trade off involving clinical sensitivity and specificity for every potential cut off for a test or an arrangement of tests.

4. Accuracy

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same

Accuracy = TP+TN/TP+FP+FN+TN

D. METHODOLOGY

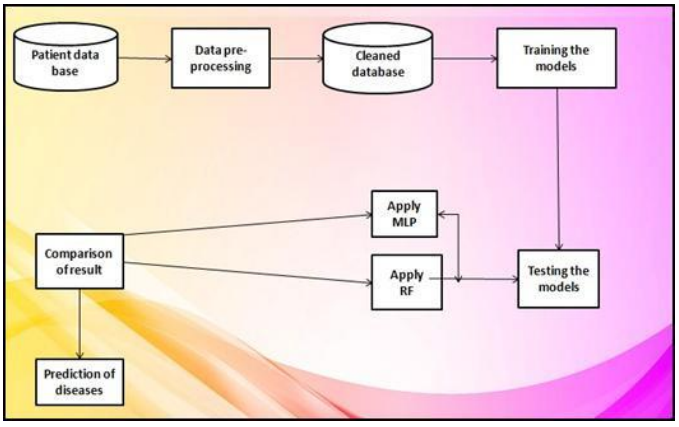

Fig 2: Methodology

The heart disease and breast cancer prediction can be performed by following the procedure which is similar to Fig.1 which specifies the research methodology for building a

classification model required for the prediction of the heart diseases and breast cancer in patients.

The following algorithm refers the actual description of the methodology of the system recall values are shown in table 1&3. Fig 1 shows the entire methodology of the system research work

1) Initialise the processes
2) Collect all the heart diseases and breast cancer database from patients(UCI machine learning repository)
3) The collected data undergoing pre-processing
4) Clean the two database by removing all the missing values and attributes.
5) Train the model
6) Test the model by using the algorithm"s such as logistic
regression and decision tree.
7) Comparising both the training result.
8) Finally predicting the diseases.

The model forms a fundamental procedure for carrying out the heart diseases and breast cancer prediction using any machine learning techniques. Firstly collect all the datasets of heart diseases and breast cancer, and train all the datasets by using classification algorithms such as decision tree and logistic regression. In the third step the training datasets undergoing testing also using decision tree and logistic regression, Finally applying these two algorithms to the sufficient datasets, prediction of result is obtained and accuracy of these algorithms are compared.

IV. RESULTS AND DISCUSSION

Our Method describes the results of both proposed and existing system. Random forest and multilayer preceptor neural network algorithms are used to predict the breast cancer and heart diseases in existing system. By comparing the two systems, proposed system depicts accurate results than existing system.

Data sets	Algorithm	Precision	Recall	Acuracy
S1	RF	0.862	0.844	85%
S1	MLP	0.875	0.862	. 86.7%
S2	RF	0.940	0.940	94.4%
S2	MLP	0.959	0.951	95.8%

Table 2 : proposed and existing system accuracy

The analysis and identification of the best classification algorithm in this research work is done and the results are provided here. Accuracy of proposed and existing precision and recall values are shown in table 3&5. Fig 2 shows the entire methodology of the system.

1. FIGURES AND TABLES
1. S1

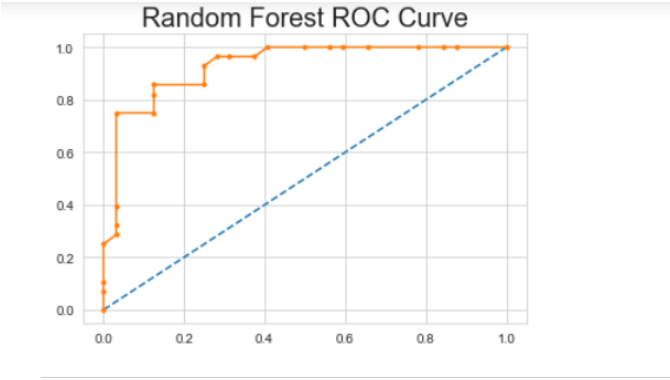

Fig 3: ROC of random forest

A receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by

plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate."

Fig 3 shows the receiver operating characteristics of random forest x axis denotes the false positive rate ,y axis denotes the true positive rate

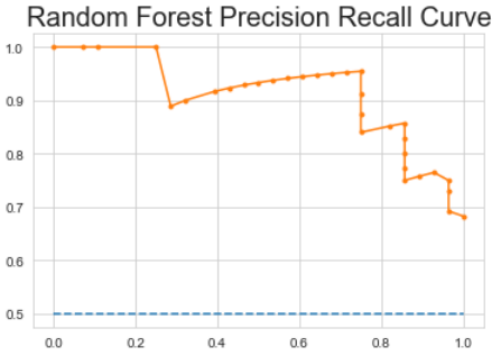

Fig 4: precision recall curve of Random forest

Precision (P) is defined as the number of true positives (Tp) over the number of true positives plus the number of false positives (Fp). Fig 4 shows the precision recall curve of Random forest classifier.

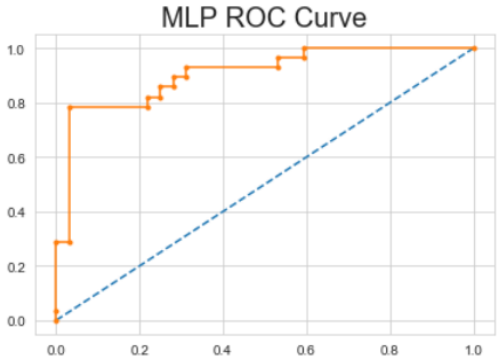

Fig 5: ROC of Multi layer perceptron

Fig 5 plot the receiver operating characterstics multi layer perceptron.

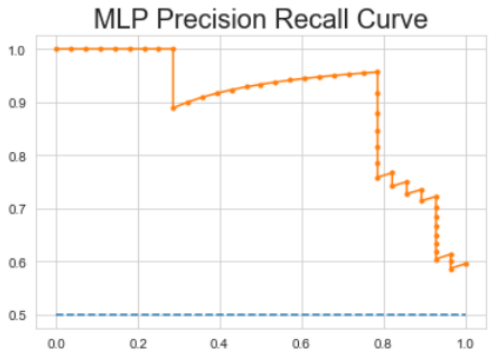

Fig 6: precision recall curve of Multi layer perceptor

Fig 5 shows precision recall curve of heart diseases using multi layer perceptor

1.2 S2

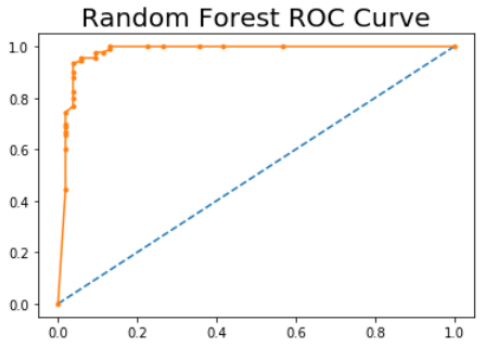

Fig 7 : Roc of random forest

Fig 7 depicts the region operating characterstics of random forest in breast cancer prediction

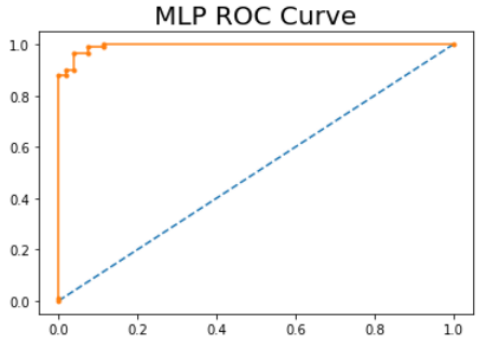

Fig 8 : precision recall curve of random forest

This graph shows the precision recall curve of breast cancer by using random forest classifier. A precision recall curve is a plot of the precision (y-axis) and the recall (x-axis) for different thresholds

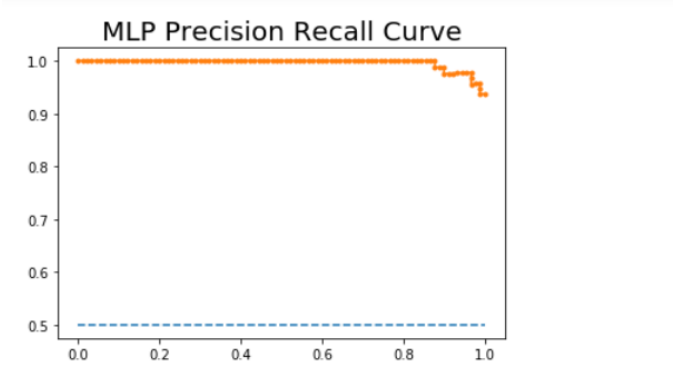

Fig 9 : ROC of Multi layer preceptor

It plots the region operating characterstics of multi layer perceptor classifier used in breast cancer prediction.It is created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate)

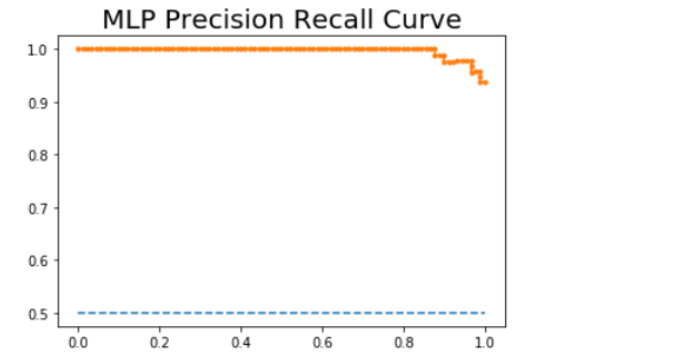

Fig 10 : precision recall of Multi layer preceptor

The graph shows the precision recall curve of multi layer paerceptor in breast cancer prediction. .If you change a binary classifier parameter, it turns out the precision and recall will change. If you graph these points (with precision on the y-axis and recall on the x-axis), you get a precision-recall curve (or equivalently, a precision-recall graph)

1.3 Existing System v/s Proposed System

By analyzing both the system"s propose system gives better result than existing system. Table 3 shows the accuracy"s of both system"s by applying machine learning techniques and also comparison of both the system.

Table 3 : Existing v/s proposed system

Data sets	Algorithm	Proposed system	Existing system
S1	Multi layer perceptron	86.7%	76%
S1	Random forest	85%	Nil
S2	Multi layer perceptron	95.8%	70.72%
S2	**Random forest**	**94.4%**	**Nil**

V. CONCLUSION AND FUTURE WORK

Random forest and Multi layer preceptor Neural Networks are powerful data mining techniques that can be used to classify cancerous tumours and Heart disease. Random forest algorithm creates user-friendly rules that indicates important attributes and requires less computation compared to other algorithms such as Neural Networks. On the other hand. Various data mining techniques are available in medical diagnosis, where the objective of these techniques is to assign patients to either a „healthy" group that does not have a certain disease. Data mining have proved the ability to reduce the number of error rate in decisions. Random forest and MLP are the most popular and effective data mining methods. DT provides a pathway to find "rules" that could be evaluated for separating the input samples into one of several groups without having to express the functional relationship directly. The proposed work can be further enhanced and expanded for the automation of Heart disease prediction. In our feature work we plan to reduce no. of attributes and to determine the attribute which contribute towards the diagnosis of disease.

VI. ACKNOWLEDGMENT

I would like to express my very great appreciation to Prof: SreeRaj Varma for his valuable and

constructive suggestion during the planning and development of this research work. His willingness to give his time so generously has been very much appreciated. I‟m sincerely thankful to our principal Dr. P Sreeraj, for providing me facilities in order to go ahead in the development of my research. I express my deep and sincere gratitude to my guide Dr. Nijil Raj N, Head of department, Computer science and engineering for providing valuable advice and timely instructions.

VII. REFERENCES

[1]. Resul das ,Turkoglu,A Sengur," Effective diagnosis of heart disease through network ensembles", Expert System with Applications36,pp7675-7680(2009)

[2]. PK Anooj," Clinical decision support system: Risk level prediction of heart disease using Weighted fuzzy rules", Journal of king saud university, CIS, 24, PP 27-40(2012)

[3]. Detrano ,Janosi,W Stein burn,et.al," International application of new probability algorithm for the diagnosis of CAD". The American Journal of Cardiology, pp 304-310,64(5),(1989)

[4]. Mai Shouman, Turner, Stocker," Using decision tree for diagnosing heart disease patients", In 9th Australian data mining conference, Australia vol 121,ACM(2011)

[5]. Tu et.al," Effective diagnosis of heart disease through bagging approach" Biomedical Engineering and approach, pp 1-4, BMEI2009, IEEE (2009)

[6]. Alaa Elsayad,Mahmoud Fakhr,"Diagnosis of cardiovasular diseases with bayesian classifier",Journal of Computer Science,vol 11(2),pp274-282(2015)

[7]. Sanjay Kumar Sen." Predicting and Diagnosing of Heart Disease Using Machine Learning" International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 6 Issue 6 June 2017, Page No. 21623-21631Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i6.14 Algorithms

[8]. M.A.Jabbar,B L Deekshatulu,Priti chandra,"Alternating decision tree for early diagnosis of heart disease "IEEE,I4C2014,pp 322-328(2014)

[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, pp. 1-8, 2008

[10]. Bray F, McCarron P, Parkin DM. The changing global patterns of female breast cancer incidence and mortality. Breast Cancer Res 2004;6:229-39

[11]. Sunita Soni , Jyothi Pillai, O.P.Vyas, An Associative Classifier Using Weighted Association Rule , IEEE proceedings of the World Congress on Nature and Biologically Inspired Computing (NaBIC'09), December 09-11, 2009, 1492-1496.

[12]. M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," Expert Syst. Appl., vol. 36, no. 2, pp. 3240– 3247, Mar. 2009.

[13]. Sellappan Palaniappan and Rafiah Awang, "Intelligent Heart Disease Prediction System using Data Mining Techniques", International Journal of Computer Science and Network Security, Vol. 8, No. 8, pp. 1-6, 2008.