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

Extrapolation of Heart Diseases and Breast Cancer using Machine Learning Approaches

Dr. Nijil Raj. N, Aby O Panicker

[*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

A major portion of the world population does not have access to proper healthcare .Heart diseases and Breast cancer is the one of the most important diseases in our day to day life. These diseases is quiet common now a days , different attributes which can relate to heart diseases and breast cancer well to find the better method to predict these disease by machine learning algorithms. Application of machine learning methods in biosciences and health care is presently, more than ever before, vital and indispensable in efforts to transform intelligently all available information into valuable knowledge. In our proposed method Decision Tree approach and Logistic Regression approach are used for predict the Heart diseases and Breast cancer. In existing method reveals that 90.2% and 77.5% accuracy by using decision tree approach in breast cancer and heart disease datasets. In our proposed method reveal that 95% and 83% accuracy in decision tree approach, and Logistic regression approach reveals that 95.8% and 88.3% respectively in breast cancer and heart disease datasets .Its seems to be that our proposed methods are better than the existing method.

Keywords: Decision Tree, Logistic Regression, Machine Learning.

I. INTRODUCTION

In our world there are so many researchers was developed to identified the diagnosis of the diseases, the pattern of diseases is an important part of this. 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 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. The term 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 two 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 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. The area under the ROC curve was used as a measurement of accuracy. Illustrated in Figure 1

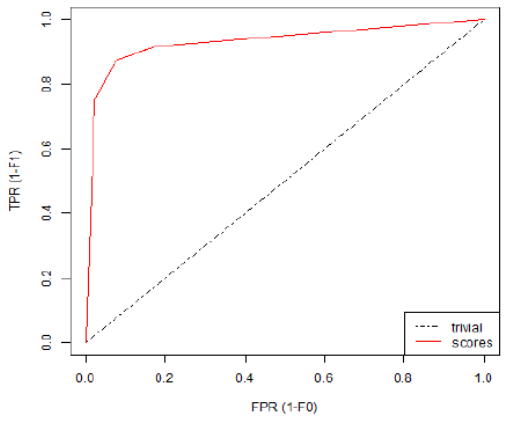

Figure 1. ROC curve

II. RELATED WORKS

During the past few years, various contributions have been made in literature regarding the predicting diseases in precise level. In recent research works, several neural network models have been developed to aid in diagnosis of heart diseases and breast cancer.

Htet Thazin Tike Thein and Khin Mo Mo Tun [1] proposed an approach for breast cancer diagnosis classification using neural network, the network is trained by DE which has been parallelized in order to achieve better performance. This paper presents a result of direct classification of data after replacing missing values using median method for the WBCD dataset by using island differential evolution algorithm. The training algorithms are compared using accuracy and computing time and finally get 94% accuracy and computing time of 16sec.

Puneet Yadav, Rajat Varshney, Vishan Kumar Gupta[2] proposed diagnosis of breast cancer using decision tree models and svm. The goal of this paper is using machine technique to predict benign cancer or the malignant one. Decision trees, Neural Networks, and SVM are powerful data mining techniques tools that can be used to achieve effective results. These algorithms construct their models using training data set then test the obtained models on the test data. In study machine learning algorithms will be tested using breast cancer Wisconsin data set, and then compared to result. By comparing the two algorithms get an accuracy of 90%

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%.

According to H. Benjamin Fredrick David and S. Antony Belcy[4] heart disease can be predicted using data mining techniques. 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 random forest, decision tree and naïve bayes are used to make predictions the analysis is done at several levels of cross validation and several percentage of percentage split evaluation methods respectively. the statlog dataset from uci machine learning repository is utilized for making heart disease 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.

Latha Parthiban and R.Subramanian[5] have proposed a method Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm, The main objective of this research is to develop a prototype Intelligent Heart Disease Prediction System with CANFIS and genetic algorithm using historical heart disease databases to make intelligent clinical decisions which traditional decision support systems cannot. The publicly available Cleveland heart-disease database consists of 303 cases where the disorder is one of four types of heart-disease or its absence.

Shiv Shakti Shrivastava , Anjali Sant , Ramesh Prasad Aharwal proposed"[6] An Overview on Data Mining Approach on Breast Cancer data. This paper gives the current overview of use of data mining techniques on breast cancer data. This paper also gives the study of data mining on medical domain which has already done from researchers. In this paper we use classification data mining techniques on breast cancer data with using data mining software. A huge amount of medical records are stored in databases. Data are produce from different sources and continuously stored in depositories. These databases are more complicated for the point of analysis. Data Mining is a relatively new field of research whose major objective is to acquire knowledge from large amounts of data.

Freddie Bray, Peter McCarron and D Maxwell Parkin[8] proposed a review about "The changing global patterns of female breast cancer incidence and mortality". In this paper we review the descriptive epidemiology of the disease, focusing on some of the key elements of the geographical and temporal variations in incidence and mortality in each region of the world. The review includes published studies and some new analyses using incidence data from population-based cancer registries and mortality data from the WHO databank. We then discuss possible explanations for the results in the light of the changing prevalence of the known aetiological factors, the impact of screening and other preventive strategies, and progress in disease management; we conclude with some comments on future prospects for prevention

Mehmet Fatih Akay[9] proposed an "Support vector machines combined with feature selection for breast cancer diagnosis" In this study, SVM with feature selection was used to diagnose the breast cancer. WBCD taken from the University of California at Irvine (UCI) machine learning repository was used for training and testing experiments . It was observed that the proposed method yielded the highest classification accuracies (98.53%, 99.02%, and 99.51% for 50–50% of training-test partition, 70–30% of training-test partition, and 80–20% of training-test partition, respectively) for a subset that contained five features. Also, other measures such as the confusion matrix, sensitivity, specificity, positive predictive value, negative predictive value and ROC curves were used to show the performance of SVM with feature selection.

Table 1. Comparison of related works

Author	Algorithms	Accuracy	Year
Htet Thazin Tike Thein and Khin Mo Mo Tun	Neural network	94%	2015
Sanjay kumar sen	1.Naive base classifier 2.SVM 3.Decision tree 4.K-Nearest	83.49% 84.15% 77.55% 67.23%	2017
Mehmet Fatih Akay	SVM	80.29%	2009
Puneet Yadav, Rajat Varshney, Vishan Kumar	Decision tree and SVM	90%	2018

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 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, 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 decision tree

Decision Tree (DT) is a simple and easy to implement classifier. The bit through feature to access in depth patients' profiles is only obtainable in Decision Trees. Decision tree builds classification or regression models in the structure of a tree making it simple to debug and handle. Decision trees can handle both categorical and numerical data. The algorithm works by finding the information gain of the attributes and taking out the attributes for splitting the branches in threes. The information gain for the tree is identified using the below given equation

$$E(S) = -P(P)\log 2P(P) - P(N)\log 2P(N) \quad (1)$$

2.Classification using logistic regression

Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

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)

$$\text{Precision} = TP/(TP+FP) \quad (2)$$

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).

$$\text{Recall} = TP/(TP + FN) \quad (3)$$

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

$$\text{Accuracy} = TP+TN/TP+FP+FN+TN$$

D . METHODOLOGY

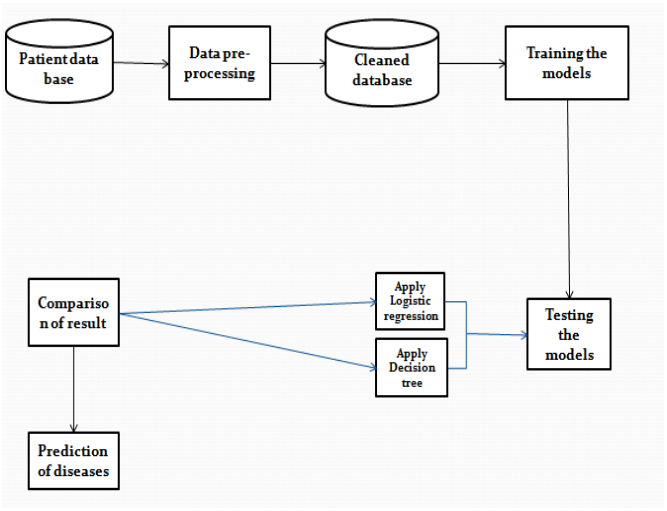

Figure 2. Methodology

The heart disease and breast cancer prediction can be performed by following the procedure which is similar to Figure 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 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. RESULT AND DISCUSSION

This section describes the results of both proposed and existing system. Logistic regression and machine learning algorithms are used to predict the heart diseases and breast cancer in proposed system and some machine learning algorithms are used to predict the breast cancer and heart diseases in existing system. By comparing these two systems(Existing and Proposed systems), proposed system depicts accurate results than existing system. The analysis and identification of the best classification algorithm in this research work is done and the results are provided here. A brief description about the data sets was presented in above section. Accuracy of proposed system, precision and recall values are shown in table 2.

Table 2. Proposed System Accuracy's

Datasets	Algorithms	Precision	Recall	Accuracy
S1	Decision tree	0.835	0.830	83%
S1	Logistic regression	0.897	0.877	88%
S2	Decision tree	0.947	0.950	95.1%
S2	Logistic regression	0.963	0.947	95.8%

1. FIGURES AND TABLES

1.1. S1

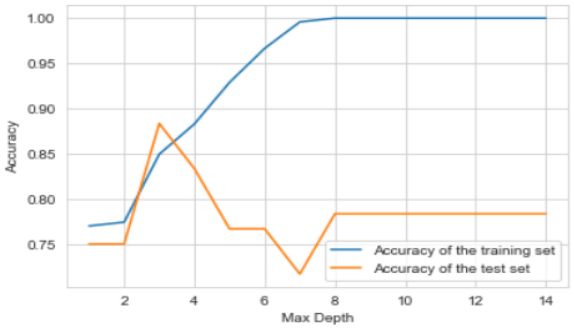

Figure 3 . Accuracy

Figure 3 shows the actual accuracy of heart diseases in both the case of training and testing cases. X axis plot the max depth and Y axis plot the accuracy and get a accuracy above 80%.

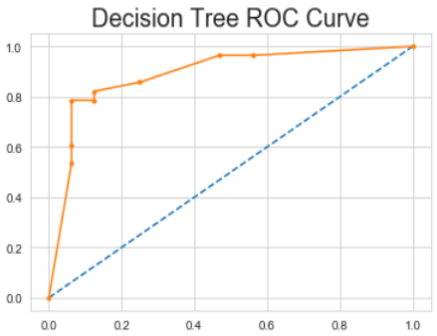

Figure 4. ROC of Decision Tree

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. Fig 4 shows the receiver operating characteristics of decision tree and get a maximum accuracy of 94%. x axis denotes the false positive rate ,y axis denotes the true positive rate

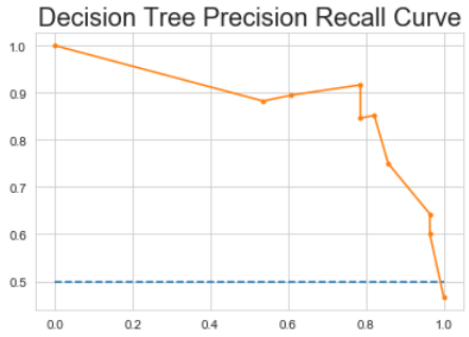

Figure 5. Precision Recall curve of Decision tree

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 5 shows the precision recall curve of decision tree classifier.

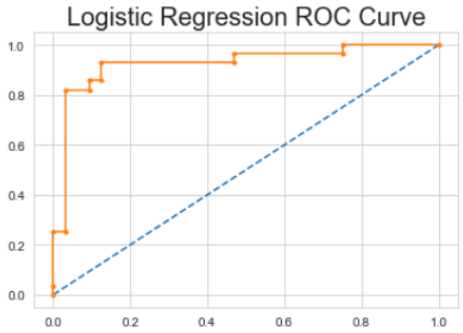

Figure 6. ROC of Logistic Regression

Figure 6 plot the receiver operating characterstics of logistic regression classifier. It shows the average of 94% accuracy.

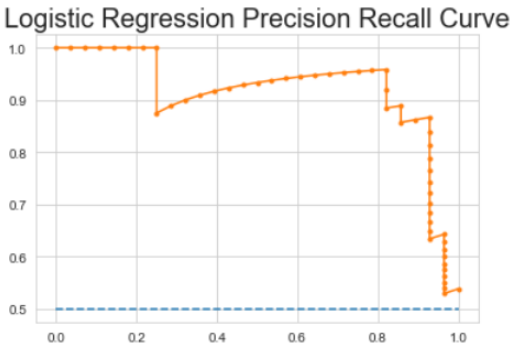

Figure 7. Precision Recall curve of Logistic regression

Figure 7 shows the precision recall accuracy of logistic regression classifier in heart diseases prediction

1.2 . S2

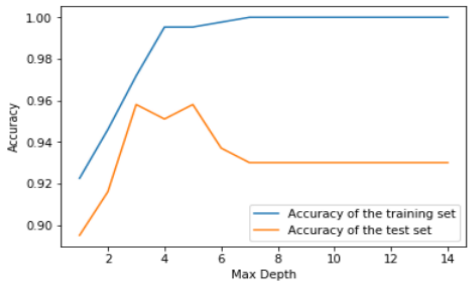

Figure 8. Accuracy

To check the accuracy we need to import confusion_matrix method of metrics class. The confusion matrix is a way of tabulating the number of mis-classifications, i.e., the number of predicted classes which ended up in a wrong classification bin based on the true classes.

Figure 8 shows the actual accuracy of breast cancer above 90%.

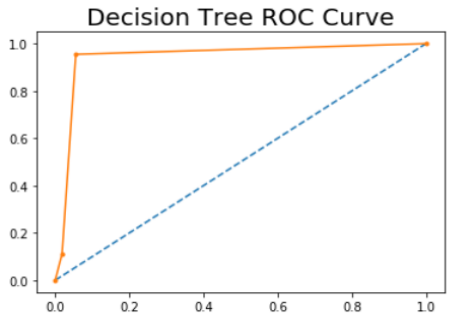

Figure 9. ROC of Decision tree

Figure 9 depicts the region operating characterstics of decision tree in breast cancer prediction and get an accuracy of 94%.

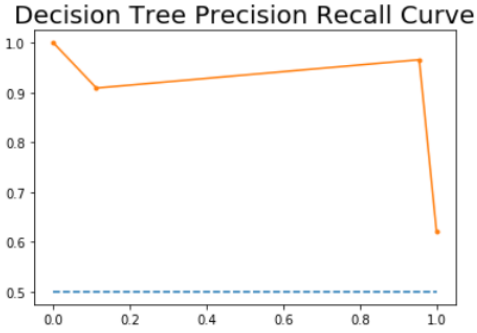

Figure 10. Precision Recall curve of Decision tree

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

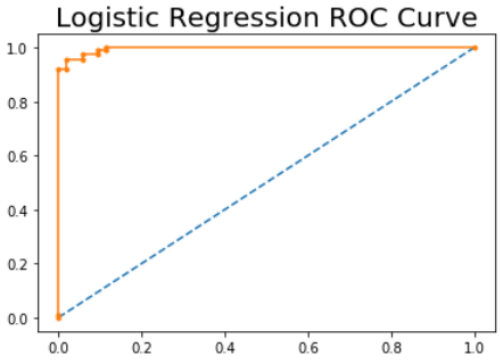

Figure 11. ROC of Logistic Regression

It plots the region operating characterstics of logistic regression classifier used in breast cancer prediction with an accuracy of 95%. 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).

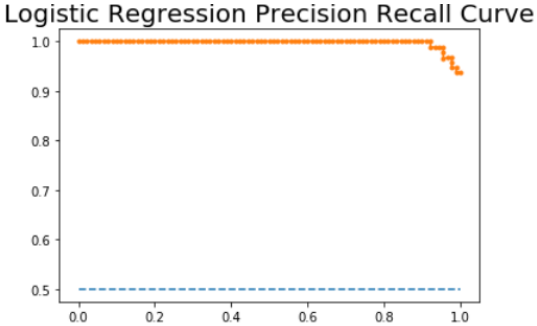

Figure 12. Precision Recall curve of Logistic Regression

The graph shows the precision recall curve of logistic regression in breast cancer prediction. 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

Datasets	Algorithms	Proposed system	Existing system
S1	Decision tree	83%	77.55%
S1	Logistic regression	88%	Nil
S2	Decision tree	95.1%	90.29%
S2	Logistic regression	95.8%	Nil

V. CONCLUSION AND FUTURE WORK

From the research work, it has been experimentally proven that Logistic regression provides perfect results as compare to Decision tree in both heart diseases and breast cancer prediction. For predicting heart diseases, significantly 15 attributes and for breast cancer, 30 attributes are listed and with basic machine learning algorithms. Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. In this paper the problem of constraining and summarizing different algorithms of data mining used in the field of medical prediction are discussed. The proposed work can be further enhanced and expanded for the automation of Heart disease and breast cancer 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. I also express my thanks to Prof. Yassir A the project coordinator,for guidance and whole hearted cooperation.

VII. REFERENCES

[1]. Tike Thein, Htet Thazin, and Khin Mo Mo Tun. "An Approach for Breast Cancer Diagnosis Classification Using Neural Network." Advanced Computing: An International Journal (ACIJ) 6 (2015) .

[2]. Puneet Yadav, Rajat Varshney, Vishan Kumar Gupta." Diagnosis of Breast Cancer using Decision Tree Models and SVM" International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 03 | Mar-2018 www.irjet.net p-ISSN: 2395-0072.

[3]. 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

[4]. H. Benjamin fredrick david and s. Antony belcy "heart disease prediction using data mining techniques" ictact journal on soft computing, october 2018, volume: 09, issue: 01.

[5]. 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

[6]. f0812effc72Shrivastava, Shiv, Anjali Sant, and Ramesh Aharwa. "An Overview on Data Mining Approach on Breast Cancer Data." International Journal of Advanced Computer Research (2013): n. page. Web

[7]. Nijil Raj N and T. Mahalekshmi, "Multilabel Classification Of Membrane Protein in Human by Decision Tree(DT) Approach." Biomedical & Pharmacology Journal Vol. 11(1), 113-121 (2018).

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

[9]. 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.

[10]. Nijil Raj N, Dr. T. Mahalekshmi, "Human Membrane Protein Classification using Multi class Support Vector Machine (MCSVM)" International Journal of Technology and Science, ISSN (Online) 2350-1111, (Print) 2350-1103

[11]. U. Ryu, R. Chandrasekaran, and V. S. Jacob, "Breast cancer prediction using the isotonic separation technique," Eur. J. Oper. Res.,vol. 181, no. 2, pp. 842–854, Sep. 2007.

[12]. Wolberg, WIlliam. "UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set." UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set. University of Wisconsin Hospitals Madison, Wisconsin, USA,Web. Oct. 2015.