JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Breast Cancer Using Data Mining Techniques

Disha Patel[1*], Bhavesh Tanwala[2], Pranay Patel [3]

[1,2,3]Computer Department, BVM Engineering College, V.V.Nagar, Anand, Gujarat, India.

**Corresponding Author: pateldisha567@gmail.com, Tel.: +91 9924766742*

Available online at: www.ijcseonline.org

Accepted: 19/Jul/2018, Published: 31/Jul/2018

Abstract— Breast cancer has a reason for the leading cause of death in women in various countries. The popular effective way to decrease breast cancer deaths is to detect it as earlier as possible. The classification of breast cancer data can be useful to predict the outcome of some disease or discover the genetic behavior of tumors. An early diagnosis method requires a more accurate and user reliable diagnosis techniques those are allow physicians to distinguish benign breast tumors from malignant ones without going for surgical biopsy. The objective of this paper is to find the classification of breast cancer as either benign or malignant .Then relative study on different cancer classification approaches viz, KNN, Decision tree and Neural Network classifiers are conducted where the accuracy of each of the classifier is also measured.

Keywords: Classification techniques; Data mining techniques; breast cancer; Diagnosis; Prognosis**.**

I. INTRODUCTION

Breast cancer is one of the most commonly occurring epithelial malignancies in women, and there are an estimated 1 million new cases and over 400,000 deaths annually worldwide. In the past 20 years, the incidence of breast cancer continues to rise. Then, the diagnosis and treatment of the breast cancer have become an extremely urgent work to do [2].

Data mining has various techniques such as classification, clustering, prediction, association rules, and regression .Among the various classification algorithms, the very famous algorithms ID3, C4.5 and latest C5.0 play an essential role in breast cancer analysis. Many researchers have attempted to apply machine learning algorithms for detecting survivability of cancers in human beings and it has been proved [3].

In short, this research is to identify the most successful data mining algorithm that helps to predict those cases of cancer, which can recur. Data mining techniques are applied to build the prediction model and in data mining fields, searching the relations between diseases and symptoms is a classification problem [4].

Symptoms of breast cancer:

- ➢ A lump in a breast
- ➢ A rash around (or on) one of the nipples
- ➢ A swelling (lump) in one of the armpits
- ➢ An area of thickened tissue in a breast
- ➢ The size or the shape of the breast changes
- ➢ One of the nipples has a discharge; sometimes it may contain blood
- ➢ The nipple changes in appearance; it may become sunken or inverted
- ➢ A pain in the armpits or breast that does not seem to be related to the woman's menstrual period
- ➢ Pitting or redness of the skin of the breast; like the skin of an orange

This paper is organized in different sections as follows; section 2 highlights the already published literature in the area of breast cancer survivability prediction models using data mining. Section 3, explains the detail description of data, various prediction algorithms and measures for performing an evaluation on the said models. The prediction results of all classification algorithms along with the accuracy, sensitivity , and specificity are presented in section 4. Section 5 concludes with a summary of results eventually leading to the future directions.

II. RELATED WORK

Performance Analysis of data mining algorithms for breast cancer cell detection using naïve Bayes, logistic regression , and decision tree Subrata Kumar Mandal information technology department Jalpaiguri government engineering college, Jalpaiguri, west Bengal, India. International journal of engineering and computer science. Feb 2017. In this paper, authors have applied techniques namely data cleaning, feature selection, data discretization and classification for predicting breast cancer as accurately as possible. Our study reveals that a logistic regression classifier gives the maximum accuracy with are a duced subset of features (four) and time complexity of this algorithm is least compared to the other two classifiers. [1]

Intelligent breast cancer prediction model using data mining techniques Runjie Shen, Yuanyuan yang, Fengfeng Shao, department of control science & engineering. Tongji University shanghais china.2014 IEEE. In this paper, we intend to build a diagnostic model of breast cancer by using data mining techniques. A feature selection method, interact is applied to select relevant features for breast cancer diagnosis and the support vector machine is used to build the classification model. Two diagnosis models are built with and without feature selection for the sake of proving the significance of feature selection. Through the experiments, the accuracy of the diagnostic model with feature selection I improved obviously compared with the model without feature selection. So as compared to other techniques with feature selection (interact) is the best accuracy 92%. [3]

A comparative survey of data mining techniques for breast cancer diagnosis and prediction. Hamid karim khani zand department of computer engineering, Iran University of science and technology, Tehran, Iran. These paper using seer dataset and compare the data mining techniques. And also they prediction on breast cancer data. So they classification algorithm is the best prediction as compared to clustering algorithms. [5]

A study on a prediction of breast cancer recurrence using data mining techniques. Uma Ojha computer science department ARSD College, Delhi University. Delhi India. And Dr.Savita Goel Sr.System programmer IIT Delhi.IEEE 2017. In this paper, they use different data mining techniques to predict all those cases of breast cancer that are recurrent using Wisconsin Prognostic Breast cancer (WPBC) dataset from the UCI machine learning repository. The C5.0 and SVM were the best predictor algorithm of 0.813 and fuzzy clustering came worst predictor of 0.3711. And also used KNN, PAM and EM. But best classifier is C5.0 and SVM to predict the highest accuracy. The result indicates the decision tree and SVM is the best predictor with 81% accuracy. [6]

Data mining techniques in multiple cancer prediction Dr.A.R Pon Periasamy Associate professor of computer science Nehru memorial college Puthanampatti, Trichy (DT) Tamilnadu, India. K.Arutchelvan Assistant professor/programmer Department of pharmacy Annamalai University, Chidambaram Tamilnadu, India. IJARC. May - 2017. In this paper, they used different data mining techniques which are classification, clustering and association mining. See5 is having the higher precision on correlation. [7]

C5.0 algorithm to improved decision tree with feature selection and reduced error pruning. Rutvija Pandya diploma computer engineering department, Gujarat Technology University, Atmiya institute of tech & Sci Rajkot and Jayati Pandya bachelor in computer science and application, saurahtra university K.P.Dholakiya info tech Amreli. IJAC May-2015. In this paper, they used different and also they provide feature selection, cross –validation, and reduced error pruning facilities. [9]

Breast cancer prediction using data mining techniques.S.padma Priya and P.Sowmiya Assistant professor & head department of information technology sri. Adi Chunchanagiri women’s college, jan-2018. In this paper, the classification algorithms ID3 and C4.5 are used to identify the various categories of breast cancer research. The works conclude that the performance of c4.5 is better than other algorithms. [11]

III. METHODOLOGY

Data Source:

In order to find out the best predictor model that can predict the recurrence cases of breast cancer, the authentic dataset has been used. In Wisconsin Breast cancer database University of Wisconsin Hospitals, There is 699 numbers of instances, 13 plus class attributes, and 16 missing values.

Data Mining:

Data mining is the process of extracting interesting patterns and knowledge from the data. This paper focuses on using some of the classification models to predict the chances of recurrences and survivability of the diseases. A short description of these algorithms and their specific accuracy and confusion matrix. [6]

There are some techniques which are classification, clustering, regression and so on. The classification is a function that assigns items in a collection to target categories or classes. The goal is to accurately predict the target classification algorithms are Decision Tree, KNN and Neural Network and so on.

Decision Tree:

In decision tree, it is a structure that includes root node, branches and leaf nodes. Each node denotes a test on an attribute, each branch denotes the outcomes on attest and each leaf nodes holds a class label. There are using different algorithm which are ID3, C4.5 and C5.0. The C5.0 algorithm is a decision tree that recursively separates observation in branches to construct a tree for the purpose of improving the prediction accuracy. [9]

The classifier is tested first to classify unseen data and for this purpose resulting decision tree is used. C4.5 algorithm follows the rules of ID3 algorithm. Similarity C5.0 algorithm has many features like:

- ➢ The large decision tree can be viewing as a set of rules which is easy to understand.
- ➢ C5.0 algorithm gives the acknowledge on noise and missing data.

- Problem of over fitting and error pruning is solved by C5.0 algorithm.
- In classification techniques the c5 classifier can anticipate which attributes are relevant and which are not relevant in classification.[10]

C5.0 model works by splitting the sample based on the field that provides the maximum information gain. Each subsample defined by the first split is the split again, usually based on a different field, and the process repeats until the subsamples cannot be split any further.

Finally, the lowest level splits are re-examined, and those do not contribute significantly to the value of the model are removed or pruned.

KNN (K-Nearest Neighbour):

The KNN is a non-parametric method used for classification and regression. It is a simple algorithm that stores all available cases and classifier new cases based on a similarity measure (e.g., distance functions) .A case is classified by a majority vote of its neighbours, with the case being assigned to the class most common amongst its K nearest neighbours measured by a distance function. If K=1 then the object is simply assigned to the class of that single nearest neighbour so they know is KNN classification and otherwise the output is the property value for the object is called KNN regression. The object is classified by a majority vote of its neighbours with the object being assigned to the class most common among K nearest neighbours.

Neural Network:

Neural networks are composed of multiple nodes, which imitate biological neurons of human brain. The neurons are composed by links and they interact with each other. The nodes can take input data and perform simple operations on the data.

The result of these operations is passed to other neurons. The output at each node is called its activation or node value. Each link is associated with weight. They are capable of learning, which takes place by altering weight values.

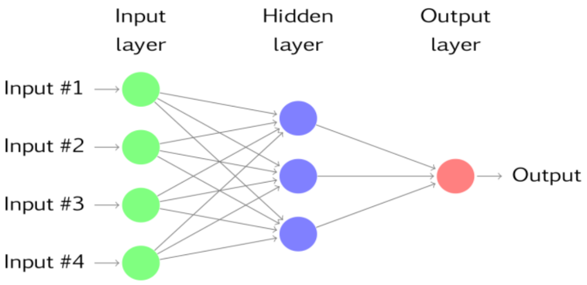
The commonest type of NN consists of three layers:

I. Input layer
II. Hidden layer
III. Output layer

In Input layer, the activity of the inputs units represents the raw information that is fed into the network.

In Hidden layer, the activity of each hidden unit is determined by the activities of the input units and the weights on the connections between the input and the hidden units.

In output layer, the behaviour of the output units depends on the activity of the hidden units and the weights between the hidden and output units.



IV. Results and Discussion

Performing of various classifiers in combination with data pre-processing methods is evaluated using breast cancer dataset. All the experiments except for the three classifier methods which are KNN, Decision tree and Neural Network were performed using RStudio software tools.

The output comes classifies whether the disease was recurrent or not and thus it was removed from the dataset so that we find how accurelty our data mining algorithm can predict all such cases. The results are show below.

Table 1 display results for three classification techniques applied on breast cancer dataset in RStudio.

Considering accuracy and Predict their 2 classes as performance measure classification techniques with highest accuracy obtained for data and the ratio is 70:30 to be applied all techniques.

Table 1 Result using RStudio

Techniques	Accuracy	Error rate
KNN	0.98	0.63
Decision Tree(C5.0)	0.90	0.62
Neural Network	0.95	0.60

Figure 1, 2 and 3 display the performance analysis of classification techniques using RStudio and their graphs of the techniques.

V. Conclusion

In this paper, we intend to a diagnostic model for breast cancer dataset. We have applied some data mining classification techniques which are KNN (K-Nearest Neighbour), Decision tree (C5.0) and Neural Network for predicting breast cancer accurelty as possible. The aim to have best or highest accuracy between those three algorithms and also using confusion matrix. So, in three algorithms best accuracy is KNN algorithm and worst accuracy is Decision tree (C5.0).and also compare their rate in Table No 2.

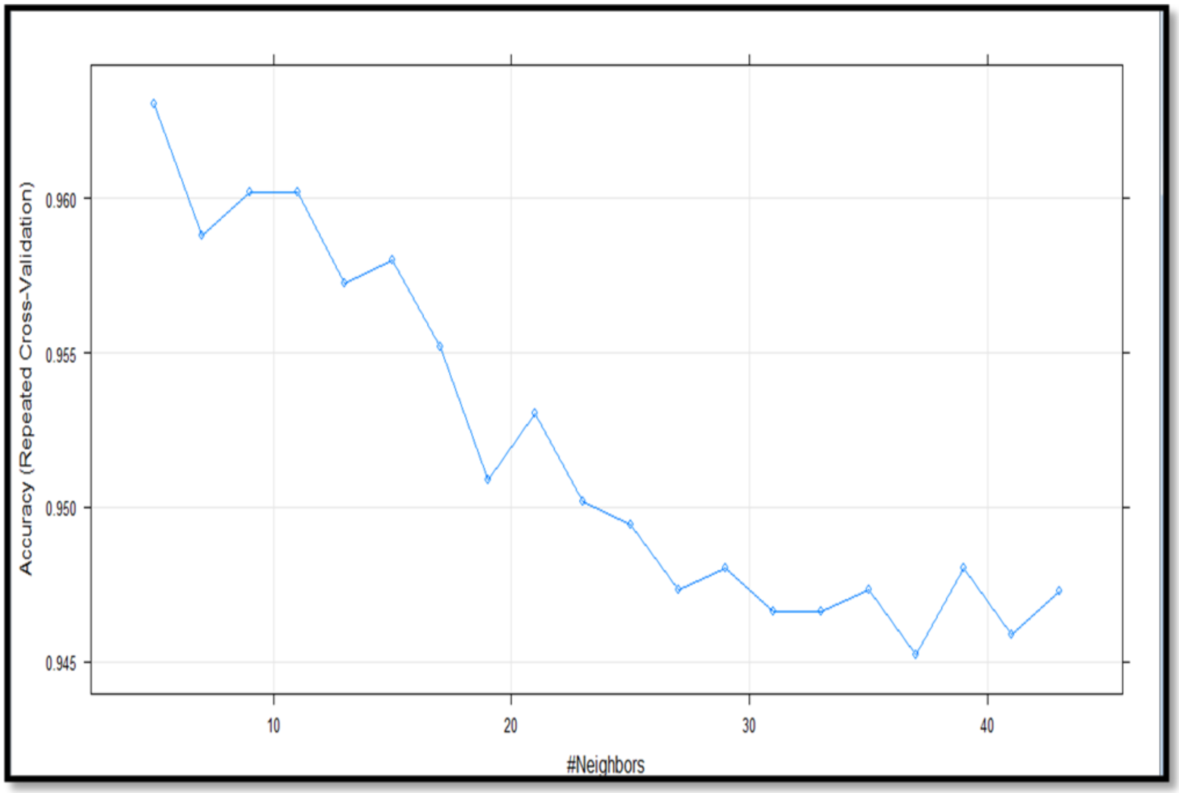


Figure 1 Result of KNN algorithm.

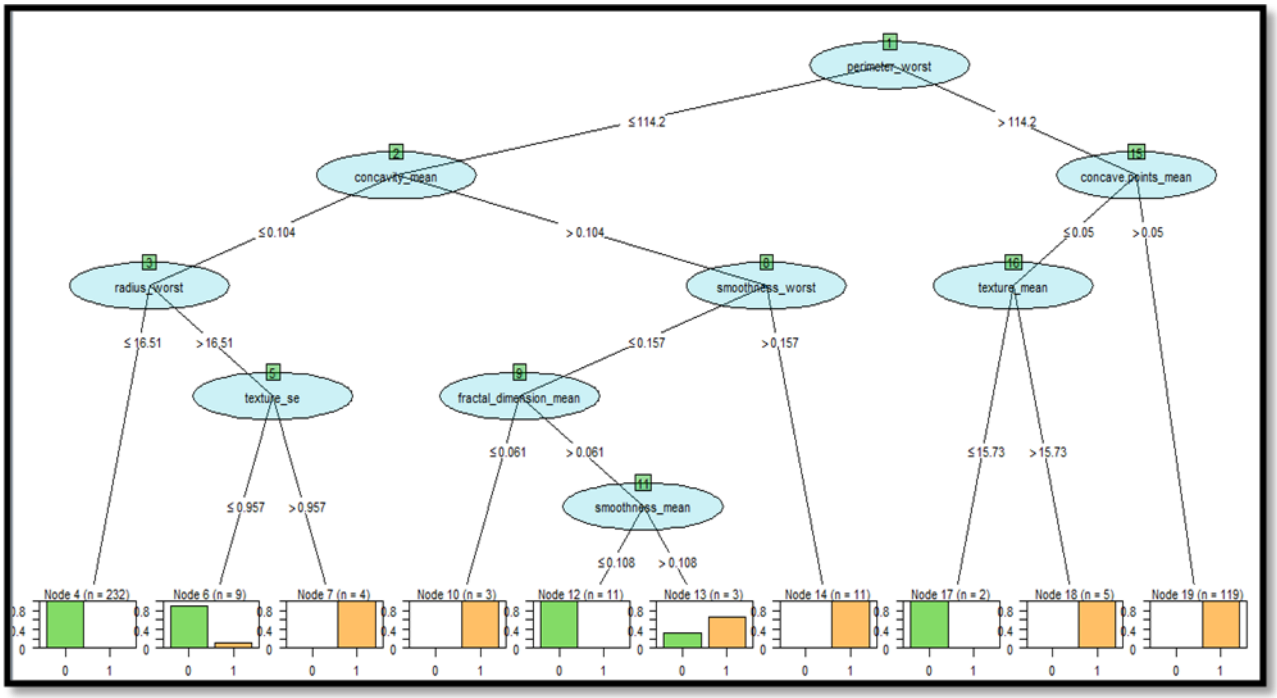
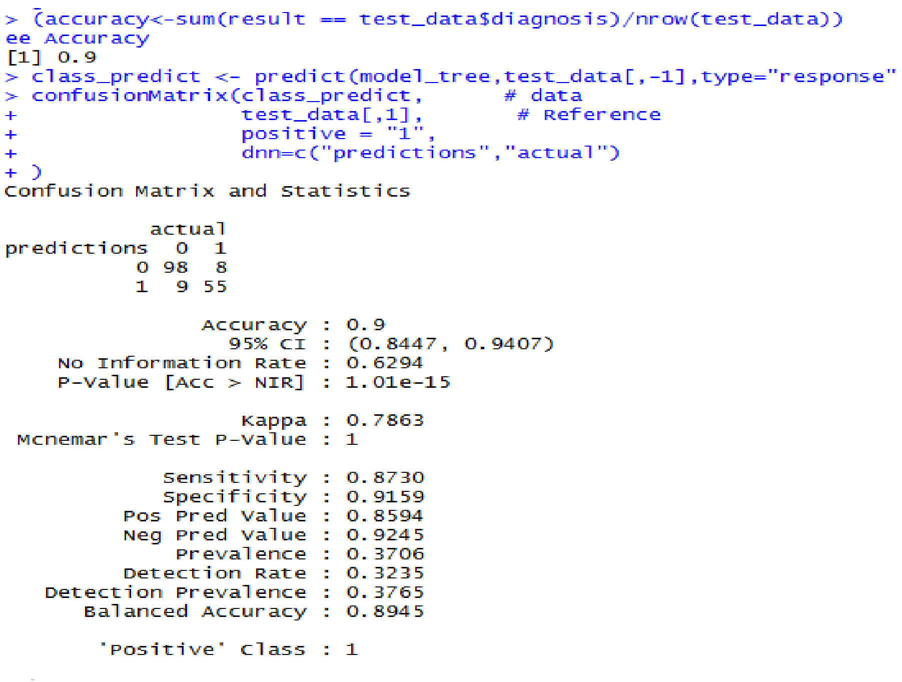


Figure 2 Result of Decision tree (C5.0)



```
> (accuracy<-sum(result == test_data$diagnosis)/nrow(test_data))
ee Accuracy
[1] 0.9
> class_predict <- predict(model_tree,test_data[,-1],type="response"
> confusionMatrix(class_predict,       # data
+                 test_data[,1],        # Reference
+                 positive = "1",
+                 dnn=c("predictions","actual")
+ )
Confusion Matrix and Statistics

           actual
predictions  0  1
          0 98  8
          1  9 55

               Accuracy : 0.9
                 95% CI : (0.8447, 0.9407)
    No Information Rate : 0.6294
    P-Value [Acc > NIR] : 1.01e-15

                  Kappa : 0.7863
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8730
            Specificity : 0.9159
         Pos Pred Value : 0.8594
         Neg Pred Value : 0.9245
             Prevalence : 0.3706
         Detection Rate : 0.3235
   Detection Prevalence : 0.3765
      Balanced Accuracy : 0.8945

       'Positive' Class : 1
```

Figure 3 confusion matrix of Decision tree (C5.0)

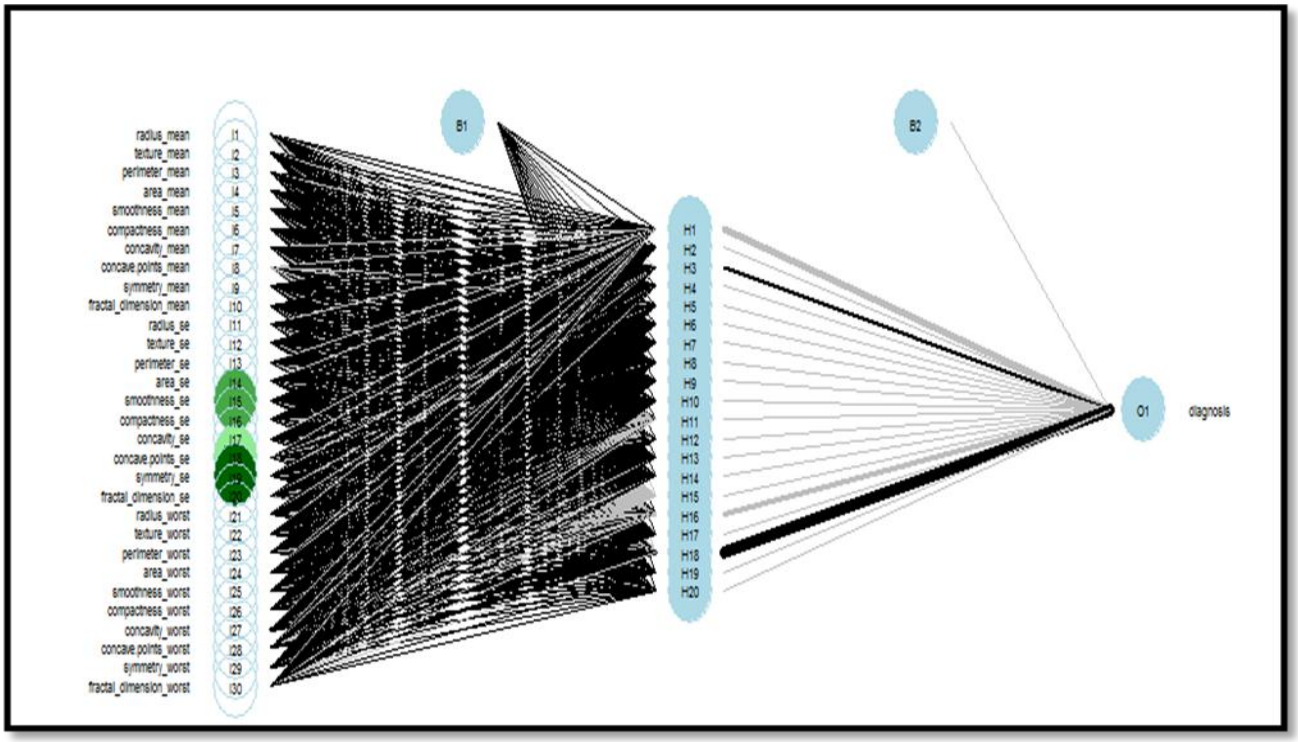


Figure 4 Result of Neural Network

ACKNOWLEDGMENT

This dissertation report has been unique and very memorable experience for me. It opened new venues of grabbing knowledge, which will certainly help me to stand in years to come. The activity of going through research has bridged between the academics and practical real life work for me. It has prepared by me to apply

myself better to become good researcher in professional as well. Naturally, it required lot of people support to complete this project. I take this opportunity to acknowledge their help to me.I wish to express my sincere gratitude towards my guide and Head of The Department **Mr. Bhavesh Tanawala and Mr. Pranay Patel** for giving me helpful support throughout this dissertation. I am glad to have the exposure of his great suggestions in research. I take this opportunity to express gratitude to all of the Department faculty members for their help and support.

References

[1]. A Study on Prediction of Breast Cancer Recurrence Using Data Mining Techniques. Uma Ojha Computer Science Department ARSD College, Delhi University Delhi-India and Dr. Savita Goel Sr.System Programmer IIT Delhi. 7th International Conference on Cloud Computing, Data Science & Engineering– Confluence. IEEE 2017.

[2]. Breast cancer prediction using data mining techniques. Padma Priya1, P.Sowmiya2. Assistant Professor & Head Department of Information Technology Sri Adi Chunchanagiri Women's College, Cumbum,(India Research Scholar, Department of Computer Science, Sri Adi Chunchanagiri Women's College, Cumbum,(India) .5th-6th janurary-2018

[3]. Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree Subrata Kumar Mandal Information Technology Department, Jalpaiguri Government Engineering College Jalpaiguri, West Bengal, India, International Journal Of Engineering And Computer Science . Volume 6 Issue 2 Feb., 2017

[4]. Intelligent Breast Cancer Prediction Model Using Data Mining Techniques Runjie Shen, Yuanyuan Yang, Fengfeng Shao, Department of Control Science & Engineering Tongji University Shanghai, China. 2014 Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics.2014 IEEE

[5]. A comparative survey on data mining techniques for breast cancer diagnosis and prediction Hamid Karim Khani Zand Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran. Indian Journal of Fundamental and Applied Life Sciences , Volume.5 2015

[6]. A Survey on Breast Cancer Analysis Using Data Mining Techniques B.Padmapriya, T.Velmurugan 2014 IEEE International Conference on Computational Intelligence and Computing Research.

[7]. A Study on Prediction of Breast Cancer Recurrence Using Data Mining Techniques. Uma Ojha Computer Science Department ARSD College, Delhi University Delhi-India and Dr. Savita Goel Sr.System Programmer IIT Delhi. 7th International Conference on Cloud Computing, Data Science & Engineering– Confluence. IEEE 2017

[8]. Data Mining Techniques in Multiple Cancer Prediction Dr. A. R. PonPeriasamy Associate Professor of Computer Science Nehru Memorial College Puthanampatti, Trichy (DT) Tamilnadu, India K. Arutchelvan Assistant Professor / Programmer Department of Pharmacy AnnamalaiUniversity,ChidamparamTamilnadu, India ,International Journal of Advanced Research in Computer Science and Software Engineering. Volume 7, Issue 5 May 2017

[9]. Breast Cancer Prediction using Data Mining Techniques Jyotsna Nakte Student, Dept. of Information Technology MCT Rajiv Gandhi Institute of Technology Mumbai, India, VarunHimmatramka Student, Dept. of Computer Engineering MCT Rajiv Gandhi Institute of Technology Mumbai, India, International Journal on Recent and Innovation Trends in Computing and Communication. Volume: 4 Issue: 11 Nov-2016.

[10]. C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning Rutvija Pandya Diploma Computer Engineering Department, Gujarat Technological University Atmiya Institute of Tech & Sci Rajkot Jayati Pandya Bachelor in Computer science and Application,Saurashtra University K.P.Dholakiya InfoTech Amreli International Journal of Computer Applications (0975 – 8887) Volume 117 – No. 16, May 2015

[11]. Performance Analysis of Data Mining Classification Techniques on Public Health Care Data Tanvi Sharma1, Anand Sharma2, Prof. Vibhakar Mansotra M. Tech Research Student, Dept. of Computer Science & I.T., University of Jammu, Jammu, J&K, India Research Scholar, Dept. of Computer Science & I.T., University of Jammu, Jammu, J&K, India Professor, Dept. of Computer Sci of Computer Science & I.T., University of Jammu, Jammu, J&K, India IJIRCE Volume: 3 Issue: 10 June 2016.

[12]. C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. Rutvija Pandya Diploma Computer Engineering Department, Gujarat Technological University Atmiya Institute of Tech & Sci Rajkot and Jayati Pandya Bachelor in Computer science and Application,Saurashtra University K.P.Dholakiya InfoTech Amreli .IJAC Volume 117 – No. 16 May 2015 .

[13]. An Overview on Data Mining Approach on Breast Cancer data. Shiv Shakti Shrivastava1, Anjali Sant, Ramesh Prasad Aharwal, International Journal of Advanced Computer Research, Volume-3 Number-4 Issue-13 Dec-2013

[14]. American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).