ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(29), DOI: 10.17485/ijst/2015/v8i29/84646, November 2015* ISSN (Online) : 0974-5645

Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification

E. Venkatesan* and T. Velmurugan

PG and Research Department of Computer Science, D.G. Vaishnav College, Chennai-600106, Tamil Nadu, India; venkatelumalai12@yahoo.co.in, velmurugan_dgvc@yahoo.co.in

Abstract

Backgrounds/Objectives: Data Mining (DM) techniques are extremely utilized for the extraction of useful information which is available in data warehouses and other database repositories. In medical diagnose, the role of DM approach rises quick recognition of disease over symptoms. To classify the medical data, a number of DM techniques are used by researchers. One of such techniques is classification. The classification algorithms predict the hidden information in the medical domain. The breast cancer is the very dangerous disease for women in developed countries like India. Most of the women death happens in the world, they are affected by the breast cancer. **Methods/Statistical Analysis**: The role of classification is importantin the real world applications in every field. Classification is used to classify the elements permitting to the features of the elements through the predefined set of classes. This research work analyses the breast cancer data using classification algorithms namely j48, Classification and Regression Trees (CART), Alternating Decision Tree (AD Tree) and Best First Tree (BF Tree). **Findings:** To find the performance of classification algorithms, this work uses cancer data as input. Particularly, this work is carried out to compare the four decision tree algorithms in the prediction of the performance accuracy in breast cancer data. All the algorithms are applied for breast cancer data to classify the data set for classification and prediction. Among these four methods, this work concludes the best algorithm for the chosen input data on decision tree supervised learning algorithms to predict the best classifier. **Applications/Improvements:** The breast cancer data is analyzed by taking the images using the same algorithms in future. Also, the microcalcifications of the breast cancer imagery are to be investigated in the same work.

Keywords: CART Algorithm, Classification Algorithms, Decision Trees, J48 Algorithm

1. Introduction

Data Mining (DM) is a technique which is used to find new, hidden and useful patterns of knowledge from large databases. From statistics, artificial intelligence and data warehouses, it is very easy to design methods and procedures to classify the data for the use of real world applications. DM concept is actually part of the knowledge discovery process[1]. DM has become a current technology in existing research and for medical field applications[2]. The data mining applications are applied to find the final result of a disease is one of the most inspiring work and a difficult task. By using some automated software tools, large volumes of medical data are being composedand readily available to the medical analysis and diagnosis groups. The DM techniques have become a widespread research instrument for medical scientist to identify and exploit patterns and interaction among large number of variable quantities and made them able to forecast the result before a disease using the past datasets[3].

Different methods of DM use different purpose of uses. The methods contribute some of its own advantages and disadvantages. To predict the tumors of images of patients, which is either a "benign" group that is non-cancerous or a "malignant" group for a cancerous group is the main use of classification algorithms. The extensively discussed classification problems in the prediction of breast cancer diagnostics are the scope of this research work. In DM, classification plays a crucial role in order to analysesthe supervised information.Classification is a supervised learning method and its objectives are predefined[3].

* *Author for correspondence*

The breast cancer is the major health issues as it is the main cause of death among the entire kinds of cancer particularly for women that too in the age of 35 to 55 years. So, far no known processes are there to avoid breast cancer. But, the way to detect the breast cancer at earlier stage will give more benefit for the affected peoples. A survey states that breast cancer became a raising issue in India where 1 in 22 women are estimated to have breast cancer. In U.S, the rate of breast cancer is very high where 1 in 8 women. Even the survival rate is also very high in U.S. For example, the survival rate of past five years of Indian women is 60% and 79-85% in the Developed countries. The current report by ICMR (Indian Council of Medical Research) states that the breast cancer cases will increase to 1,06,124 at 2015 and at 2020 it will raise to 1,23,634. The earlier stage of cancer detection can enhance the cure rate of breast Cancer[4]. As a survey 8.2 million cancer-related deaths occurred in 2012 according to The International Agency for Research on Cancer (IARC) statistics. In India, number of breast cancer cases will be estimated to double in 2025. Early stage of detection is the only way to prevent and protect us from breast cancer. Among the different techniques for identifying breast cancer, digital mammography is the most widely used screening tool. CAD technique could act as a second reader for assisting radiologist to identify the abnormalities in the mammogram[5].

DM methods in medical field are increasing faster due to the effective of the approaches to the prediction systems. A classification method is used and help medical practitioners in their decision making process to find diseases. In adding together to its significance in decision ways to develop patient results, reduce the cost of medicine for a patient and help in improving clinical studies and analyzing results. It is very pathetic to know that more than one million females is affected by breast cancer, the most common aggressive cancer in females and among them 600,000 proved to be fatalevery year[6]. Predictions given by the algorithm actually categorize the patients into two type of tumors -"benign" for non-cancerous or a "malignant" for cancerous[7]. Breast cancers developed due to the growth of uncontrolled cells in the tissues of breast which cause tumor in an abnormal cell growth[8]. Female breast is made up of fatty and fibrous connective tissues. It is very surprising to note that some men also prone to breast cancer, but frequency are very less[1].

For breast cancer details of genitourinary system that are usually extracted from the patient are Pregnancy Age, Number of Delivery (number of children born), Delivery age (Frequency of delivery based on number of children born), Usage of contraceptives (to avoid pregnancy), Miscarriages, Menarche (Attaining puberty – getting the first menstruation period) and Menopause (last menstruation). All these ages play a very vital role and it can be used for further analysis[9]. There is a possibility that even unmarried women can be affected by breast cancer because of the hereditary genes like BRCA1 and BRCA2.

This research work examines the various classification algorithms compared using WEKA tool environment and results are discussed[10]. This would classify the breast cancer dataset into the three breast cancer type (categories), depending on their characteristics, performance and other features. Several types of classification algorithms were selected and the dataset was applied with these algorithms. The classifiers used in this research work consist of common decision tree algorithms J48, CART, AD Tree and BF Tree. Here after, the information about the previous work done by various researchers in the comparative analysis among classification algorithms is described. The performance statistics of different dataset for medical and some other related applications were discussed. The main focus of this research work is making the possibilities for the selection of algorithms and dataset category to design proper medical applications in future. This work also creates simple strategy for the researcher or programmer to select the input parameter for classification creation of breast cancer data in medical domain.A research work done by S. Aruna, et al. uses supervised learning classifiers such as Support Vector Machine (SVM), Naïve Bayes, Radial Basis Function (RBF) Kernel, for breast cancer analysis which will lead to a best classifier.According to the results derived SVM RBF Kernel performs well than in parameters like sensitivity, precision, accuracy and specificity for both binary and multiclass datasets[11].

A research work was carried out by N. Poomani and R. Porkodi titled as "A Comparison of Data Mining Classification Algorithms using Breast Cancer Microarray Dataset". They compared on various supervised learning algorithms to predict the best classifier. The experimental result shows that the highest accuracy is found in J48 classifier, which gives 0.979 with the lowest error rate 0.9587 among various classification algorithms[12]. Another research work done by Gouda I. Salama, et al. titled as "Breast cancer finding on three altered datasets using multi-classifiers[6]". Their results show that fusion between

MLP and J48 classifiers with features selection Prognosis Breast Cancer (PCA) is excellent than other classifiers. Their results proved that the combination of IBK, J48, SMO and MLP works well in Wisconsin Prognosis Breast Cancer WPBC dataset[7].

Comparison of various breast cancer data sets is done by Saabithet et al. in his research work[13]. The classification algorithms such as J48, MLP and Rough set to estimate the proportion of exactness with and without characteristics selection methods for breast cancer data. They concluded that the characteristics choice technique is the most dependable significant process to improve the exactness of different categorization techniques, to achieve minimum Mean Standard Error (MSE) and maximum Recipient Operating Characteristics (ROC) to identify the breast cancer disease. An analysis is focused by means of two datamining techniques by Peter Adebayo Idowu, et al. in their research work to calculate breast cancer risks in Nigerian patients using the naive Bayesalgorithm and the J48 decision trees algorithms. The performance of both classification techniques was assessed in arrange to decide the most capable and useful model[14]. A research work compares DM techniques to model breast cancer data by S. Syed Shajahaan, et al. They explored the application of the decision tress in order to predict the presence of the breast cancer and also, the performance measurement of conservative supervised learning algorithms via, CART, C4.5, ID3 and Naive Byes. The experimental results proved that the random tree of ID3 is the best one with the high accuracy.

This research paper has been organized with Section 2 describing the materials and methods applied for this research work, Section 3 explaining the results of the four algorithms we have dealt and section 4 brings out the findings of this research.

2. Materials and Methods

There is a number of classification algorithms that has been proposed by several researchers in the field of classification applicationsand investigated in breast cancer data using decision tree algorithms. They used these algorithms to predict classification of breast cancer data. They selected classification algorithm to find the most suitable one for predicting cancer[15]. To classify breast cancer data set with high accuracy, efficiency and supervised learning algorithms via simple CART, J48, AD Tree and BF Tree in this work. Data pre-processing is performed in this research work by WEKA tool. In this research work, the breast cancer datais to be classified using the four classification algorithms and the classification based on the age of patientsand categories of cancer type.

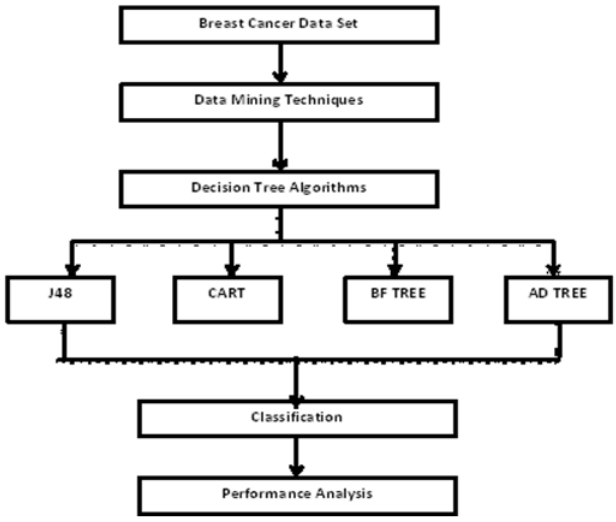

Figure 1. Methodology.

2.1 Dataset

In this research work, it is used nine attributes namely sex, age, present problem, past history, medical diagnosis, Occupation, food habit, Height and weight, the data taken from 220 patients at Swamy Vivekananda Diagnostic Centre Hospital, Chennai at D.G.Vaishnav College, India. Breast cancer images in DICOM (Digital Imaging and Communications in Medicine) format are taken for analysis. In this data set DICOM, the above said attributes are available. DICOM data set is the international standard for medical images and related information. Also, another one format used in this work is CSV (Comma Separated Value) format. The CSV format of breast cancer data is given as input via age, sex, modality, study description, date of image taken, image size and type etc. The data records have been created in Excel data sheet and saved in the format of CSV and then converted into the accepted WEKA format ARFF. In breast cancer data, benign and malignant are two types of properties. These data are processed in this research work. The breast cancer data attributes are summarized in Table 1. The method analyses only the main parts of breast cancer data by the classification algorithms J48, CART, ADTree and BFTree. The best performance of these algorithms is analyzed. In this research work, the breast cancer data has been analyzed considering the ages between 20 and 72.

Table 1. Description of the attributes

Variables	Description	Possible Values
sex	Patients	Female
Age	Teen, Middle and old	Age between 20 and 72 and above
present problem	Benign and malignant	Normal and Cancer
past history	Past disease	Nil or continues
medical diagnosis	Breast cancer	change in the size or shape of the breast
Occupation	Life style	Job
food habit	Eating	Both veg and non-veg
Height	Human height measure	Centimetres
weight	Human weight measure	Kilograms

2.2 Classification Algorithms

Classification is the most important task in Data Mining. It maps the data in to predefined targets. The goal is to build a classifier based on test cases with attributes to illustrate the items to designate the group ofthe objects. Next the group attributes based on the area values will be computed by the classifier and build a decision tree in which each non-leaf node will denote a test or decision on the considered data item. So to do classification, the scanning starts from the root node and traverse until a leaf node is reached. A decision will be built on reaching the terminal node[16].

2.3 J48 Algorithm

Quinlans C4.5 algorithm implements J48 to generate a trimmed C4.5 decision tree. The every aspect of the data is to split into minor subsets to base on a decision. J48 examine the normalized information gain that actually the outcomes the splitting the data by choosing an attribute. In order to make the decision, the attribute utmost standardized information gain is used. The minor subsets are returned by the algorithm. The splitting methods stop if a subset belongs to the same class in all the instances. J48 constructs a decision node using the expected values of the class. J48 decision tree can handle specific characteristics, lost or missing attribute values of the data and differing attribute costs. Here precision can be increased by pruning[17].

2.3.1 The Algorithm

Step 1: The leaf is labelled with the same class if the instances belong to the same class.
Step 2: For every attribute, the potential information will be calculated and the gain in information will be taken from the test on the attribute.
Step 3: Finally the best attribute will be selected based on the current selection parameter.

2.4 Classification And Regression Tree (CART)

In 1984, Leo Breiman, Jerome Friedman, Richard Olsen and Charles Stone jointly developed Classification and Regression Tree (CART) and derived a common method for developing statistical models from simple feature data. CART is powerful since it deals with data that is not fully finished, data with predicated and input features. The tree developed by CART will contain human readable rules. The algorithm will consider the set of samples-question about the data features will lead to the data minimization and continues till some stop criteria is reached[18]. CART is capable of handling both numerical and categorical variables. Gini index measures how well a given attribute separates training samples into targeted class. Here binary splitting of attributes takes place. It is most widely used statistical procedure. It provides a hierarchy of univariate binary decision.

2.4.1 The Algorithm

Step1: The first is how the splitting attribute is selected.
Step2: The second is deciding upon what stopping rules need to be in place.
Step3: The last is how nodes are assigned to classes.

2.5 ADTREE (Alternating Decision Tree)

Yoav Freund and Llew Mason introduced Alternating Decision Tree (ADTree), a machine learning method for classification, which generalizes decision tree and data structure. This tree predicts the nodes in the leaves and roots. The classification is done by traversing through all paths for all decision nodes. The binary classification trees are distinct and the AD Tree is different among that. In Binary Classification Trees, only one path is considered[19] while here all paths are considered.

2.6 Best First Tree (BF TREE)

Judeay described the best-first search which searches up to the collected point and additional knowledge about the problem domain. It expandsthe most promising node chosen based on specified rules. The following extended algorithm is to use an additional list, consists of all nodes that have been evaluate. It will avoid the node which evaluated twice and it is not have the concept of infinite loops.The inputs or instances to the decision tree algorithm are $(x_1, y_1)\ldots.(x_m, y_m)$ with x_i =vector of attributes and y_i = -1 or 1.Rule,the fundamental element of the ADTree algorithm may be aprecondition, a condition andtwo scores. A condition is of form "attribute <comparison> value" and aprecondition is a logical conjunction of conditions.

2.6.1 The Algorithm

Step1: Eliminate the best node from open, call it n, and it too closed.
Step2: if n is the goal state backtrack path to n (through-out recorded parents) and come back path.
Step3: create n's successors.
Step4: for each successor do:
Step5: if it is not in blocked and it is not in open: estimate it, add it to open andrecord its parent,
Step6: otherwise, if this fresh path is better than previous one, modify its recorded parent.
Step7: if it is not in open add it to open
Step8: otherwise, correct its main concern in open using this new evaluation.

3. Experimental Results

The main aim of this research proposal is to analyze the classification algorithms' performance for breast cancer data (output) based on the numerous input parameters as per Table 1. They are analyzed using decision tree j48,CART, AD Tree and BF Tree algorithms. The WEKA application is used for the performance evaluation. Each classifier is applied for two testing beds - Cross Validation which uses 10 folds with 9 folds used for training each time and 1 fold is used for testing and Percentage Split which uses 2/3 of the dataset for training and 1/3 for testing is given as output. The screen shot of the WEKA preprocessing stage is shown in Figure 1. In this, the classification of pre-processing is carried out based on all the values of taken nine attributes. A comparative study of classification accuracy in J48, CART, ADTree and BFTree algorithm is carried out in this work. The TP Rate FT Rate and precision analysis is also carried out. The various formul as used for the calculation of different measures are as follows. The following formula is used to calculate the proportion of the predicted positive cases, Precision P using TP = True Positive Rate and FP = False Positive Rate as,

$$\text{Precision P} = \frac{TP}{TP + FP} \tag{1}$$

It has been defined that Recall or Sensitivity or True Positive Rate (TPR) means the proportion of positive cases that were correctly identified. It will be computed as

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

Where FN =False Negative Rate

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

The above formula will calculate the accuracy (the proportion of the total number of predictions that were correct) with TN = True Negative

Sensitivity is the percentage of positive records classifiedcorrectly out of all positive records.

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{4}$$

Specificity is the percentage of positive records classified correctly out of all positive records.

$$\text{Specificity} = \frac{TN}{(TN + FP)} \tag{5}$$

The F-Measure can be computed as some average of the information retrieval precision and recall metrics.

$$F = \frac{|2 * \text{Re}call * precision}{precision + \text{Re}call} \tag{6}$$

ROC stands for Receiver Operating Characteristic. A graphical approach for displaying the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) of a classifier are given as follows.

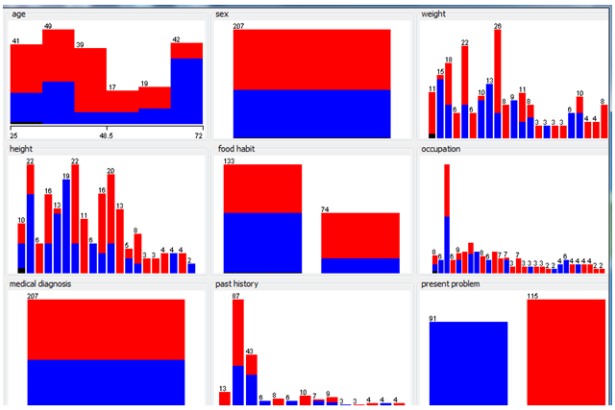

Figure 2. Data distribution in the preprocessing stage.

The experimental results of basic classifiers are discussed in this section. Breast cancer data contains tumors which represents the severity of the disease. To classify the tumors correctly from the training data set, the error rates and accuracy are calculated using classifiers. The accuracy of J48 algorithm is 99%, BF Tree 98%, AD Tree 97% and CART 96%. The confusion matrix helps us to find the various evaluation measures like accuracy, recall and precision, F-Measure and ROC Area etc. The performance of algorithms J48, Cart, AD Tree and BF Tree are given the Tables 2, 3, 4 and 5. The classification accuracy of four algorithms J48, CART, AD Tree and BF Tree are observed from the Tables 2, 3, 4 and 5 via values of weighted average, which is available in the last row of each table. The Table 6 depicts the error report of the four classification algorithms. Table 7 and Figure 4 shows the weighted average accuracy of the classification algorithm for the breast cancer data. The Figure 5 represents the comparison of the J48, CART, AD Tree and BF Tree classification algorithms based on the Table 8 values.

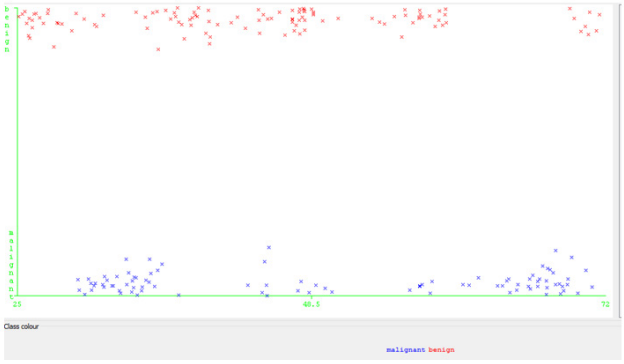

Figure 3. Age wise classification of breast cancer data

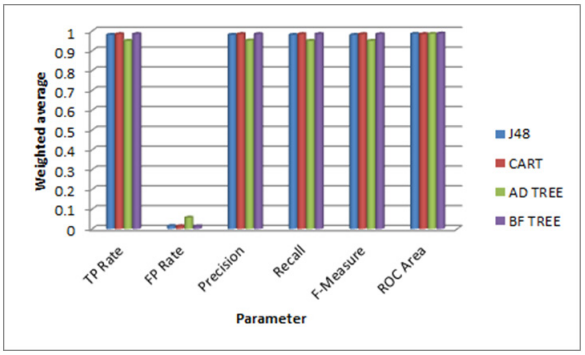

Figure 4. Weighted average of various parameters.

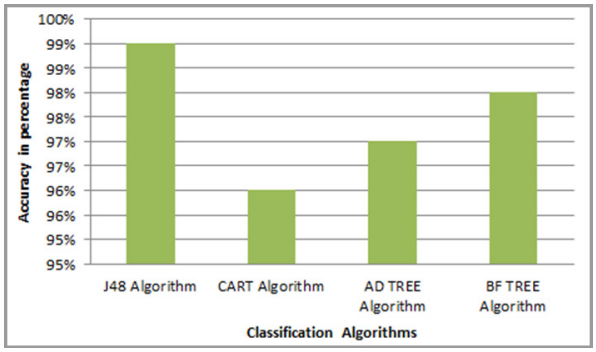

Figure 5. Performance comparison of algorithms.

Table 2. Results of J48

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Malignant	1	0.035	0.958	1	0.978	0.989
Benign	0.965	0	1	0.965	0.982	0.983
Weighted average	0.981	0.015	0.981	0.981	0.981	0.986

Table 3. Results of CART

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Malignant	1	0.017	0.978	0.989	0.984	0.985
Benign	0.983	0.011	0.991	0.983	0.987	0.985
Weighted average	0.985	0.014	0.985	0.985	0.985	0.985

Table 4. Results of AD Tree

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Malignant	0.912	0.071	0.976	0.912	0.943	0.986
Benign	0.983	0.088	0.934	0.983	0.958	0.986
Weighted average	0.951	0.057	0.953	0.951	0.951	0.986

Table 5. Results of BF Tree

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Malignant	0.989	0.017	0.978	0.989	0.989	0.988
Benign	0.983	0.011	0.991	0.983	0.987	0.988
Weighted average	0.985	0.014	0.985	0.985	0.985	0.988

Table 6. Error Reports

STATISTIC	J48	CART	AD TREE	BF TREE
Kappa statistic	0.9608	0.9705	0.9009	0.9705
Mean absolute error	0.0287	0.0239	0.1802	0.0258
Root mean squared error	0.1348	0.1201	0.2432	0.126
Relative absolute error	5.8164	4.8438	36.5192	5.231
Root relative squared error	27.1449	24.185	48.9692	25.376

Table 7. Accuracy by weighted average

S.No	Parameter	J48	CART	AD TREE	BF TREE
1	TP Rate	0.981	0.985	0.951	0.985
2	FP Rate	0.015	0.014	0.057	0.014
3	Precision	0.981	0.985	0.953	0.985
4	Recall	0.981	0.985	0.951	0.985
5	F-Measure	0.981	0.985	0.951	0.985
6	ROC Area	0.986	0.985	0.986	0.988

Table 8. Performance accuracy of algorithm

J48 Algorithm	CART Algorithm	AD TREE Algorithm	BF TREE Algorithm
99%	96%	97%	98%

4. Conclusion

This research work evaluate the performances in terms of classification accuracy of J48, AD Tree, BF Tree and regression trees (CART) algorithms using various accuracy measures like FP rate, TP rate, Recall, Precision, ROC Area and F-measure. Decision trees are standard constructs and easy to understand from which rules can be extracted. In the implementation process, it is considered only the numerical values of some attributes in the breast cancer data.The experimental results shows that thehighest accuracy 99% is found in J48 classifier and accuracy 96% is found in CART algorithm, 97% in AD Tree algorithm and98% in BF Tree algorithm. Based on the classification results of all the four algorithms, the performance of J48 is better than the other three algorithms for the chosen data set.

5. References

1. Syed SS, Shanthi S, Chitra VM. Application of Data Mining techniques to model breast cancer data. International Journal of Emerging Technology and Advanced Engineering. 2013 Nov; 3(11):362–9.
2. Shiv Shakti S, Sant A, Aharwal RP. An Overview on Data Mining Approach on Breast Cancer data. International Journal of Advanced Computer Research. 2013; 3(13):256–62.
3. Shweta K. Using data mining techniques for diagnosis and prognosis of cancer disease. International Journal of Computer Science, Engineering and Information Technology. 2012 Apr; 2(2):55–66.
4. Takiar R, Nadayil D, Nandakumar A. Projections of number of cancer cases in India (2010-2020) by cancer groups. Asian Pac J Cancer Prev. 2010; 11(4):1045–9.
5. Vaidehi K, Subashini TS. Breast tissue characterization using combined K-NN classifier. Indian Journal of Science and Technology. 2015 Jan;8(1):23–6.
6. Salama GI, Abdelhalim MB, Abd-elghany Zei Md. Breast cancer diagnosis on three different datasets using multi-classifiers. International Journal of Computer and Information Technology. 2012 Sep;1(1):36–43.
7. Gupta S, Kumar D, Sharma A. Data mining classification techniques applied for breast cancer diagnosis and prognosis. Indian Journal of Computer Science and Engineering. 2011 Apr-May; (2):188–95.
8. Rajesh K, Anand S. Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm. International Journal of Advanced Research in Computer and Communication Engineering. 2012 Apr; 1(2):72–7.
9. Saravana Kumar K, Arthanariee AM. Evaluate the multiple breast cancer factors and calculate the risk by software tool breast cancer risk evaluator. Indian Journal of Science and Technology. 2015 Apr;8(S7):686–91.
10. Arora R, Suman S. Comparative analysis of classification algorithms on different datasets using WEKA. International Journal of Computer Applications. 2012; 54(13):21–5.
11. Aruna S, Rajagopalan SP, Nandakishore LV. Knowledge based analysis of various statistical tools in detecting breast cancer. Computer Science & Information Technology. 2011; 2:37–45.
12. Poomani N, Porkodi. R. A comparison of Data Mining classification algorithms using breast cancer microarray dataset: A study. International Journal for Scientific Research and Development. 2015; 2(12):543–7.
13. Saabith ALS, Elankovan S, Abu Bakar A. Comparative study on different classification techniques for breast cancer dataset. 2010; 3(10):185–91.

14. Williams K, Idowu PA, Balogun JA, Oluwaranti A. Breast cancer risk prediction using data mining classification techniques. Transactions on Networks and Communications. 2015; 3(2):1–11.
15. Bellaachia A, Guven E. Predicting breast cancer survivability using data mining techniques. Society for Industrial and Applied Mathematics. 2006; 58(13):1–4.
16. Aloraini A. Different machine learning algorithms for breast cancer diagnosis. International Journal of Artificial Intelligence and Applications. 2012 Nov; 3(6):21–30.
17. Kaur G, Chhabra A. Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications. 2014 Jul; 98(22):13–7.
18. Wang KJ, Adrian AM. Breast cancer classification using hybrid synthetic minority over-sampling technique and artificial immune recognition system algorithm. International Journal of Computer Science and Electronics Engineering.2013; 1(3):408–12.
19. Sharma AK, Sahni S. A comparative study of classification algorithms for spam email data analysis.International Journal on Computer Science and Engineering.2011 May; 3(5):1890–5.