International Journal of Data Mining Techniques and Applications
Volume 5, Issue 1, June 2016, Page No.43-49
ISSN: 2278-2419

# Classification Algorithm Based Analysis of Breast Cancer Data

B.Padmapriya[1], T.Velmurugan[2]
[1]Research Scholar, Bharathiyar University, Coimbatore, Tamil Nadu, India,
[2]Associate Professor, PG.and Research Dept. of Computer Science, D.G. VaishnavColllege, Chennai-600106, India,
Email: ppwin74@gmail.com[1], velmurugan_dgvc@yahoo.co.in[2]

Abstract- The classification algorithms are very frequently used algorithms for analyzing various kinds of data available in different repositories which have real world applications. The main objective of this research work is to find the performance of classification algorithms in analyzing Breast Cancer data via analyzing the mammogram images based its characteristics. Different attribute values of cancer affected mammogram images are considered for analysis in this work. The Patients food habits, age of the patients, their life styles, occupation, their problem about the diseases and other information are taken into account for classification. Finally, performance of classification algorithms J48, CART and ADTree are given with its accuracy. The accuracy of taken algorithms is measured by various measures like specificity, sensitivity and kappa statistics (Errors).

Keywords- Classification Algorithms, Breast Cancer Analysis,J48 Algorithm, CART Algorithm and AD tree Algorithm**.**

## I. INTRODUCTION

Data mining (DM) is the use of automated data analysis techniques to uncover previously undetected relationships among data items. DM often involves the analysis of data stored in a data warehouse. Three of the major data mining techniques are regression, classification and clustering. DM techniques are applied into the scientific, business, intelligent data analysis, mathematical and the medical applications. DM refers to extracting or "mining" knowledge from large amounts of data or database.It can be defined as the non-trivial extraction of potentially useful information from a large volume of data where the information is implicit. The analytical tools of Data Mining are drawn from a number of disciplines, which may include machine learning, pattern recognition, machine discovery, statistics, artificial intelligence, human-computer interaction and information visualization. The DM technique classification is used for this research work to analyze the cancer affected mammogram images. There are generally two stages of cancer, benign andMalignant. If cancer is detected in benign stage, life expectancy of a patient increases. The malignant stage develops when cells in the breast tissue divide and grow without the normal controls on cell death and cell division. Cancer on breast tissue is called breast cancer [1].

Recently many new cancer detection and treatment approaches were developed, the cancer incidences and death of breast cancer decreased constantly.The advance of cancer diagnosis and treatments, reduce the number of mortalities and increase the survival time for patients [2].The classification of Breast Cancer data can be useful to predict the result of some diseases or discover the genetic behavior of tumors. There are many techniques to predict and classification in breast cancer pattern. This research work empirically compares performance of different classification algorithms that are suitable for direct interpretability of their results [3].Many researchers taking different attributes for the analysis of cancer affected images. One of the research article in [4] uses attribute selection is a popularly useddimensional reduction technique, which has been the focus ofresearch in machine learning and data mining and used inmedical image mining and analysis. It helps to construct simple, comprehensive classification models with classificationperformance. Even though several models exist for featurereduction and selection process only few will be suitable for anenvironment application. Thus, it is necessary to study thesuitability for attribute selection methods for our mammogram data base.

Vikas Chaurasia, Saurabh Pal [5] has developed, the accuracy of classification techniques is evaluated based on the selected classifier algorithm.Specifically, and they used three popular data mining methods: Sequential Minimal Optimization (SMO), IBK, and BestFirst Tree. They give that it is an important challenge in data mining and machine learning areas to build the precise and compute efficientclassifiers for Medical applications. The performance of SMO shows the high level comparisonand concrete results in Breast Cancer disease. Therefore the SMO classifier issuggested for diagnosis of Breast Cancer disease based on the classification to get better results with accuracy, low error rate and performance.

Zarei.S, and et.al [6] have analyzed about the multivariate linear regression, logistic regression, the K-nearest neighbor (KNN) method and discriminate analysis to determine tumor type in a patient using Wisconsin Breast Cancer (WBC) database. Stepwise method for variable selection is used in regression method.With respect to coefficient of variation, although model 1 has three independent variables less than complete model and there is no significance difference between these two models. Despite of this, they used, new model to reduce the cost and time of diagnosis of tumor type. KNN classifier has no restricted assumption and has a very good precision, andsuggested this method as auxiliary for breast cancer diagnosis.In this paper, reviewed about the various research works carried out using J48, CART and ADTree algorithms, done by different researchers. This will identify to predict general and individual performance of patients. The remaining of this paper is organized as follows. Section II discusses about different research articles via its literature survey. The Classificationalgorithms are illustrated in section III. Section IV, the experimental work conducted using Weka tool is discussed. Finally, section V concludes the research work.

## II. REVIEWOF LITERATURE

This section contains the description of the literature that has been done on Breast Cancer Analysis and classification techniques. Comparative study of different classification techniques is summarized with advantages and disadvantages. Most of the key features, methods are mentioned below with respective limitations and benefits that make our work unique.Clinical diagnosis of breast cancer helps in predicting the malignant cases. Various common methods used for breast cancer diagnosis are Mammography, Positron Emission Tomography, Biopsy and Magnetic Resonance Imaging. This section includes the review of various technical and review articles on data mining techniques applied in breast cancer diagnosis.Ireaneus.Y et al., discussed about the tumors in early detection of Breast Cancer Using SVM Classifier Technique. The mammogram is divided into three main stages [7]. The first step involves an enhancement procedure and image enhancement techniques.To makes certain features it is easy to modify the colors or intensities in image by increasing the signal and noise ration. Then the features are extracted from the segmented mammogram. The next stage involves the classification using SVM classifier.K.Rajesh, et.al [8] classified about SEER breast cancer data into the groups of "Carcinoma in situ" and "Malignant potential" using C4.5 algorithm. They obtained an accuracy of 94% in the training phase and an accuracy of 93% in the testing phase. They have compared the performance of C4.5 algorithm with other classification techniques.VanajaS., K. Rameshkumar, [9] discussed about theC4.5 classification algorithm is the extractionof ID3 algorithm. It supportcontinuous attributes and shows the best accuracy onattribute with missing values. Theinformation gainby attribute measurement, which indicates the percentage by given attribute and separate dataset according to their finalclassification.Both C4.5 and C5.0 can produce classifiersexpressed either decision trees or rulesets. In manyapplications, rule sets are preferred because they aresimpler and easier to understand than decision trees, butC4.5's ruleset methods are slow and need more memory.C5.0 embodies new algorithms for generating rulesets and the improvement is substantial.

The alternating Decision Tree (ADTree) is a successful machine learning classification technique that combines many decision trees. It uses a meta-algorithm boosting to gain accuracy. The induction algorithm is used to solve binary classification problems. The splitter node and two prediction nodes are added to generate a decision tree. The algorithm determines a place for the splitter node by analyzing all prediction nodes. Then the algorithm takes sum of all prediction nodes to gain overall prediction values. A positive sum represents one class and a negative sum represents the other in two class data sets. A special feature of ADTree can be merged together. In multiclass problems the alternating decision tree can make use of all the weak hypotheses in boosting to arrive at a single interpretable tree from large numbers of trees were discussed byChandrahasan, R. Kalaichelvi, et al [10].Comparison of robustness against missing values of alternative decision tree and multiple logistic regressions for predicting clinical data in primary breast cancer were discussed bySugimoto et.al, [11]. Mammogram based on multiple logistic regressions (MLR) is a standard technique for predicting diagnostic and treatment. To overcome these issues, they have developed prediction models using ensembles of alternative decision trees (ADTree) andthey compare the performance of MLR and ADTree models in terms of robustness against missing values. In this case study, employ datasets including pathological complete response (PCR) of neo adjuvant therapy, is one of the most important decision-making factors in the diagnosis and treatment of primary breast cancer. Ensemble ADTree models are more robust against missing values than MLR.

Nithya, R. and B. Santhi, [12] discussed about the decision tree classifiers for mass classificationfrom the mammogram is important for breast cancer diagnosis. This paper proposes the classification method for breast masses using the decision tree techniques. This paper presents the comparison result of 12 decision tree algorithms including ADTree, BFTree, DecisionStump, FT, C4.5, LADTree, LMT, NBTree, RandomForest, RandomTree, REPTree and CART. In comparison, four performancesof metrics were used. The aim of the study is to determine the best decision tree classifier for mass classification from BI–RADS features (mass shape, mass margin, assessment and subtlety). In the experimental studies, decision tree algorithms are applied on the UCI data set. This work employed various decision tree algorithms. Twelve decision tree algorithms are evaluated using UCI mass data set. LADTree and LMT decision tree algorithms are performed well. Four features (assessment, margin, shape and patient age) are identified to be most important feature for mass classification. In a research work done by J.S.Saleema, et al., [13], discussed about the Lung cancer, breast cancer, colon cancer and colorectal cancer using the base classifiers have been found in the literature. Prediction of cancer patients are of two types. First one is based on the severity of diseases from the date of diagnosis; a patient's survival period is defined. The second one predicts using the follow-up features after the diagnosis. Less work is found in the survival of combined data would be similarities in age, morphological data, stage and extent of disease that lead as a motivation for this paper. The SEER data set has been generated from different sources and the data recoding strategy based on the medical dictionary among the SEER community leaves data sparse and skewed.

There are several algorithms to classify the data using decision trees. The frequently used decision tree algorithms are ID3, C4.5 and CART. The CART algorithm is chosen to classify the breast cancer data because it provides better accuracy for medical data sets than ID3. It is based on Hunt's algorithm. CART handles both categorical and continuous attributes to build a decision tree. It also handles missing values. CART uses Gini Index as an attribute selection measure to build a decision tree. ID3 and C4.5 algorithms, CART produces binary splits. Hence, it produces binary trees. Gini Index measure does not use probabilistic assumptions like ID3 and C4.5. CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy, were discussed by Lavanya, D., and Dr K. Usha Rani. [14] Analysis of feature selection with classification inbreast cancer datasets.H.S.Hota,discussed about the CART (Classification and Regression Tree) classification algorithm [15] which is based on decision tree induction. The Classification and Regression tree method uses recursive partitioning to split the training records into segments with similar output field values.

The CART tree node starts by examining the input fields to find the best split, measured by the reduction in impurity index, which results from the split. CART uses Gini index splitting records measures in selecting the splitting attribute. Pruning is done in CART by using a training data set. Li, J., &Gramatica P., [16] discussed about the Classification and virtual screening of androgen receptor antagonists. Alternating decision trees (ADTree) is a kind of option tree and used as a tool to mine the NCI human tumor cell line databaseand analyze the mass spectrometry data, etc. Option trees differ from decision trees; they contain two types of nodes, a decision node and a prediction node, while decision trees just contain a decision node. When a query reaches a decision node, the sign of this node will be assigned to the query, like in the decision tree. So in an alternative decision tree, they find different branches (multipath). This is commonly done by using the boosted algorithm, and the resulted trees are usually called ADTree instead of option trees.Decision tree provides a powerful technique for classification and prediction in Breast Cancer diagnosis problem [17]. Various decision tree algorithms are available to classify the data, including ID3, C4.5, C5, J48, CART and CHAID. In this paper,they have chosen J48 decision tree algorithm to establish the model. 10-fold cross validation is used to prepare for the training and test data. After data pre-processing, the J48 algorithm is employed on the dataset using WEKA tools after which data are divided into "benign" or "malignant" depending on the final result of the decision tree that is constructed. Evaluation of Decision Tree Classifiers on Tumor Datasets was discussed by G. Sujatha and Dr. K. Usha Rani [18].Frequently used classifiers are ID3, C4.5 and CART.These experiments are conducted on those classifiers forbetter accuracy and execution time to construct the tree. They observed that C4.5 performs well for tumor datasets, ifavailable datasets are used as it is. Among these threealgorithms, C4.5 is the best one for enhanced dataset of Primary tumor and for enhanced Colon tumor dataset both ID3 and C4.5 exhibit equal classificationaccuracy. In future performance of this experiment with ensemble technique specified decision tree classifiers for analysis.

## III. MATERIALS AND METHODS

Classification algorithms had been proposed by several researchers in the field of classification of application and investigated in breast cancer data using decision tree algorithms. They used algorithms to predict classification of breast cancer data and find the most suitable one for predicting cancer [19]. To classify breast cancer data set with high accuracy and efficiency of learning algorithms in the way of simple methods such as J48, CART, ADTree. Data pre-processing was performed in this research work by WEKA tool for modeling breast cancer data, and this tool was accustomed for the data mining software. The data mining methods proposes many exploratory data analysis, statistical learning, machine learning and database analysisfor the real world problems [20].In this research work, the breast cancer data is classified using the three classification algorithms and it is based on the age of patientsand categories of cancer type.

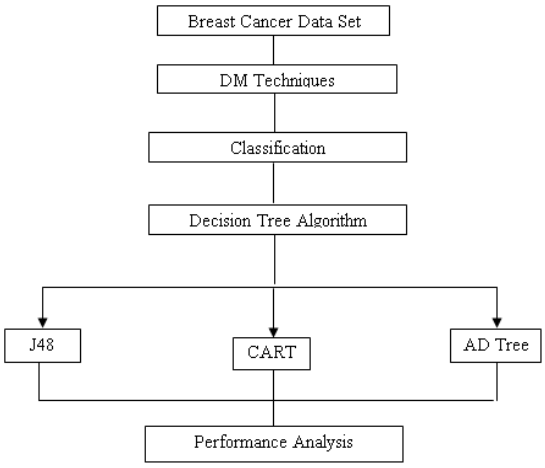


Figure 1: The proposed Method

Table 1: Description of the Data Set

| Attributes | Possible Values | Description |
|---|---|---|
| Sex | Female | Patients |
| Age | Age between 20 and 72 and above | Teen, Middle and old |
| Height | Centimeters | Human height measure |
| Weight | Kilograms | Human weight measure |
| Present problem | Normal and Cancer | Benign and malignant |
| Past history | Nil or continues | Past disease |
| Medical diagnosis | change in the size or shape of the breast | Breast cancer |
| Occupation | Job | Life style |
| Food habit | Both veg and non-veg | Eating |

### *A. Description of Data Set*

In this research work, nine attributes are used namely age, sex, present problem, past history, medical diagnosis, Occupation, food habit, Heightand weight. The dataset comprises of 250 instances of breast cancer patients with each, either having malignant or benign type of tumor. The data has been taken for 250 patients at the Cancer Institute, Adyar, Chennai, Tamilnadu, India. Breast cancer images in DICOM (Digital Imaging and Communications in Medicine) format are taken for analysis. The above said attributes are available in the DICOM. DICOM data set is the international standard for medical images and related information and supporting the encapsulation of any Information Object definition. We used another one format in this work is CSV (Comma Separated Value) format. This CSV format, breast cancer data is given as input via age, image size, sex, modality, study description, date of image taken, and type etc. These data records are created in Excel data sheet, saved in the format of CSV and then converted into the accepted WEKA format ARFF. There are two types of properties namely benign and malignant in breast cancer data. 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 and AD Tree. In this research work, the breast cancer data has been analyzed considering the ages between 20 and 72.

### *B. Classification Algorithms*

Classification is one of the data mining methodologies used to predict and classify the predetermined data for the specific class. There are different classifications methods proposed by researchers. The data mining consists of various methods. Hence, the breast cancer diagnostic problems are basically in the scope of the widely discussed by Shelly Gupta et al. [21] about the classification problems. Classification is a method used to extract models describing important data classes or to predict the future data.Classification is two step processes:

(i) Learning or training step where data is analyzed by a classification algorithm.
(ii) Testing step where data is used for classification and to estimate the accuracy of the Classification [22].

### *C. J48 Algorithm*

Decision tree (J48):Decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label.

The decision tree classifier has two phases:

i) Growth phase or Build phase.
ii) Pruning phase.

The tree is built in the first phase by recursively splitting the training set based on local optimal criteria until all or most of the records belong to each partition.The pruning phase handles the problem of over fitting the data in the decision tree. It removes the noise and outliers. The accuracy of the classification increases in this phase [22].J48 or C4.5: This algorithm is based on Hunt's algorithm [23]. It handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. It uses Gain Ratio as an attribute selection measure to build a decision tree. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

***The Algorithm***

Step1: In case the instances belong to the same class the tree denotes a leaf so the leaf is returned by labeling with the same class.
Step 2: The potential information is calculated for every attribute, given by a test on the attribute. Gain in information is considered that would result from a test on the attribute.
Step 3: The best attribute is found on the basis of present criterion and that attribute selected for branching [23].

### *D. Classification And Regression Tree (Cart)*

CART (Classification and Regression trees) was introduced by Breiman in 1984 [24]. It builds both classifications and regressions trees. It is also based on Hunt's model of Decision tree construction and can be implemented serially. It uses gini index splitting measure in selecting the splitting attribute. Pruning is done in CART by using a portion of the training data set. CART uses both numeric and categorical attributes for buildingthe decision tree and has in-built features that deal with missing attributes. The CART approach is an alternative to the traditional methods forprediction .In the implementation of CART, the dataset is split into the two subgroups that are the most different with respect to the outcome. This procedure is continued on each subgroup until some minimum subgroup size is reached.The Algorithmic steps are given below:

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

### *E. Adtree (Alternating Decision Tree)*

The ADTree is considered as another semantic for representing decision trees [25]. In the ADTree, each decision node is replaced by two nodes: a prediction node and splitter node. The decision tree in ADTree algorithm is identical while the prediction node is associated with a real valued number. As it is stated in the decision tree, an instance is mapped into a path along with the tree from the root to one of the leaves. The classification in ADTree that is associated with the path is not the label of the leaf. Instead, it is the sign of sum of the prediction along the path.This is different from binary classification trees such as CART or C4.5 which an instance follows only one path through the tree.The Algorithmic steps are given below.

Step1: if (precondition)
Step 2: if (condition)
Step3: return score one
Step4: else
Step5: return score two
Step6: end if
Step7: else
Step8: return 0
Step9: end if

## IV. EXPERIMENTAL RESULTS

In this experiment the medical data related to breast cancer is considered because the breast cancer is one of the leading causes of death in women. The experiments are conducted using Weka tool. In this study ADTree algorithm is chosen to analyze the breast cancer datasets because it provides better accuracy for medical data sets than the other two frequently used decision tree algorithms C4.5 and CART. With an intension to find out whether the same feature selection method may lead to best accuracy for various datasets of same domain, various experiments are conducted on three different breast cancer datasets. The preprocessing the data set is shown in figure 1. In this, the classification of pre-processing is carried out based on all the values of taken nine attributes. The two types of tumors, benign and malignant is serious cancer. A comparativestudy of classification accuracy in J48, CART, and ADTree algorithm is carried out in this work. The TP Rate FT Rate and precision analysis is also carried out.The various formulas used for the calculation of different measures are as

International Journal of Data Mining Techniques and Applications
Volume 5, Issue 1, June 2016, Page No.43-49
ISSN: 2278-2419

follows. PrecisionP is the proportion of the predicted positive cases that were correct, as calculated using the formula

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

Where TP is the True Positive Rate, FP means False Positive Rate.
Recall or Sensitivity or True Positive Rate (TPR): o it is the proportion of positive cases that were correctly identified, as calculated using the equation

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

Here FN meansFalse Negative Rate
Accuracy is the proportion of the total number of predictions that were correct. It is determined using the equation.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Where TNstands for True Negative Sensitivity is the percentage of positive records classified correct output of all positive records.

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

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

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

The F-Measure computes some average of the information retrieval precision and recall metrics.

$$F = \frac{|2 * \text{Re}\,call * precision}{precision + \text{Re}\,call} \quad (6)$$

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.

TPR = positives correctly classified/total positives
FPR = negatives incorrectly classified/total negatives
TPR is plotted along the y axis o FPR is plotted along the x axis

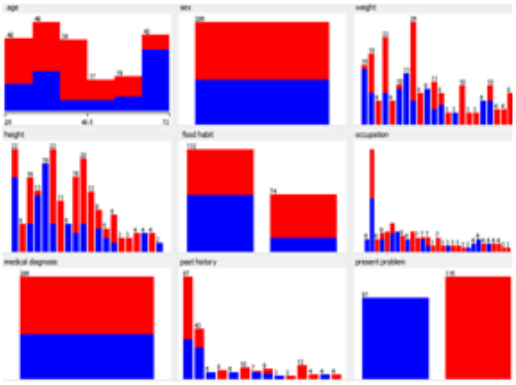

Figure 1: Data Distribution

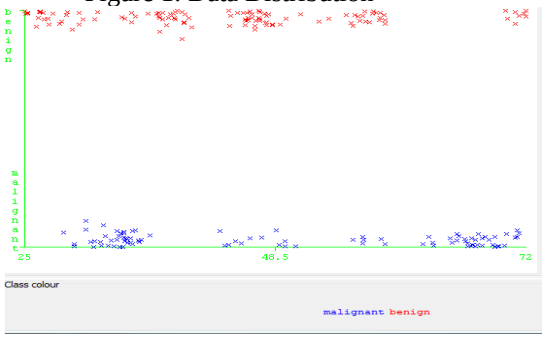


Figure 2: Age wise Classification of Breast cancer data

The experimental results of basic classifiers are discussed in this section. Breast cancer data contains tumors which represents the severity of the disease. The two kinds of tumors are benign and malignant. To classify them correctly from the algorithms J48, Cart, and AD Tree are given the table 2, 3 and 4. The classification accuracy of four algorithms J48, CART, and AD Tree are observed from the tables 2, 3 and 4 via values of weighted average, which is available in the last row of each table. The table 5 depicts the error report of the three classification algorithms. Table 6 and figure 3 shows the weighted average accuracy of the classification algorithm for the breast cancer data. The figure 4 represents the comparison of the J48, CART and ADTree classification algorithms based on the table 7 values.

Table 2: Results of CART

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area |
|---|---|---|---|---|---|---|
| Malignant | 0.989 | 0.0174 | 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 3: Results of AD Tree

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area |
|---|---|---|---|---|---|---|
| Malignant | 0.968 | 0.017 | 0.978 | 0.968 | 0.973 | 0.987 |
| Benign | 0.983 | 0.032 | 0.975 | 0.977 | 0.977 | 0.987 |
| Weighted Average | 0.977 | 0.025 | 0.977 | 0.977 | 0.977 | 0.987 |

International Journal of Data Mining Techniques and Applications
Volume 5, Issue 1, June 2016, Page No.43-49
ISSN: 2278-2419

Table 4: Results of J48

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area |
|---|---|---|---|---|---|---|
| Malignant | 0.978 | 0.017 | 0.978 | 0.978 | 0.978 | 0.99 |
| Benign | 0.983 | 0.022 | 0.983 | 0.983 | 0.983 | 0.99 |
| Weighted Average | 0.981 | 0.02 | 0.981 | 0.981 | 0.981 | 0.99 |

Table 5: Error Reports

| STATISTIC | CART | AD TREE | J48 |
|---|---|---|---|
| Kappa statistic | 0.9705 | 0.9524 | 0.9606 |
| Mean absolute error | 0.0239 | 0.1666 | 0.0247 |
| Root mean squared error | 0.1201 | 0.2222 | 0.1188 |
| Relative absolute error | 4.8438 | 33.8907% | 5.016% |
| Root relative squared error | 24.185 | 44.8179% | 23.9293% |

Table 6: Accuracy by Weighted Average

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Time |
|---|---|---|---|---|---|---|---|
| Malignant | 0.985 | 0.014 | 0.985 | 0.985 | 0.985 | 0.985 | 0.01 |
| Benign | 0.977 | 0.025 | 0.977 | 0.977 | 0.987 | 0.987 | 0.02 |
| Weighted Average | 0.981 | 0.02 | 0.981 | 0.981 | 0.981 | 0.981 | 0.04 |

Table 7: Performance accuracy of algorithm

| Classifiers | Accuracy (%) | Time (in Seconds) |
|---|---|---|
| CART | 98.50 | 0.01 |
| AD TREE | 97.70 | 0.02 |
| J48 | 98.10 | 0.04 |

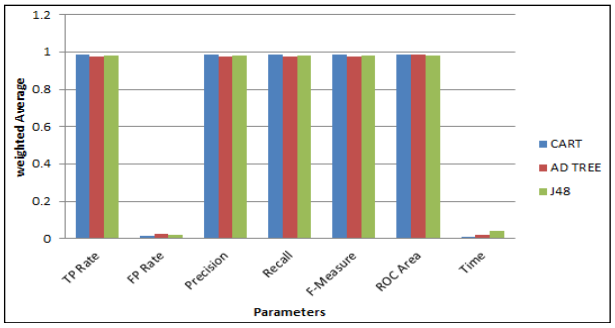


Figure 3: Weighted average of various parameters

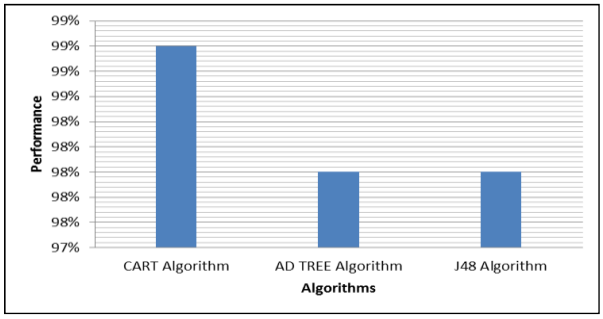


Figure 4: Performance comparison of Algorithms

## V. CONCLUSION

An important challenge in data mining is to build precise and to find efficient classifiers for medical applications.In this research work, the accuracy of classification techniques is evaluated based on the selected attributes of mammogram images. Specifically, three popular data mining methods J48, AD Tree and CART has been used for the analysis. The classifiers J48 have accuracy of 98.1 %, ADTreehave 97.7% and highest accuracy value 98.5 % is found in CART. It is observed that from the obtained results, the CARTalgorithm performs well for classifying mammogram images. The algorithm ADTree gives low performance and J48 act as intermediary.In future, the other classification algorithms are utilized for the analysis of the same mammogram images to predict their performances.

## References

[1] JaiminiMajali,RishikeshNiranjan,VinamraPhatak,OmkarTadakhe,"Data MiningTechniques for Diagnosis And Prognosis of Cancer", Int. Journal of Advanced Research in Computer and Communication Engg., Vol. 4, Issue 3, 2015, pp. 613-614.

[2] K.R.Lakshmi, M.Veera Krishna, S.Prem Kumar, "Performance Comparison of Data Mining Techniques for Prediction and Diagnosis of Breast Cancer Disease Survivability", Asian Journal of Computer Science and Information Technology, Vol. 3, 2013, pp. 81 - 87.

[3] Joshi, Miss Jahanvi, and Mr.RinalDoshi, Dr.Jigar Patel. "Diagnosis And Prognosis Breast Cancer Using Classification Rules", Int. Journal of Engineering Research and General Science, 2014, Vol. 2, Issue 6, pp. 315-323.

[4] VasanthaM., &Bharathy, V. S. "Evaluation of Attribute Selection Methods with Tree Based Supervised Classification-A Case study with Mammogram Images", Int. Journal of Computer Applications, Vol. 8, No. 12, 2010, pp. 35.

[5] VikasChaurasia, Saurabh Pal, "A Novel Approach for Breast Cancer Detection using Data Mining Techniques",

Int. Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 1, 2014, pp. 2464.

[6] Zarei, S., Aminghafari, M., HakimehZali, "Application and comparison of different linear classification methods for breast cancer diagnosis", International Journal of Analytical, Pharmaceutical and Biomedical Sciences, Vol. 4, Issue2, 2015, pp. 123-128.

[7] Y.Ireaneus Anna Rejani, Dr.S.ThamaraiSelvi, "Early Detection Of Breast Cancer Using Svm Classifier Technique", Int. Journal on Computer Science and Engineering, Vol. 1, Issue 3, 2009, pp. 127-130.

[8] Rajesh,k., Dr.SheilaAnand,"Analysis of SEER Dataset for Breast Cancer Diagnosis usingC4.5 Classification Algorithm", IJARCCE,Vol.1, Issue 2, 2012, pp. 2278-1021.

[9] Vanaja, S., K. Rameshkumar, "Performance Analysis of Classification Algorithms on Medical Diagnoses-a Survey", Journal of Computer Science, Vol. 11, 2015, pp. 32-33.

[10] Chandrahasan, R. Kalaichelvi, et al. "An Empirical Comparison of Boosting and Bagging Algorithms", International Journal of Computer Science and Information Security, Vol. 9, Issue 11, 2011, pp. 147-152.

[11] Sugimoto, Masahiro, Masumi Takada, and Masakazu Toi, "Comparison of robustness against missing values of alternative decision tree and multiple logistic regressions for predicting clinical data in primary breast cancer",IEEE, 2013, pp. 3054-3057.

[12] Nithya, R., and B. Santhi, "Decision tree classifiers for mass classification", Int. Journal of Signal and Imaging Systems Engineering, Vol. 8, No. 1-2, 2015,pp. 39-45.

[13] J.S.Saleema, N.Bhagawathi, S.Monica , P.DeepaShenoy, K.R.Venugopal and L.M.Patnaik, Int. Journal on Soft Computing, Artificial Intelligence and Applications, Vol.3, No. 1, 2014, pp. 9-10.

[14] Lavanya, D., and Dr K. Usha Rani. "Analysis of feature selection with classification: Breast cancer datasets." Indian Journal of Computer Science and Engineering, Vol. 2, No. 5, 2011, pp. 756-763.

[15] H.S.Hota, "Diagnosis of Breast Cancer Using Intelligent Techniques", Int. Journal of Emerging Science and Engg.,ISSN: 2319–6378, Vol.1, Issue-3, 2013, pp.48-49.

[16] Li,J.,& Gramatica P., "Classification and virtual screening of androgen receptor antagonists", Journal of chemical information and modeling, Vol. 50, Issue 5, 2010, pp.861-874.

[17] Sumbaly R., Vishnusri N., &Jeyalatha, S. "Diagnosis of Breast Cancer using Decision Tree Data Mining Technique", International Journal of Computer Applications, Vol. 98, No.10, 2014, pp. 16-24.

[18] G. Sujatha1, Dr. K. Usha Rani, **"**Evaluationof Decision Tree Classifiers on Tumor Datasets", International Journal of Emerging Trends & Technology in Computer Science,Vol. 2, Issue 4,2013, pp. 422.

[19] Abdelghani Bellaachia and ErhanGuven, "Predicting breast cancer survivability using data mining techniques", Society for Industrial and Applied Mathematics, Vol. 58, Issue 13, 2006, pp. 1-4.

[20] Shajahaan, S. Syed, S. Shanthi, and V. M. Chitra, "Application of Data Mining Techniques to model Breast Cancer Data", International Journal of Emerging Technology and Advanced Engineering, Vol. 3, Issue 11, 2013, pp. 362-369.

[21] Shelly Gupta et al., Data Mining Classification Techniques Applied For Breast Cancer Diagnosis and Prognosis, Indian Journal of Computer Science and Engineering, Vol. 2, No. 2, 2011, pp. 188-195.

[22] Sujata Joshi, S. R. PriyankaShetty, "Performance Analysis of Different Classification Methods in Data Mining for Diabetes Dataset Using WEKA Tool", International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 3, Issue 3, 2015, pp. 1168 – 1173.

[23] Salzberg, Steven L., "C4. 5: Programs for machine learning by J. rossquinlan, Morgankaufmann publishers, inc.", Machine Learning,Vol. 6, Issue 3,1994, pp. 235-240.

[24] G.Sujatha, Dr. K. Usha Rani, "Evaluation of Decision Tree Classifiers on Tumor datasets", Int. Journal of Emerging Trends & Technology in Computer Science, Vol. 2, Issue 4, 2013, pp. 419-420.

[25] Aloraini, Adel, "Different Machine Learning Algorithms for Breast Cancer Diagnosis", International Journal of Artificial Intelligence & Applications, Vol. 3, Issue 6, 2012, pp. 21-30.