Prediction of Tumor in Classifying Mammogram images by k-Means, J48 and CART Algorithms

E.Venkatesan[1], T.Velmurugan[2]
[1]Research Scholar, [2]Associate Professor
PG and Research Department of Computer Science, D.G.Vaishnav College, Arumbakkam, Chennai, India
E-Mail: venkatelumalai12@yahoo.co.in, velmurugan_dgvc@yahoo.co.in

Abstract- The Breast cancer is one of the leading cancers for women in world countries including India. It is the second most common causes of cancer death in women. The high incidence of breast cancer in women has increased significantly in the last few years. Detecting cancer in the later stages, leads to very complicated surgeries and the chances of death is very high nowadays. Early detection of Breast Cancer helps in less complicated procedures and early recovery. Many tests have been found so as to detect cancer. Some of those tests are mammography, ultrasound etc. Mammography is a method that helps in early detection of Breast Cancer. But finding the mass and its spread from mammographic images is very difficult. Expert radiologists were needed for accurate reading of a mammogram image, and analyses have been for k-Means algorithm which helps for easy detection and extraction tumor area. The mammography image helps to provide some criteria in order to help the physicians to decide whether a certain disease is abnormal or normal. This research work is to identify the breast cancer tumor area and find its affected region by splitting the images into five clusters. The tumor area has been identified in the last cluster and classified with the help of decision tree algorithms J48 and CART.

Keywords-J48 Algorithm, CART Algorithm, Mammogram Images, k-Means Algorithm.

I. INTRODUCTION

Among the various diseases, the cancer is the second leading causes of death for women all over the world and more than 8%of women suffer from this disease during their lifetime. According to the latest world health organization (WHO) statistics, cancer causes around7.9 million deaths worldwide each year. It has been reported that one of four women in the world are affected with breast cancer. This is the most rampant disease in the world. In general the newly diagnosed ten cancers, one is breast cancer. Since the cause of breast cancer is still unknown, early detection is the key to reduce the death rate [3].Cancer is one of the most dread diseases and it has been studied for years. A new cancer is diagnosed every 2 minutes. Late detection of breast cancer causes the cancer cells to spread to other body parts and organs. This will result in complicated surgeries and also increases the chances of death. Because of these reasons, for the women after the age of 50 met with cancer were advised by doctors to conduct tests to detect it at its early stage. Mammography is one such test that helps to detect cancer at its early stage [1]. The Cancer is an abnormal, continual multiplying of cells. The cells divide uncontrollably and may grow into adjacent tissue or spread to distant parts of the body. Early detection of breast cancer through periodic screening has noticeably improved the outcome of the disease. The mass of cancer cells will eventually become large enough to produce lumps, masses, or tumors that can be detected. Tumor is uncontrolled growth of cells which can be either Benign or Malignant. Benign Tumors are not cancers. Benign tumors may grow larger but do not spread to other parts of the body. Malignant Tumors are cancers. Malignant tumors can attack and destroy nearby tissue and spread to other parts of the body. These tumor can be easily identified in mammogram because tumor part is highly bright (having high intensity) compared to other part (background) of the mammogram image [2].

Data mining (DM) has become a fundamental methodology for computing applications in the domain area of medicine. Evolution of data mining applications and its implications are manifested in the areas of data management in healthcare administrations, epidemiology, patient care and intensive care systems, significant image analysis to information extraction and automatic identification of unknown subjects. In the recent years, the data from several domains including banking, retail, telecommunications and medical diagnostics contains valuable information and knowledge which is often hidden. DM has various techniques such as Classification, Clustering, Prediction, Association Rules, Decisions Tress, and Neural Networks. [12]

Classification is a supervised Machine Learning technique which assigns labels or classes to different objects or groups. Classification is a two-step process. First step is model construction which is defined as the analysis of the training records of a database. Second step is model usage; the constructed model is used for classification. The classification accuracy is estimated by the percentage of test samples or records that are correctly classified. Classification is playing an important role in the field of data mining as well as in the studies of machine learning, neural network, statistics and many expert systems over many years. Different classification algorithm has been successfully implemented in various applications. Among them some of the popular implications of classification algorithms are scientific experiments, credit approval, weather prediction, fraud detection, medical diagnosis, image processing, target marketing and lots more. Medical dataset such as cancer dataset contains large number of gene expression values [10].In the recent years medical data classification especially cancer data classification caught a huge interest amidst the researchers. It is necessary to discuss some of the research work in the area of cancer data is must. The significance of this work can be shown by the recent researches on cancer classification by many researchers. The method of clustering organizes the objects into groups based on some feature, attribute and characteristic. There are two types of clustering, one supervised and the other is unsupervised. In supervised clustering, cluster criteria are specified by the user. In unsupervised type, the cluster criteria are decided by the clustering system itself. The k-Means algorithm is a popular unsupervised data clustering algorithm. The k-means algorithm is a simple interactive method to partition of given dataset into a user specified number of clusters; k-Means clustering is suitable for biomedical image as the number of clusters is usually known for images of particular regions of the human anatomy. Clustering is a vital technique used in pattern recognition and region based segmentation approaches. Group of objects which have same similar features between them and are dissimilar to the objects belonging to other groups is defined as cluster. Clustering can be defined as an unsupervised learning process of organizing objects into groups whose members possess some similar feature [5].The segmenting the breast tumor region is frequently difficult due to the wide variation of tumor size, shape and relative location within the mammogram. In the meantime, the low contrast and poor definition of tumor edge, and the surrounding fatty tissues and veins are also contributing to the difficulty of tumor segmentation [11]. One of the affected mammogram images of abnormal and normal are shown in figure 1 and figure 2

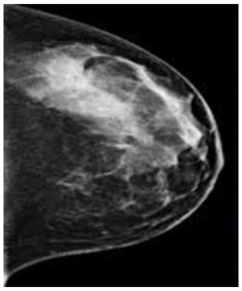

Figure1: Abnormal image.

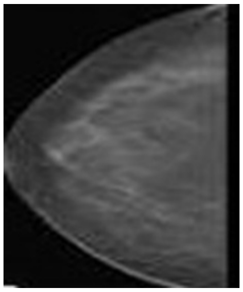

Figure 2: Normal image

This research paper is organized as follows. Section 2 discusses about methods and materials used for this research work namely k-Mean algorithm, Decision tree algorithms j48 and CART. The results of Breast cancer tumor extraction, tumor area value classification of decision tree performance are illustrated in section 3. Finally, the work is concluded in section 4.

II. MATERIALS AND METHODS

The detection of blocks in breast cancer tumor area carries various methods in the last decade. The k-Means algorithm is a well-known algorithm suited for the mammogram breast cancer tumor area extraction. Breast cancer tumor is a collection of growth of abnormal cells in the breast and the area closer to it. Many different types of breast tumors are exist. The Benign tumors which are an early stage it can't spread to the other regions. But the Malignant spreads in other regions and other parts of the body. There are a number of classification algorithms that has been proposed by several researchers in the field of classification applications. They used these algorithms to predict classification of tumor area by means of finding pixels in mammogram images data. They selected classification algorithm to find the most suitable one for predicting cancer. To classify tumor area pixel values with high accuracy via supervised learning algorithms CART, J48 in most of their research work. The data mining methods proposes many exploratory data analysis, statistical learning, machine learning and database analysis for the real world problems. This research work is carried out to identify and analyze the breast cancer tumor in mammogram images and find its affected region perfectly [13]. The classification of tumor area by decision tree algorithms j48 and CART and compare the results based on the performance. Data pre-processing is performed in this research work by WEKA tool for model the tumor area pixels. WEKA is an open source data mining software mainly used for algorithms performance analysis in data mining.

a) The k-Means Algorithm

The k-means algorithm is a simple and most used partition based clustering algorithm used for many researches in the current world. The k means algorithm is an interactive technique which is used to split an image into k clusters. In statistics and machine learning, k-Means clustering is a method of cluster analysis which can partition n observations into k cluster, in which each observation is in the right place to the cluster with the adjacent mean. The k-Means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to categorize a given data set through a certain number of clusters. There are many algorithms for clustering the chosen datasets.

The k-means clustering is a popular method used to divide n patterns $\{x_1, \ldots, x_n\}$ in d dimensional space into k clusters(assume k clusters). The result is a set of k centers, each of which is located at the centroid of the partitioned dataset. This algorithm can be summarized in the following steps:
Step 1: Give the number of cluster value as k.
Step 2: Randomly choose the k cluster centers
Step 3: Calculate mean or center of the cluster
Step 4: Calculate the distance between each pixel to each cluster center
Step 5: If the distance is near to the center then move to that cluster.
Step 6: Otherwise move to next cluster.
Step 7: Re-estimate the center.
Step 8: Repeat the process until the center doesn't move.
Many Clustering algorithms prove their efficiency in different fields. The k-Means algorithm is the widely used algorithm in all domains [9].

b) 2.2 Decision Tree Algorithm

A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically, each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. Decision trees are powerful classification algorithms that are becoming more and more popular with the growth of data mining in the field of information systems. Popular decision tree algorithms include Quinlan's ID3, C4.5, C5 and CART algorithm [6].

c) 2.3 CART Algorithm

CART (Classification and Regression trees) was introduced by Breiman in 1984. 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 building the decision tree and has in-built features that deal with missing attributes. CART is unique from other Hunt's based algorithms as itis also used for regression analysis with the help of their egression trees tumor area classification [7].

d) 2.4 J48 Algorithm

J48 algorithm is open source algorithm in data mining. J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. J48 classifier is the most popular tree classifier till today. WEKA classifier package has its own version of C4.5 classifier known as j48 and it is used in WEKA platform [8].

e) 2.5 Performance Measures

The various formulas used for the calculation of different measures are discussed below. Precision 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 means False 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 TN stands for True Negative

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

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

Specificity is the percentage of positive records classified correctly 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.

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

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

III. EXPERIMENTAL RESULTS

Mammogram images are analyzed by many of the peoples for different applications. This research work mainly focuses on the analysis of mammogram images to identify the tumor area. Totally 30 images are taken for analysis in this work. First, the images are clustered into 5 groups by k-Means clustering algorithm. In each and every cluster, the number of pixels are calculated which are affected by calcifications. The affected area of tumor is identified in the final stage of the clustering work. The clustering process is carried out to identify the tumor, which is done by k-Means algorithm. After processing all the 30 images, the identified tumors are verified by the classification algorithms. The k-Means algorithm source code was written in MATLAB software. After the clustering process, for classification, the classification algorithms J48 and CART are implemented in WEKA software. The output of WEKA tool is given figure 3 and 4.

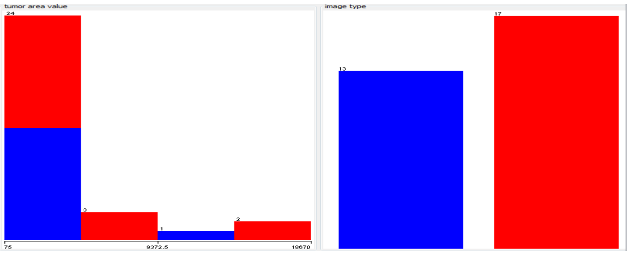

Figure 3: Preprocessing Output

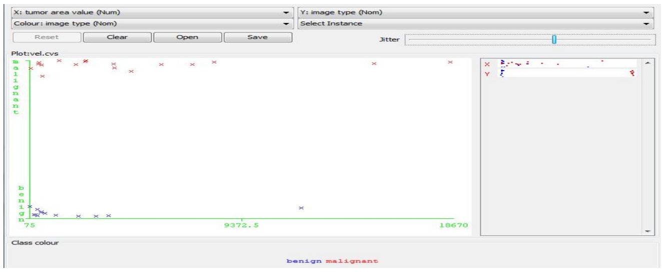
X: tumor area value (Num)
Y: image type (Nom)
Colour: image type (Nom)
Select Instance
Plot:vel.cvs

Figure 4: Classification Output

Table 1: Results of classification images of Pixel Values

No of Total Images	Tumor area Pixels Value between 75 to 4723.75	Tumor area Pixels Value between 75 to 4723.75	Tumor area Pixels Value between 9372.5 to 14021,25	Tumor area Pixels Value between 14021.25 to 18670
30	24 image both benign and malignant	3 image malignant	1 image benign	2 image malignant

a. Data Set

The data set used for this research work is described in this section. The mammogram images are taken from the Digital Database for Screening Mammography repository, which is available at the link: http://marathon.csee.usf.edu/ Mammography/ Database.html, This DDSM (Digital Database for Screening Mammography) data set has been extensively used by the research community. It is maintained at the University of South Florida for purposes of keeping it accessible on the web by the research community. The mammogram images in DICOM format is taken for analysis. DICOM means Digital Imaging and Communications in Medicine. It is an International Standard for medical images and related information of DICOM defines the formats for medical images. This DICOM image format applies into k-Means Algorithm to find breast cancer tumor area. The number of pixels in each cluster is tabulated and applied for classification.

b. The Proposed Method

Totally 30 images are considered for analysis for the identification of affected region of mammogram images. There are three types of images are available in the data set, which are benign, malignant and normal. All the three types are categorized by clustering and classification algorithms in this work. The MATLAB (R2008a) was used for writing the source code. The steps involved in clustering the Mammogram images by k-Means algorithm and Decision tree Classification algorithm j48 and CART are given below.

Step 1: Insert the original images as input.
Step 2: Convert the fetched mammogram DICOM format file into .JPG.
Step 3: Cluster the images.
Step 4: Find out ‘k’ value in the images by algorithm itself.
Step5: Get the clustering images of tumor area in every cluster.
Step 6: Apply Decision tree classification algorithm J48 and CART.
Step 7: Find the number of pixel values in each and every clusters.
Step 8: Predict the performance.

c. Results and Discussion

The input images of 30 persons are given in the figure 3. The input images are taken from different patients. The figure 3 contains normal, malignant and benign images. Some of the images are affected by cancer and some are not affected by cancer in the given set of images. The images are taken from different age group patients. There are a number of classification algorithms that has been proposed by several researchers in the field of classification applications and investigated result of cancer tumor area. They used these algorithms to predict classification of tumor area by means of identifying pixel data. They selected classification algorithm to find the most suitable one for predicting cancer. The classification of tumor area pixel value data is done by supervised learning algorithms via simple CART, J48. The accuracy and efficiency of both the algorithms are analyzed and the experimental results of chosen classifiers are discussed in this section. The two kinds of tumors are Benign and Malignant are classified correctly from the training data set with the error rates and accuracy.

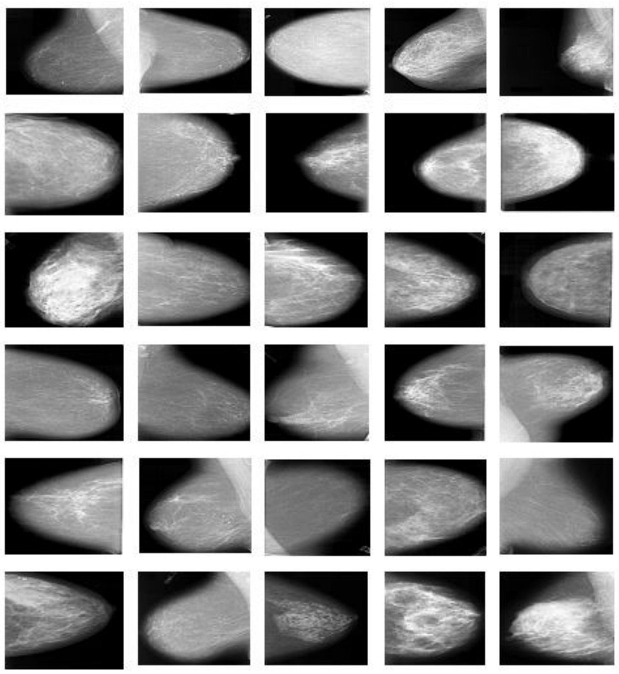

Figure 5: Mammogram input images

The accuracy of J48 algorithm is 96.341% and CART is 98.175%. The confusion matrix helps us to find the various evaluation measures like accuracy, recall, precision, F-Measure and ROC Area etc. The results of the performance of algorithms J48 and CART are given the table 3 and table 4 respectively. Table 5 gives the error rate of both the algorithms. Table 6 shows that the weighted average accuracy of the classification algorithms based on the various parameters for the breast cancer data and it also shown in figure 5. The Figure 6 represents the comparison of the J48 and CART classification algorithms based on the table 6 values.

Initially, the k-Means algorithm is applied in the 30 images and clustered based on the tumor area. The number of clusters in every image is considered as 5. After clustering by k-Means algorithm, the final results of the algorithm are given in figure 4. The number of pixels in each category of images is listed in table 2. In the table 2, C1 means cluster 1, C2 means cluster 2 and so on. The number of pixels in each and every cluster is given this table. The total number of pixels in all the five clusters is equal to the number of pixels in the original images.

It is easy to identify that some of the clusters are having very less number of pixels and a highest values of pixels are available in some clusters. Wherever the numbers of pixels are very less, that particular portion of the mammogram image has been affected by cancer tumor. The other areas are not very seriously affected.

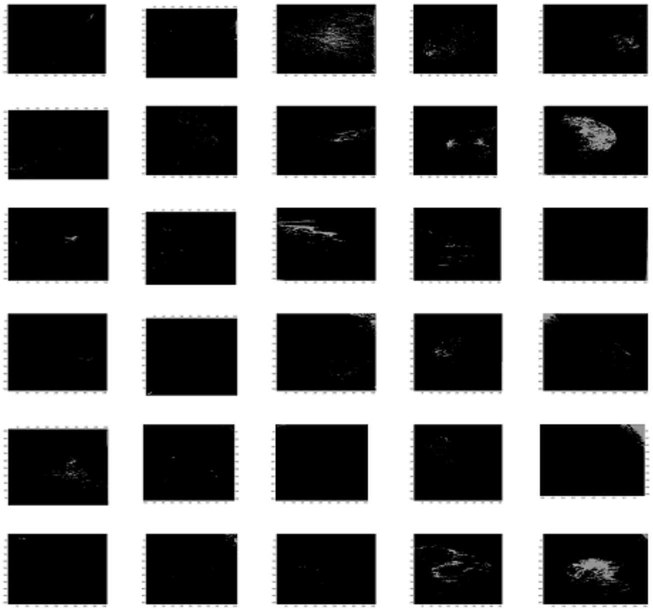

Figure 6: Results of tumor area in 5th cluster

Table 2: Results of clusters (Pixel Values)

Image No.	Total Pixels	No. of Pixels in C1	No. of Pixels in C2	No. of Pixels in C3	No. of Pixels in C4	No. of Pixels in C5
1	182228	26958	97418	34908	22726	218
2	69409	15614	20601	12465	19321	1408
3	69460	14314	10667	15309	16705	12465
4	179976	21963	46412	56422	52838	2341
5	100426	20073	17223	23928	35424	3778
6	154054	24493	28492	94576	6147	346
7	90884	13824	16916	10452	49148	544
8	107599	20862	26357	36268	21914	2198
9	108029	17956	12225	34193	40004	3651
10	112689	14389	11254	43486	43408	143
11	244105	112341	83776	41715	5452	821
12	202200	22243	41483	105460	32844	170
13	192015	22960	27011	89230	44658	8156
14	167448	27929	27514	70708	40092	1205
15	162420	31804	50036	76969	1439	2172
16	71339	22801	29653	14543	4249	93
17	47243	27407	10540	6829	2273	194
18	89871	27758	44873	10029	4495	2716.
19	164552	21129	52737	76106	13012	1568.
20	60223	19534	23688	10705	3776	2520
21	150360	24362	2874	81108	38173	3843
22	96831	24482	26873	11508	33776	192
23	205615	41213	137901	20466	5885	150
24	206756	22764	25702	123517	34449	324
25	223810	35060	29802	113119	39925	5904
26	153559	32240	53883	50098	17216	122
27	195905	31401	25557	116315	21518	1114
28	109477	4151	68664	30712	5875	75
29	192378	35523	46936	59566	43139	7214
30	191849	28179	37154	47259	60587	18670

Table 3: Results of J48

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Benign	0.919	0	1	0.919	0.958	0.987
Malignant	1	0.081	0.938	1	0.968	0.987
Weighted average	0.963	0.044	0.966	0.963	0.963	0.987

Table 4: Results of CART

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Benign	0.959	0	1	0.959	0.979	0.998
Malignant	1	0.041	0.968	1	0.984	0.998
Weighted average	0.982	0.22	0.982	0.982	0.982	0.998

Table 5: Error Reports

STATISTIC	J48	CART
Kappa statistic	0.9256	0.9626
Mean absolute error	0.0552	0.0205
Root mean squared error	0.1722	0.1287
Relative absolute error	11.1522	4.1367
Root relative squared error	34.5948	25.8523

Table 6: Accuracy by Weighted Average

Parameter	J48	CART
TP Rate	0.5	0.6
FP Rate	0.618	0.469
Precision	0.304	0.593
Recall	0.5	0.6
F-Measure	0.378	0.57
ROC Area	0.285	0.516
Accuracy	96.925	98.175

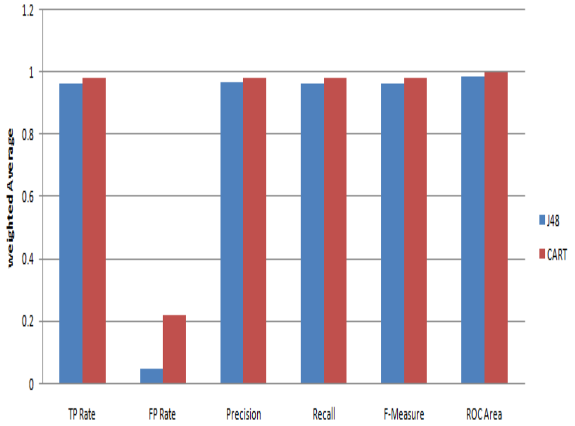

Figure 5: Weighted average of various parameters

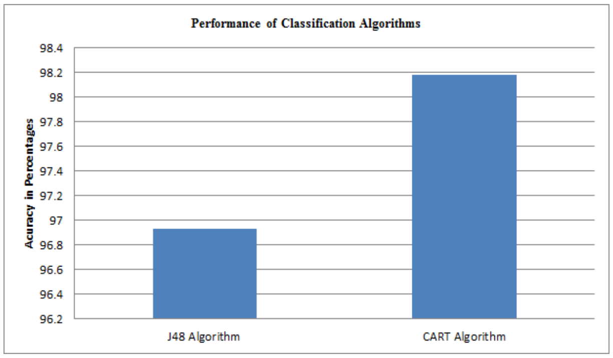

Figure 6: Performance of Algorithms

IV. CONCLUSIONS

The expert’s radiologists are needed for accurate reading of a mammogram images. The analysis has been done by k-Means algorithm help for easy detection and extraction of tumor area. The mammogram image helps to provide some criteria in order to help the physicians to decide whether a certain disease is in abnormal or normal. This research work is to analyze the breast cancer tumor area extraction and its affected region. The given 30 images are clustered by k-Means algorithm based on its pixel values by taking k value as 5. The tumor area was identified by clustering the images in the fifth cluster by k-Means algorithm. This research work evaluate the performances in terms of classification accuracy of J48 and CART algorithms using various accuracy measures like TP rate, FP rate, Precision, Recall, F-measure and ROC Area. In the implementation process, it is considered only the numerical values in the breast cancer tumor area pixel values data. The experimental results shows that the highest accuracy 98.175% is found in CART classifier and accuracy 96.341% is found in J48 algorithm. Based on the classification results of both the algorithms, the performance of CART is better than the J48 algorithm.

REFERENCES

[1] Lisha Sara Varughese, Anitha J,”A Study of Region Based Segmentation Methods FOR Mammograms”, International Journal of Research in Engineering and Technology, Vol.02, Issue 12, pp. 421-425, 2013.

[2] Pradeep, N., H. Girisha, and K. Karibasappa, “Segmentation and feature extraction of tumors from digital mammograms”, Computer Engineering and Intelligent Systems, Vol. 3 No.4, pp.37-46, 2012.

[3] Sindhuja, A., and V. Sadasivam,”Automatic Detection of Breast Tumors in Sonoelastographic Images Using DWT”, World Academy of Science, Engineering and Technology, Vol. 7, No.9, pp. 596-602, 2013.

[4] S. Syed Shajahaan, S. Shanthi, V. ManoChitra, “Application of Data Mining Techniques to Model Breast Cancer Data”, International Journal of Emerging Technology and Advanced Engineering, Vol.3, Issue 11, pp.362-369, 2013.

[5] Kawadiwale, Ramish B., and Milind E. Rane, “Clustering Techniques for Brain Tumor Detection”, Association of Computer Electronics and Electrical Engineers, Vol.5, No.49,pp.299-305, 2014.

[6] Shrivastava, Shiv Shakti, Anjali Sant, and Ramesh Prasad Aharwal, “An Overview on Data Mining Approach on Breast Cancer data”, International Journal of Advanced Computer Research, Vol. 3, issue 13, No. 4, pp.256-262,2013.

[7] Sujatha, G., and K. Usha Rani, “Evaluation of Decision Tree Classifiers on Tumor Datasets”, International Journal of Emerging Trends & Technology in Computer Science, Vol. 2, Issue 4, pp.418-423, 2013.

[8] N.Poomani, Dr.R.Porkodi,” A Comparison of Data Mining Classification Algorithms using Breast Cancer Microarray Dataset: A study”, International Journal for Scientific Research & Development, Vol.2, Issue 12, pp.543-547, 2015.

[9] Ramani, R., S. Valarmathy, and N. SuthanthiraVanitha., “Breast Cancer Detection in Mammograms based on Clustering Techniques-A Survey”, International Journal of Computer Applications, Vol.62, No.11, pp.17-21, 2013.

[10]Abidhasan., ”Evaluation of Decision Tree Classifiers and Boosting Algorithm for Classifying High Dimensional Cancer Datasets”, International Journal of Modeling and Optimization, Vol. 2, No. 2, pp.92-96,2012.

[11]Sheng-Wen Zheng., Jui Liu., Chen- Chung Liu,” A Random-Walk Based Breast Tumors Segmentation Algorithm for Mammograms”, International Journal on Computer, Consumer and Control, Vol. 2, No.2, pp.66-74,2013.

[12]Padmapriya, B., and T. Velmurugan, “A Survey on Breast Cancer Analysis Using Data Mining Techniques”, IEEE International Conference on Computational Intelligence and Computing Research, pp.1234-1237, 2014.

[13]Mahalakshmi, S., and T. Velmurugan, “Detection of Brain Tumor by Particle Swarm Optimization using Image Segmentation”, Indian Journal of Science and Technology, Vol. 8, Issue 22, pp.1-7, 2015.