ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 9(30), DOI: 10.17485/ijst/2016/v9i30/93851, August 2016*

Extraction of Cancer Affected Regions in Mammogram Images by Clustering and Classification Algorithms

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

Objectives/Backgrounds: The breast cancer has increased significantly in the last few years. It is one type of cancer and is the second deadliest disease in the world wide. Recently, cancer is diagnosed by various test such as mammography, ultrasound, etc. Mammography is used to breast imaging to help in detecting breast cancer. **Methods/Statistical Analysis:** The Mammogram images are taken for the analysis to find the tumor affected regions by data mining techniques in this research work. This work uses the Median filter method for noise removal and Gaussian filter for image enhancement of preprocessing the images. The k-Means algorithm, which is easily detected and extracts tumor area by means of intensity values by segmenting the mammography images. Two types of mammography images; normal and abnormal are given as input to the algorithms. After clustering the images by k-Means algorithm, the results found are classified by J48, JRIP, Support Vector Machines (SVM), Naïve Bayes and CART algorithms to verify the accuracy of the results based on its pixel values. **Findings**: The performance of taken classification algorithms is compared and find out the best classifier in terms of its accuracy, sensitivity and specificity. **Improvements**: In the future, the other classifiers and feature selection algorithms are applied to extract the mammography images. Also, it gives more than fifty images for analysis.

Keywords: Classification Accuracy, Classification Algorithms, k-Means Algorithm, Mammogram Images

1. Introduction

The breast disease is the second common disease in India. It is the most frequent kind of disease and affected one out of 22 women in India and is probable to suffer from breast cancer[1]. The deficiency of awareness initiatives, structured viewing and affordable treatment facilities continues to result in poor survival[2]. Breast tumor is the type of non-symptoms disease. Maximum number of peoples discovery this disease later on arriving the level of the high phase of distortion. Breast cancer phase is applied to explain the illness of malignant neoplastic disease, namely the position, the size, where it is spreaing and the range of its effect on another organs[3]. Actions for breast cancer is distinguished into two core forms, native and organized. Surgical procedure and energy are examples of native treatments, whereas chemotherapy and hormonal therapies are the examples of taxonomic treatments. Frequencies for the finest solutions, the two cases is to discourses are used together[4]. 40,000 women die in a year affected by this illness and one woman in every 13 minutes dying from this illness everyday[5]. The early detection of breast cancer is the effective way to shorten it. The best path to see the difference between

**Author for correspondence*

benign breast cancer and malignant type is without surgical biopsy is to recover the early diagnosis of accurate and reliable diagnosis procedure. The primary objective of this research study is to find cancer affected regions of patients to either a noncancerous kind group or a cancerous hateful group. Therefore, breast tumor diagnostic and predictive problems are mainly discussed in the course of study. Likewise, many researchers applied various techniques like data mining, statistics and many other areas to detecting the cancer in mammography images[6].

Knowledge Discovery in Databases (KDD) comprises Data Mining (DM) which is a general tool to forecast the effect of a illness using the past cases deposited within datasets[6]. Nowadays, the statistics from many areas, containing retail, investment, telecoms and medicinal diagnostics covers valuable info and information which is frequently obscured. Data Mining has numerous methods such as categorization, gathering, forecast, association rules and neural networks[7].

It is a great challenge to the researcher to employ the appropriate data mining algorithms such as clustering and classification to diagnose medical related problems. The exact method can be taken only after examining all the available classification methods and determining its performance in term of accuracy. Several inquiries have been borne out in the field of medical diagnoses by using a classification methodology. The most significant fact in medical diagnostic system is the accuracy of the classifier[8]. Most of the work uses the categorical data for prediction. Compartmentalization is a two-step process; exercise stage and the challenging stage. In the preparation stage, the predetermines statistics and the associated class label are applied to sorting. The tuples used in training stage is called training tuples. This is likewise known as supervised learning[8]. Classification and clustering are the two most basic techniques of Data Mining which are used in the area of medical science[4]. The uncontrolled growth of tumor in the breast tissues is the chief reason of breast malignant neoplastic disease. The neoplasm is a typical growth of tumor that can be either kind or malicious. They are the noninvasive while malicious tumors are cancerous and feast to other regions of the physical structure. The initial diagnosis and action helps to stop the feast of malignant neoplastic disease. Approximately 90% of breast malignant neoplastic disease is occurring due to the genetic abnormalities that occurs because of the aging process and the ducts 5-10% of cancers are due to an abnormality which is inherited from the parents[9]. The cancer investigation is usually scientific and/or organic in nature, information driven statistical investigation has become a shared compliment. One of the most challenging and interesting tasks is to anticipate the result of a disease anywhere to grow information mining applications.

Mammogram image's exposure can be taken out of by applying a diversity of methods in the literature. For popular action of the patient, the breast cancer is discovered in the starting phase and therefore the patient can be recouped rapidly. Various techniques used in breast cancer recognition are defined below[10]. In[11] organized, described the clustering approach based on the evaluation among the different classifiers such as decision tree (J48), Multilayer Perception (MLP), Naive Bayes (NB), Sequential Minimal Optimization (SMO), and Instance Based for K-Nearest neighbor (IBK) on all the three different databanks of breast malignant neoplastic disease is better classifiers analysis.

A research study was done by[12] in their work. In this inquiry, they compare the presentation standard of 5 supervised knowledge classifiers such as the techniques should be practical on micro-array information to invent techniques that should calculate the incidence of the neoplasm. Though, the correctness of techniques differs according to the organization techniques used. Recognizing the best classification techniques among all accessible is a stimulating job. This work which is presented here, they had produced a inclusive relative examination of fourteen different classification techniques and their working has been assessed by using three different cancer data banks.

A research work done by[14] they present clustering algorithm to study digital mammogram. They find the Probabilistic clustering algorithm performed better than hierarchical clustering algorithm. Another research work carried out by[15] is based on Digital mammography which is the mutual method or initial breast cancer detection. Since the manual examination is very slow, they find that the automated examination of the imageries is very significant. Nowadays, for radiologists, only eight slides of mammography images are permitted; it is inconsistent

and very costly. In this paper, they analyzed two classifiers with feature extraction and conclude with the best accuracy classification[15].

Another research carried out by[16]. In this work, the processor based analysis and classification system can decrease avoidable biopsy. These type of learning examines a novel advance to the organization of mammogram imagery founded on pixel strength mean structures. Future technique for the organization of regular and irregular (cancerous) design is a 2-step procedure. Initial step is feature removal. Strength based features are removed from the digital mammograms. Next step is the classification process, which is the differentiation between regular and irregular design. With these discussions about the basic concepts of DM techniques, next the structure of this article is elaborated. Section 2 deals about materials and methods used for this research work, namely k means algorithm for segmentation of cancer area, and classification algorithms such as J48, CART, SVM, Naïve Bayes and JRIP to find accuracy. The experimental results are illustrated in the Section 3. Finally, this work concludes with its innovative information in Section 4.

2. Materials and Methods

The objective of this work is to analyze totaling 50 mammography images for breast cancer disease, detection of the affected region of mammogram images tumor area. There are two types of images are available in the data set, which are benign and malignant. These two types are given as input to the pre-pressing techniques via median filter and then the mammogram image enhancement is done by using the Gaussian filters. The k-Means techniques is used for division of the cancer area by clustering and classification algorithms to find accuracy in this work. The MATLAB (R2008a) was used for writing the source code. The steps are involved in this process, in clustering the Mammogram images by this algorithm for prediction and classification algorithm to find the accuracy. The description about the methods used for clustering and classification algorithms Naive Bayes, JRIP, SVM, J48and CARTare given below.

2.1 The k-Means Algorithm

This algorithm is a humble and most used separation based clustering algorithm used for many researches in the current world. The k-means algorithm is a repetitive technique which is used to tear a cancer affected image into k clusters. In statistics method and machine learning method, this clustering is a method of group approach which we can divider n number of comments into k cluster, in which each number of observations is in the correct place in then number of clusters with the adjacent mean. This is the best humblest unsupervised learning techniques which are used to solve the well-known gathering problem. The procedure follows a humble and easy method to classify a given data set through a sure number of groups. There are plenty of algorithms used for the chosen data set. To divide n number of patterns $\{x_1,\ldots,$ Xian$\}$ in d-dimensional space into k- groups (assume k clusters) using this clustering algorithm. The result set of this will be the set of k centers, in which each of the clusters is positioned at the centroid of the divided dataset. It is outlined in the given succeeding steps:

Step 1: The number of cluster values for k should be entered.

Step 2: After that, choose the k group centers randomly.

Step 3: Mean or center of the cluster is calculated.

Step 4: The distance between each pixel to each cluster center is then calculated.

Step 5: If the distance is near to the center, then we should move to that cluster.

Step 6: Else move to following group.

Step 7: Re-evaluation the cluster midpoint.

Step 8: Recap the same process till the middle doesn't move.

Many clustering algorithms show their efficiency in different field. This technique is the extensively used tech-

nique in all domains[17]. Particularly, in the medical field, the efficiency of k-Means proved in many applications through its performance.

2.2 CART Algorithm

CART (Classification and Regression Trees) is developed by Bremen in 1984. It builds all the classifications and regression trees. It is also based on Hunt's replica of a decision tree building and can be implemented in sequence. It uses GINI index splitting amount in choosing the intense characteristic. Prune is used in CART by using a helping of the preparation data set. CART uses both numeric and definite attribute for structure the decision tree and has inbuilt applications that deal with lost attribute. CART is single from other Hunt's based techniques as it is also used for with the assist of regression tree to find the regression analysis. For forecasting the dependent variable from a set of forecaster variables over a given phase of time by means, the regression analysis feature. The CART approach is another to the established method for prediction[18].

2.3 J48 Algorithm

J48 algorithm is the open source method in data mining techniques. J48, similar to C4.5 decision tree which is used to create the binary tree. The most commonly used approach to find the classification problem is the decision tree approach. With the help of this technique, a binary tree is built to model the classification process. J48 is the most famous tree classifier till date. WEKA package has its version of C4.5 classifier called as J48 and it is used in the same platform[19].

2.4 Naive Bayes Algorithm

It implements a probabilistic Naïve Bayes classifier. Naive means provisional liberty among the attributes of features. The "naive" assumption significantly reduces the computational difficulty to simple development of probabilities. It handles numeric attribute using supervised discretization and uses kernel thickness estimators that will get better the performance. It wants only a small set of preparation, data to develop exact limit estimations because it requires only the computation of the frequencies of attribute and attribute outcome pair in the training data set[20].

2.5 Support Vector Machines (SVM)

Support Vector Machines (SVMs) are a new pattern recognition tool supposedly found in Vapnik's statistical learning theory (Vapnik, 1998). It is initially intended for binary classifying; employ administered culture to find the optimal separating hyper plane between both the groups of data. Having found a plane, it can then forecast the organization of an unlabeled instance by asking on which part of the unraveling plane for the examples. It acts as a linear classifier in a high dimensional characteristic space originated by a ledge of the unique input space, the resultant classifier has in common nonlinear in the given input space and it achieved very good simplification performance by exploiting the boundary between the classes[21].

2.6 JRIP

JRIP is a quick technique for learning "IF-THEN" regulations. The JRIP algorithm was proposed by William Cohen (1995) and such as decision trees regulations learning technique are famous because the knowledge demonstration is very easy to construe repeated. The abbreviation for RIPPER is Incremental Pruning to Produce Error Reduction is one of the essential and most famous techniques. Classes are inspected in growing size and an early set of rules and regulations for the class is generate using incremental reduced-error pruning[22].

3. Experimental Results

The experimental work have preprocessing the image by median filter methods, image enhancement by Gaussian filter method, processing by k-Means algorithm and classification by classification algorithms for its accuracy of results. This technique used to identify those affected region of mammography and the source code was written in MATLAB software. The investigational results illuminate the several measures that have been used to evaluate the model for image extraction by clustering and classification. In this work, to assess the presentations in

terms of classification correctness by J48, CART, JRIP, Naive Bayes and SVM techniques using several accuracy actions like Sensitivity, Specificity and time comparison. After the clustering process, for classification, algorithms are implemented in WEKA software. The classification algorithm is trained and tested in10 times. Accuracy is measured and created using 10 fold cross confirmation method. Ten-fold cross confirmation is the normal way of computing the error degree of a learning system on a specific data set; for consistent outcomes, ten times the data set is executed by 10-fold cross-confirmation. In 10-fold cross confirmation method, the data set is arbitrarily subdivided into ten equivalent sized dividers. Among the dividers nine of them have been used as training set and the remaining one set is used as a test set[20].

3.1 Data Set

This research work uses 50 mammogram images of two types such as normal and abnormal patient. The mammogram images are taken from the Swami Vivekananda Diagnostic Centre (SVDC) Lab Centre in Chennai at Dwaraka Doss Goverdhan Doss Vaishnav College campus. An irregularity of the mammogram imageries was patented by SVDC head. The mammogram breast cancer imageries in DICOM format are taken for examination. Abbreviation of DICOM is Digital Imaging and Communications in Medicine. It is a global Standard for medical imageries and linked info of DICOM describes the formats for medical imageries. The DICOM format file supports the encapsulation of the info object description. This kind of imageries can hold images with patient information like age, gender, modality, description of study, date of imageries taken, imageries size and type of imageries etc. Regular and irregular data are preserved in this procedure. Figure 1 shows that a normal image of a mammogram.

3.2 Preprocessing

First, the input image is preprocessed by filtering techniques. In pre-processing, first the noise can be removed using the median filter method, and then the mammogram image enhancement is done by using the Gaussian filter method. Noise refers to the unwanted area of the mammogram images. The pre-processing has been divided into two phases. The first phase is noise removal, and second is enhanced. Out of the four filtering technique Mean filters, Median filter, Wiener filter and linear filter; this work uses median filter to remove the noises. The Mean filter is a simple sliding-window three-dimensional filter that eliminates the original image by filtering the unwanted areas of the image. The window is usually square, but can appear in any shape. Nonlinear digital filtering is done by the Median filter technique; it is used to eliminate the noise. The discount is a characteristic preprocessing stage to recover the results. It is one of the best filtering techniques for removing the noise. The second phase is the Mammogram enhancement and which is done using the Gaussian filter. The Gaussian filter is a fil-

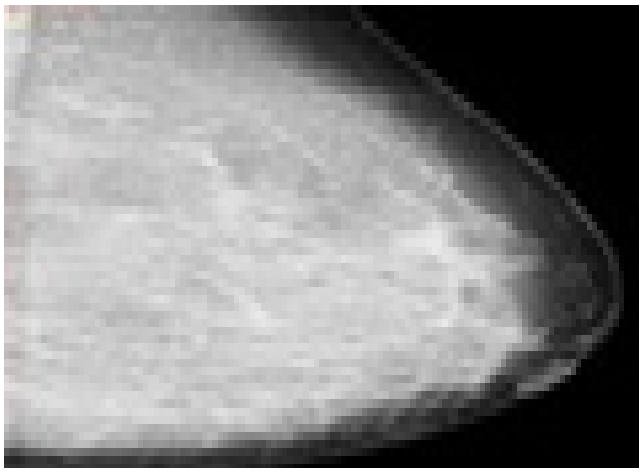

Figure 1. Input image.

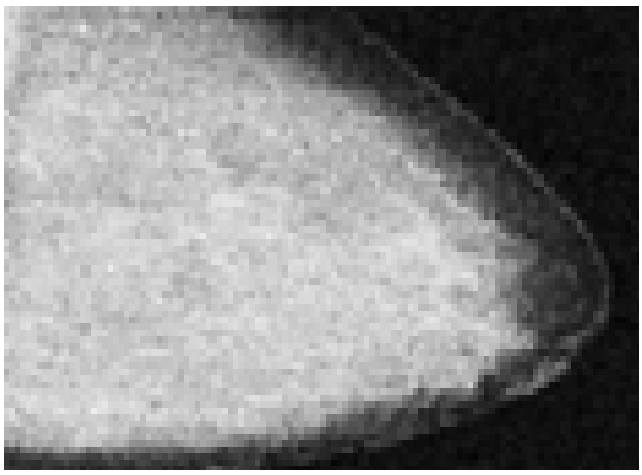

Figure 2. Noise removed image.

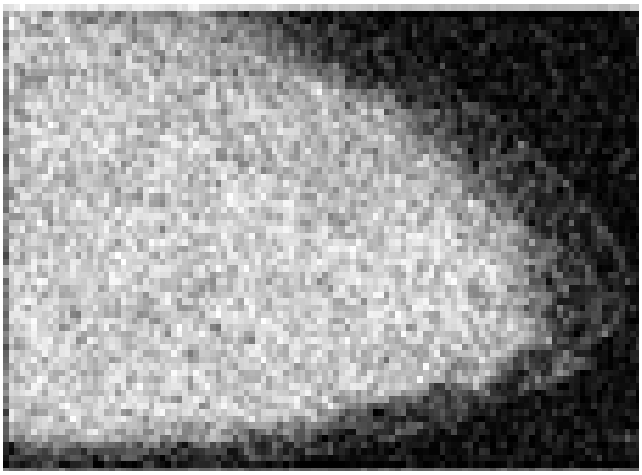

Figure 3. Enhanced image.

tering technique whose instinct reply is a Gaussian filters. Gaussian filters had the things about having no overshoot to a stage purpose input while reducing the increase and reduction time. It is used to enhance the mammogram image[23]. The output of Figure 1 after the noise removal is given in Figure 2 and the enhanced image is given in Figure 3.

3.3 The Proposed Method

In this research, the proposed method has three stages; pre-processing, image segmented and classified. Imageries pre-processing methods are necessary, in order to find the positioning of the mammogram, to eliminate the noise and to improve the excellence of the imageries. Before applying any imageries processing algorithm, it can be applied on pre-processing and uses the result image for processing. The segmentation is done by using the k-means clustering algorithm by giving k value as 5. Next, the classification algorithms are applied to find accuracy of results produced by clustering algorithm based on its pixel values[23]. The steps involved in this exploration work by clustering the Mammogram imageries by k-Means algorithm and Classification by JRIP, SVM, Naive Bayes, J48 and CART are agreed here.

Step 1: Insert the given input image for preprocessing.

Step 2: Change the gotten mammogram in DICOM files format into .JPEG format.

Step 3: Cluster given images.

Step.4: Find out 'k' value of the images by algorithm.

Step.5: Get the taken clusteredimagesand its pixel values in every cluster.

Step.6: Find the number of pixels in each and every cluster.

Step 7: Apply classification algorithm J48, JRIP, SVM, Naive Bayes and CART.

Step 8: Predict the performance.

3.4 Results and Discussion

The segmentation of imageries by the k-means clustering algorithm is carried out in this work. Initially, this algorithm is then applied to the mammogram images and clustered based upon the intensity. The number of clusters in every image is considered as 5. After clustering by this k-Means algorithm, the final results of the technique are given in Figure 4. It is easily identified that certain of the groups are having actual less amount of pixels based on intensity and maximum standards of pixels are obtainable in certain other groups. It is easy to identify that the 5th cluster has some differences in the intensity and which is the most affected portion of the cancer image. As per the suggestions of medical practitioners, the affected region is found in 5th cluster. After the clustering process, the number of pixels in each and every cluster is classified by classification algorithms.

There are a several number of algorithms of classification that has been projected by several scholars in the arena of classification and examined result of cancer disease. They are used to forecast organization by means of identifying breast cancer disease evaluation. They are the designated organization technique to find the greatest suitable one for forecasting tumor extraction area. The classification is completed by managed knowledge techniques via simple CART, J48, JRIP, SVM and Naïve Bayes. The accuracy and efficiency of all the algorithms

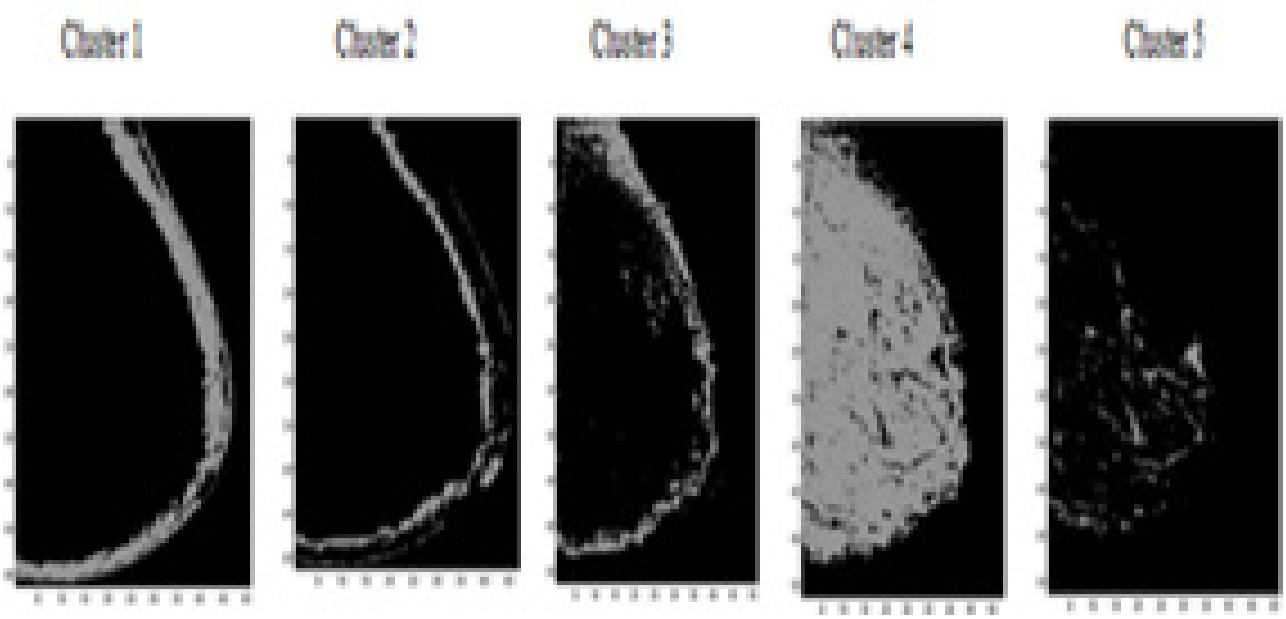

Figure 4. Results of k-means algorithm.

are analyzed and the results of these experiments of selected classifiers are deliberated in this segment. Benign and Malignant are two kinds of tumors; which are classified properly from the exercise data set with the accuracy, Sensitivity, Specificity. Table 1 shows that the values of sensitivity, specificity and accuracy. The run time for the algorithms is given in the last column of Table 1. The performance of all the five algorithms is shown in figure 5

Table 1. Error measures of various predictive models

Predictive Model	Accuracy	Sensitivity	Specificity	Time Milliseconds
JRIP	95.45	0.985	0.006	10
Naïve Bayes	93.61	0.926	0.361	30
SVM	92.40	0.92	0.401	20
J48	92.85	0.929	0.095	40
CART	93.75	0.938	0.104	30

based on its accuracy. Figure 6 show that the error measurements. Figure 7 is the performance of classification algorithms based on its implementation time. After that the Table 1, it is experiential that the shortest time which is around 10 milliseconds is found for JRIP algorithm compared with the other algorithms. Also, the accuracy of JRIP is significantly high. The results displays that the highest accuracy 95.45% is originate in JRIP classifier, second accuracy 93.75% is found in CART algorithm, third 93.61% is found in Naive Bayes algorithm, fourth 92.85% is originate in J48 algorithm and finally accuracy 92.40% is found in SVM algorithm.

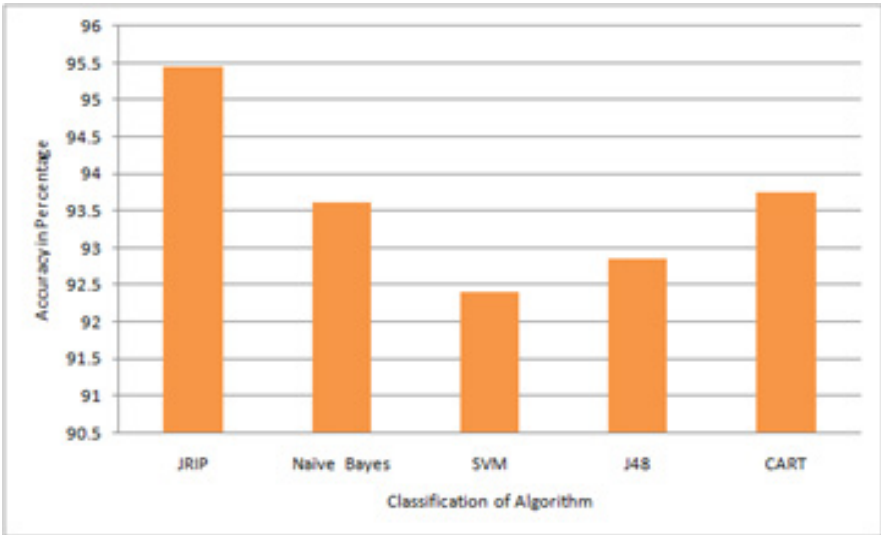

Figure 5. Performance of algorithms.

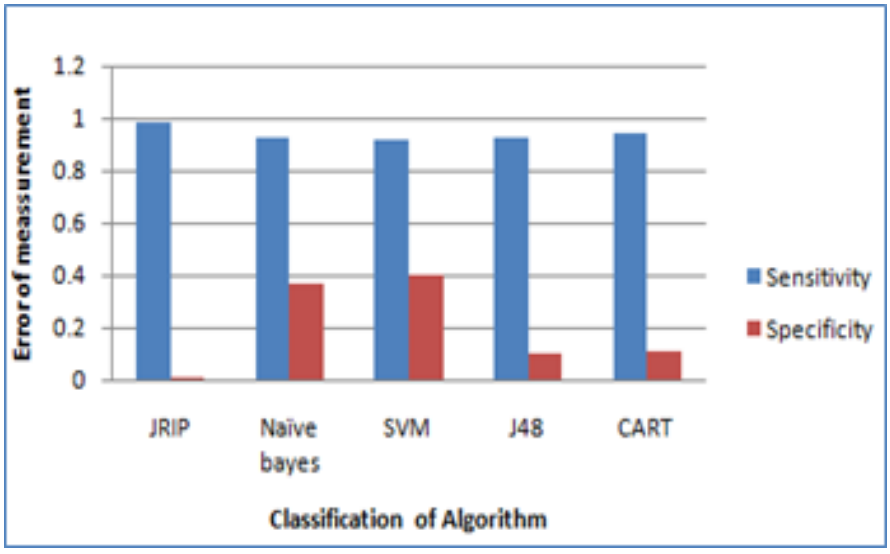

Figure 6. Performance of error measurement.

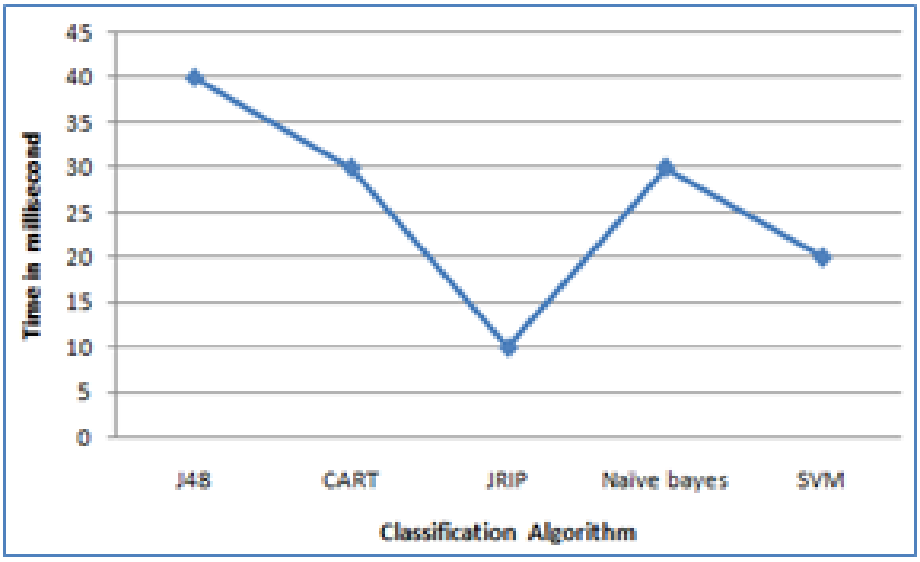

Figure 7. Performances in time.

4. Conclusions

The mammogram images of normal and abnormal are extracted by k-Means algorithm to help for easy detection of breast cancer tumor area identification. The major findings of this investigation work are to determine the breast cancer area via finding its affected region. Totally, 50 mammogram images are taken for analysis and clustered by k-Means algorithm based on its intensity by taking k value as 5. Before applying k-Means algorithm, the images are preprocessed by means of noise removal by median filter method and for the enhancement of images by Gaussian filter method. By k-Means algorithm, the tumor affected area was identified in the fifth cluster based on the intensity of images. This work evaluates the performances of classification accuracy in J48, JRIP, SVM, Naive Bayes and CART algorithms by means of various accuracy measures like sensitivity, specificity and time comparison. In the carrying out process, it is measured only the pixel values of final images produced by k-Means algorithm in the estimate of breast cancer. The results founded on the results of classifications of all these algorithms, performance of the JRIP is better than the SVM, CART, Naïve Bayes and J48. The other clustering and classification algorithms are applied to identify the cancer affected regions in future. Also, some of the feature selection algorithms are implemented for further analysis.

5. References

1. Boss RS, Thangavel K, Daniel DA. Mammogram image segmentation using rough clustering. International Journal of Research in Engineering and Technology. 2013; 2(10):66–77.
2. Shrivastava SS, 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. Karmilasari, Widodo S, Hermita M, Agustiyani NP, Hanum Y, Lussiana ETP. Sample k-means clustering method for determining the stage of breast cancer malignancy based on cancer size on mammogram image basis. International Journal of Advanced Computer Science and Applications. 2014; 5(3):86–90.
4. Zand HKK. A comparative survey on data mining techniques for breast cancer diagnosis and prediction. Indian Journal of Fundamental and Applied Life Sciences. 2015; 5(s1):4330–9.
5. Ronak S, Vishnusri N, Jeyalatha S. Diagnosis of breast cancer using decision tree data mining technique. International Journal of Computer Applications. 2014; 98(10):16–24.

6. Gupta S, Kumar D, Sharma A. Data mining classification techniques applied for breast cancer diagnosis and prognosis. IJCSE. 2(2):188–95.
7. Padmapriya B, Velmurugan T. A survey on breast cancer analysis using data mining techniques. IEEE International Conference on Computational Intelligence and Computing Research; 2014. p. 1234–7.
8. Vanaja S, Rameshkumar K. Performance analysis of classification algorithms on medical diagnoses a survey. Journal of Computer Science. 2015; 11(1):30–52.
9. 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; 1(2):72–7.
10. Karmilasari, Widodo S, Hermita M, Agustiyani NP, Hanum Y, Lussiana ETP. Sample k-means clustering method for determining the stage of breast cancer malignancy based on cancer size on mammogram image basis. International Journal of Advanced Computer Science and Applications. 2014; 5(3):86–90.
11. Gouda IS, Abdelhalim M, and Zeid MA. Breast cancer diagnosis on three different datasets using multi-classifiers. International Journal of Computer and Information Technology. 2012; 1(1):36–43.
12. Aruna S, Rajagopalan SPA, Nandakishore LV. An empirical comparison of supervised learning algorithms in disease detection. IJITCS. 2011; 1(4):81–92.
13. Nookala GKM, Pottumuthu BK, Mudunuri SB. Performance analysis and evaluation of different data mining algorithms used for cancer classification. International Journal of Advanced Research in Artificial Intelligence. 2013; 2(5):49–55.
14. Halawani SM, Alhaddad M, Ahmad A. A study of digital mammogram by clustering algorithm. Journal of Scientific and Industrial Research, 2012; 71:594–600.
15. Jog N, Pandey A. Implementation of segmentation and classification techniques for mammogram images. International Journal of Innovative Research in Science, Engineering and Technology. 2015; 4(2):422–6.
16. Nithya R, Santhi B. Mammogram analysis based on pixel intensity mean features. Journal of Computer Science. 2012; 8(3):329–32.
17. Ramani R, Valarmathy S, Vanitha NS. Breast cancer detection in mammograms based on clustering techniques- A survey. International Journal of Computer Applications. 2013; 62(11):17–21.
18. Sujatha G, Rani KU. Evaluation of decision tree classifiers on tumor datasets. International Journal of Emerging Trends and Technology in Computer Science. 2013; 2(4):418–23.
19. Halawani SM, Alhaddad M, Ahmad A. A study of digital mammogram by clustering algorithm. Journal of Scientific and Industrial Research. 2012; 71:594–600.
20. Karthikeyan T, Thangaraju P. Genetic algorithm based CFS and naive bayes algorithm to enhance the predictive accuracy. Indian Journal of Science and Technology. 2015; 8(26):1–8.
21. Khashei M, Eftekhari S, Parvizian J. Diagnosing diabetes type II using a soft intelligent binary classification model. Review of Bioinformatics and Biometrics. 2012; 1(1).
22. 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).
23. Akila K, Sumathy P. Early breast cancer tumor detection on mammogram images. Indian Journal of Computer Science and Engineering. 2015; 5(9):334–6.