© 2018 IJSRSET | Volume 4 | Issue 8 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Breast Cancer Detection in Mammogram Using FuzzyC-Means And Random Forest Classifier

Aleena Johny[1], Jincy J Fernandez[2]

[1]M.Tech Scholar, Department of Computer Science and Engineering, Rajagiri School of Engineering and Technology Kakkanad, Kochi, India

[2]Assistant Professor, Department of Computer Science and Engineering, Rajagiri School of Engineering and Technology Kakkanad, Kochi, India

ABSTRACT

Breast Cancer is one of the important reasons for death among ladies. Many research has been done on the diagnosis and detection of breast cancer using various image processing techniques. The proposed work deals with a technique for extracting the malignant masses in the mammography image for the earlier detection of breast cancer. The mammography images are complex, and also because of the noisy, inconsistent and incomplete data, several pre-processing techniques are used to enhance and make clear the targeted areas in the mammogram images. After segmenting the images into specific regions, based on its homogeneous characteristics, features are extracted which helps the classification more accurate. In this work, Fuzzy C-Means method is combined with Random Forest classifier to improve the accuracy.

Keywords: Pre-Processing, Segmentation, Post-Processing, Random Forest Classifier, Fuzzy C-Means.

I. INTRODUCTION

Nowadays, the usage of image processing techniques in medical science are increasing day by day for the better diagnosis and treatment of a patient. Medical imaging helps in revealing the internal organs, which is useful for the medical practitioners to do laparoscopic surgeries for viewing body parts without opening the body. The development of various medical imaging methods such as CT, MRI, PET, [1] helps the physicians to find the disease affected area.

Due to the inaccuracy of some image acquisition systems, noisy images are captured which affects the overall diagnosis of the patient. So pre-processing [2,3] plays a key role in image processing which improves the image quality by suppressing unwanted distortions in the captured image. Instead of processing the entire image which increases the complexity in terms of time and space, the image is divided into segments/parts based on few important characteristics [4]. The processing such as feature extraction are done after the extraction of region of interest (ROI). Breast cancer is a major cause of death among all cancers for women aged between 35 to 55 years and continues to be the leading cause of non-preventable cancer deaths. The proposed work deals with an approach for extracting the malignant masses in the mammography image for the detection of earlier breast cancer. The problem with mammography images are they are complex. Thus, image processing and feature extraction methods are used to assist radiologist for detecting tumour. Features extracted from suspicious regions in mammography images can assist medical doctors to find out the existence of the tumour at real time. Detecting breast cancer can be quite a challenging job. Specially, as most cancers is no longer a single disease but is a collection of multiple diseases. Thus, every cancer is dierent from every other. Also, the same drug may have dierent reaction on similar type of cancer. Thus, cancer vary from person to person. Depending on only one technique or one algorithm to detect breast cancer may not provide best result.

The structure of the paper is as follows: section II reviews various existing works in the detection of breast cancer. The proposed architecture is explained in section III. The experimental results are discussed in section IV. And finally conclusion is included in section V.

II. LITERATURE SURVEY

A. **Early Breast Cancer Detection using Mammogram Images: A Review of Image Processing Techniques**

Breast cancer is one of the most frequent cancers worldwide among female so that one in eight female is affected by means of this ailment at some point of their lifetime [5]. Mammography is the most high quality imaging modality for early detection of breast cancer in early stages. Because of poor contrast and low visibility in the mammographic images, early detection of breast cancer is a significant step to efficient treatment of the disease. Dierent computer aided detection algorithms have been developed to help radiologists provide an accurate diagnosis. The main focus is on image segmentation methods and the variables used for early breast cancer detection. Various pre-processing techniques discussed in the literature are image enhancement, noise removal, image smoothing, edge detection and enhancement of contrast. There are many segmentation algorithms discussed are thresholding, edge based segmentation, region based segmentation, clustering, classifier based segmentation and deformable model based segmentation.

B. **Detection of malignant tissue in mammography image using morphology based segmentation technique**

Breast cancer is the leading cause of the death among the women. Mammography is the best diagnostic technique for the breast cancer. But not all breast cancer can be seen by mammogram. Although breast cancer can be mortal, people have the highest chances to survive if cancer could be detected at the early stages. But there are certain limitations of the segmentation technique it is difficult to find the effected region perfectly. The proposed work deals with an approach for extracting the malignant masses in the mammography image for the detection of earlier breast cancer. The steps involved are removal of noise from the background information, thresholding and retrieving the largest region of interest, performing morphological operations and extracting the ROI and identifying the malignant masses from the image. Various pre-processing techniques used are Initial cropping, Intensity adjustment, CLAHE (Contrast limited adaptive histogram equalization), Noise reduction, Remove background information, Thresholding, Elimination of noise by locating connected segment larger area, Erosion, Perform subtraction, Removing connected components corresponding to the pectoral muscles.

C. **Breast Cancer Detection in Mammograms based on Clustering Techniques- A Survey**

Cancer is one of the most leading causes of deaths among the ladies in the world [6]. Doctor or radiologists can miss the abnormality due to inexperience in the area of cancer detection. Segmentation is very precious for doctor and radiologists to analysis the facts in the mammogram. Accuracy rate of breast cancer in mammogram relies upon on the image segmentation. The primary reason of clustering is to divide a set of objects into massive groups. The clustering of objects is based totally on measuring of correspondence between the pair of objects using distance function. Thus, result of clustering is a set of clusters, where object inside one cluster is further comparable to every other, than to object in every other cluster. The cluster evaluation has been widely used in numerous applications, inclusive of segmentation of medical images, pattern recognition, data analysis, and image processing. Clustering is also called data segmentation in some applications because clustering partitions huge data sets into groups according to their resemblance.

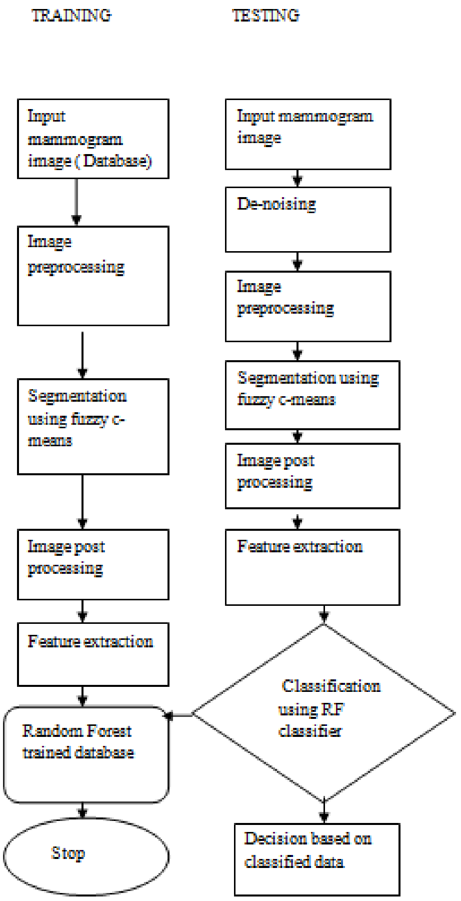

TRAINING
TESTING
Input mammogram image (Database)
Image preprocessing
Segmentation using fuzzy c-means
Image post processing
Feature extraction
Random Forest trained database
Stop
Input mammogram image
De-noising
Image preprocessing
Segmentation using fuzzy c-means
Image post processing
Feature extraction
Classification using RF classifier
Decision based on classified data

Fig.1 Proposed Architecture

III. PROPOSED SYSTEM

After the pre-processing step to get mammogram images of visually good quality by removing unwanted noises, the image is partitioned into several segments based on its homogeneous characteristics. Fuzzy C-Means clusting method is used. Then after extracting a set of features, Random Forest classifier is used to classify the affected area benign and malignant. The general architecture of the proposed work is given in the fig.1 and high level design in fig.2.

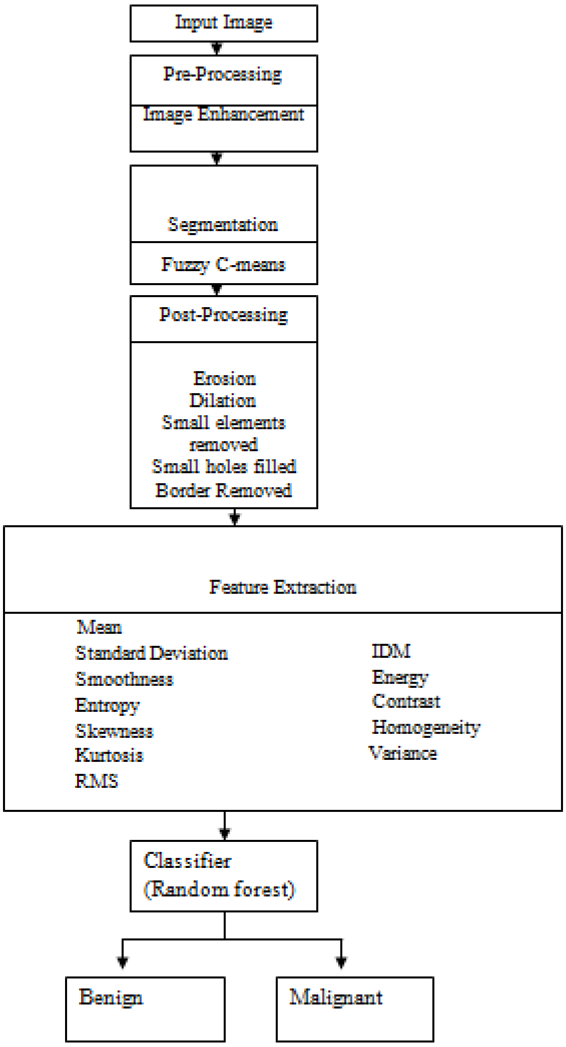


Fig.2 High Level Design

A. IMAGE ACQUISITION

Input a mammogram image of M×N size. Convert it into gray scale. Fig. 3 shows the input image.

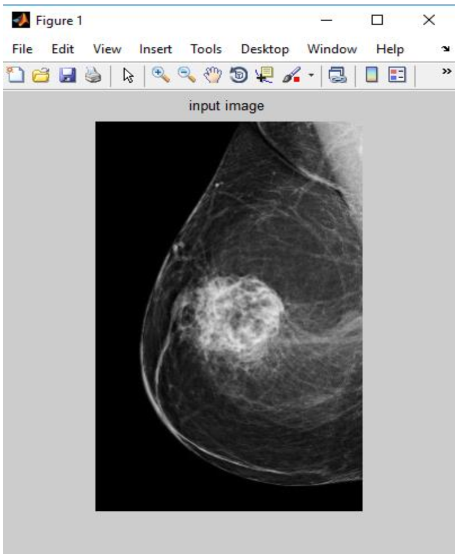


Fig.3 Input Image

B. IMAGE PRE-PROCESSING

Pre-processing improves the visual quality of mammogram image and makes them more suitable for analysis [7]. It is one of the basic steps that give high accuracy and also help to remove artifacts. Noise in the medical image negatively affect the aberration detection rate and accuracy. Different noise that reduce the quality of the image are adaptive noise, speckle noise, gaussion noise, impulsive noise, raison noise etc. Speckle noise is inherent property of ultrasound images where data got corrupted and it will blur the anatomical details hence it is an important task in image processing. Parts of our body such as breast, lungs, kidney contain soft tissues, in which lesion detection and boundary detection is an important part of ultrasound screening and diagnosing get affected by noise. So an adaptive median filter is used to remove the noise from the mammogram image. Adaptive Median Filter [8] is used to smoothen non repulsive noises in the absence of edge blurring and also retain edge information whenever there is a high density impulse noise. The pixels are differentiated by comparing each pixel with its neighbourhood, after setting size of the neighbourhood and threshold for the comparison. A pixel whose intensity not matched with majority of its neighboring pixels is considered to be an impulse noise and these noise pixel get replaced with the median pixel values in that neighbourhood. Fig.4 shows pre-processed image.

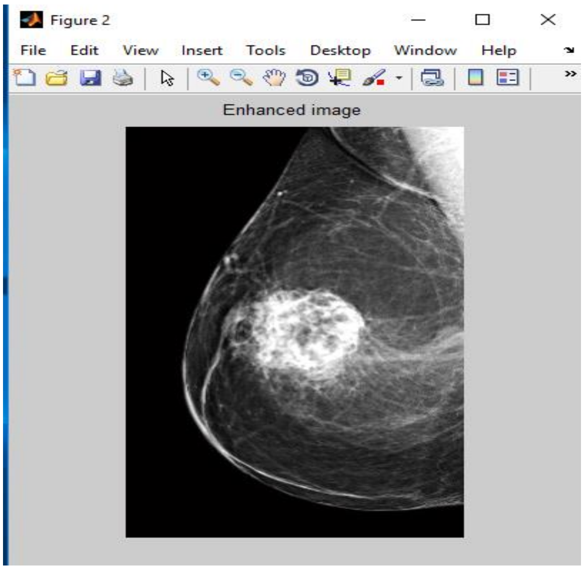


Fig.4 Pre-processed Image

C. IMAGE SEGMENTATION

Image segmentation [9, 13, 14] is used to identify suspicious areas (ROIs) in the mammogram images. With the use of segmentation, ROIs containing all abnormalities are obtained and suspicious lesions can be located from the ROIs. The image get divided into various segments by considering some features such as color, texture, brightness, contrast and gray level. The output of the process is irregularity. Fuzzy C-Means is used as the segmentation method. Fuzzy C-Means is a soft clustering method where each data points can be member of more than one class which are homogeneous in characteristics.

Fuzzy C-means [7] algorithm as follows:

1. One of the pixel is placed as constant from group of clusters.
2. Identify the distance among the pixels and calculate given dimension of input image.
3. Start the iterations and if probable iteration is reached, then stops the process and get segmenting image otherwise iteration process is continued. Fig.5 shows the segmented image.

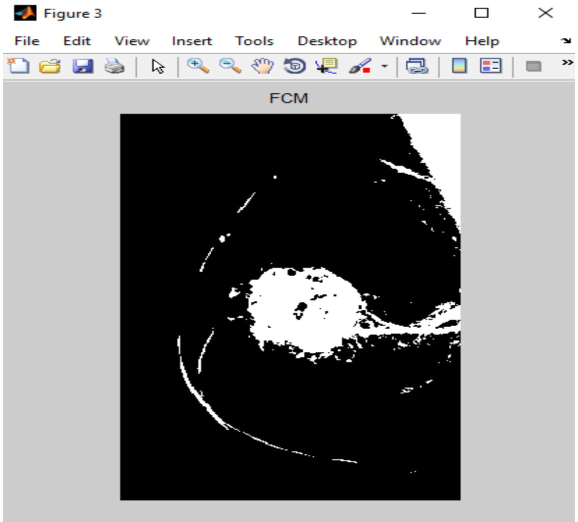


Fig.5 Image After Segmentation

D. IMAGE POST-PROCESSING

Post-processing helps to find the boundary of the object from background areas. Morphological operations such as erosion and dilation. Morphological Techniques consider an image with a small arrangement called structuring element [11]. Which is placed at all feasible locations in the image and it is correlated with neighbourhoods of pixels. Two morphological operations are Erosion and Dilation. Erosion reduces the size of region of interest and it also removes small fine points from an image. The dilation enlarges the shapes contained in the image. Morphological opening and closing are two other operations that are defined by specific combinations of dilation and erosion. Erosion comes after dilation in the case of opening and reverse process for closing. Morphological opening is generally used to smoothen region contours and removes thin projections in images. Similarly, morphological closing adds smoothness to image contours; however, it generally fuses two large regions separated by narrow breaks. This effect is opposite to that caused by morphological opening that breaks the narrow links between two large regions. Fig.6 shows the post-processed image after doing the erosion, dilation, small elements removed and border removal. Fig.7 show the final segmented image which is input to the feature extraction stage.

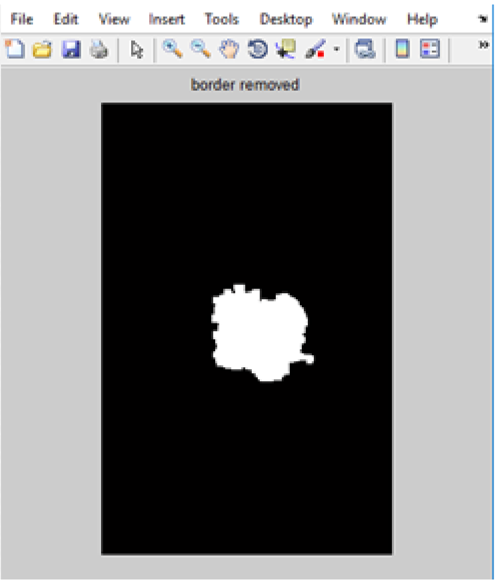


Fig.6 Image After Post-processing

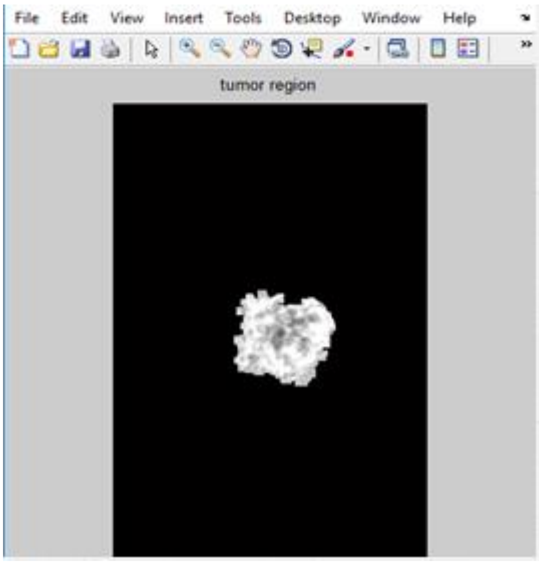


Fig.7 Tumour Region

E. FEATURE EXTRACTION

Features [15] helps to identify the patterns in an image and it give related information about the image. The accuracy of the classification depends on the feature extraction stage. In the proposed work, 13 features are extracted and are stored in vector format.

1. MEAN

Mean of the pixel value is used to calculate the value in the image in which central clustering occurs.

$$\mu = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} P(i,j)$$

Where P(i,j) is the pixel value at point (i,j) of an image of size MxN.

2. STANDARD DEVIATION

The Standard Deviation σ is used to calculate the mean square deviation of grey pixel value P(i, j) from its mean value m of the image. Standard deviation describes the dispersion inside a neighbourhood region.

$$\mu = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(P(i,j)-\mu)^2}$$

3. SMOOTHNESS

Relative smoothness 'R' is a measure of grey level contrast that can be used to establish descriptors of relative smoothness.

$$R = 1 - \left(\frac{1}{1+\sigma^2}\right)$$

where σ is the standered deviation.

4. ENTROPY

Entropy H can also be used to describe the distribution variation in a region.

$$H = -\sum_{k=0}^{L-1} P_{rk}\,(\log_2 P_{rk})$$

Where, P_{rk} is the probability of the kth grey level, which can be calculated as Zk /m*n, Zk is the total number of pixels with the kth grey level and L is the total number of grey levels.

5. SKEWNESS

Skewness S characterizes the degree of asymmetry of a pixel distribution in the specified window around its mean. Skewness is a pure number that characterizes only the shape of the distribution.

$$S = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(\frac{p(i,j)-\mu}{\sigma}\right)^3$$

Where, p(i, j) is the pixel value at point (i,j), μ and σ are the mean and standard deviation respectively.

6. KURTOSIS

Kurtosis measures the peakness or flatness of a distribution relative to a normal distribution.

$$S = \left\{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(\frac{p(i,j)-\mu}{\sigma}\right)^4\right\} - 3$$

Where, p(i,j) is the pixel value at point (i,j), μ and σ are the Mean and Standard Deviation respectively. The subtracting 3 makes the value zero for a normal distribution.

7. RMS

RMS (Root Mean Square) computes the RMS value of every row or column of the input, along vectors of a detailed dimension of the input, or of the whole input. The RMS value of the jth column of an M×N input matrix u is given by below equation:

$$y = \sqrt{\sum_{i=1}^{M}\left|\mu_{ij}\right|^2}$$

8. INVERSE DIFFERENCE MOMENT (IDM)

It is a measure of image texture. IDM ranges from 0.0 for an image that is highly textured to 1.0 for an image that is untextured.

$$H = \sum_{i,j}\frac{p(i,j)}{1+|i-j|}$$

9. ENERGY

Energy is also known as uniformity. The range of energy is [0 1]. Energy is 1 for a constant image.

$$H = \sum_{i,j} p(i,j)^2$$

10. CONTRAST

Contrast returns a measure of the intensity contrast between a pixel and its neighbour over the whole image. Contrast is 0 for a constant image.

$$C = \sum_{i,j}|i-j|^2 p(i,j)$$

11. CORRELATION

Correlation returns a measure of how each pixel is correlated to its neighbour over the whole image. The range of correlation is [-1 1]. Where the means and standard deviations of pi and pj, the partial probability density functions.

$$corr = \sum_{i,j} \frac{(1-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i \sigma_j}$$

12. HOMOGENEITY

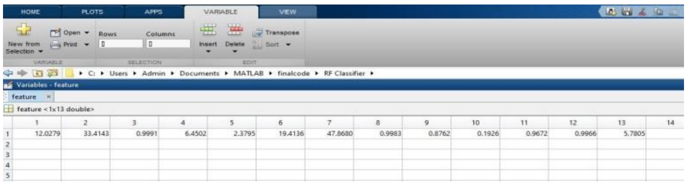
Homogeneity returns a value that measures the closeness. The range of homogeneity is [0 1].

$$H = \sum_{i,j} \frac{p(i,j)}{1+|i-j|}$$

13. VARIANCE

Variance is the square root of standard deviation. It is the difference by which set of random numbers are spread with respect from their average values.

$$var = \left(\sqrt{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (P(i,j)-\mu)^2} \right)^{1/2}$$



C: ▸ Users ▸ Admin ▸ Documents ▸ MATLAB ▸ finalcode ▸ RF Classifier ▸

Variables - feature

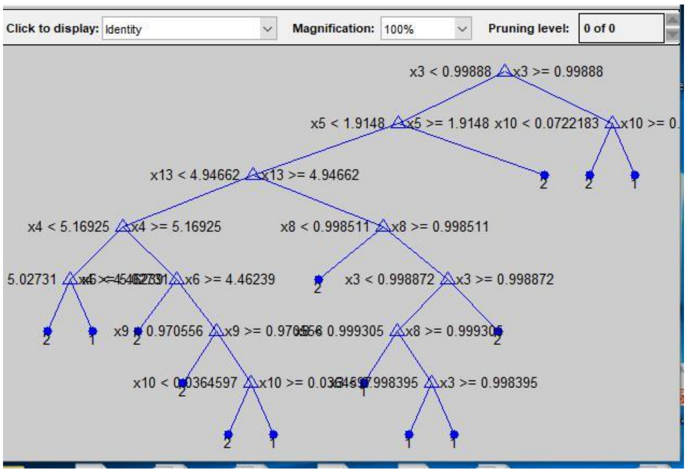
feature <1x13 double>

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	12.0279	33.4143	0.9991	6.4502	2.3795	19.4136	47.8680	0.9983	0.8762	0.1926	0.9672	0.9966	5.7805	
2														
3														
4														
5														

Fig.8 Extracted Features

F. RANDOM FOREST CLASSIFIER

Random forests or random selection forests are an ensemble gaining knowledge of approach for classification, regression and other tasks, that function via establishing a multitude of decision trees at training time and outputting the type that is the mode of the instructions (classification) or mean prediction (regression) of the individual trees. Random Forest develops plenty of decision trees primarily based on random resolution of variables and random resolution of data. It provide the type of structured variable primarily based on many trees. Many such random trees together forms a random forest.



Click to display: Identity

Magnification: 100%

Pruning level: 0 of 0

Fig.9 Tree Formed Based On Random Forest Algorithm

Individual decision trees tend to overfit. Bootstrap-aggregated (bagged) decision trees combine the results of many decision trees, which reduces the effects of overfitting and improves generalization. The figure above shows Tree Bagger grows the decision trees in the ensemble using bootstrap samples of the data.

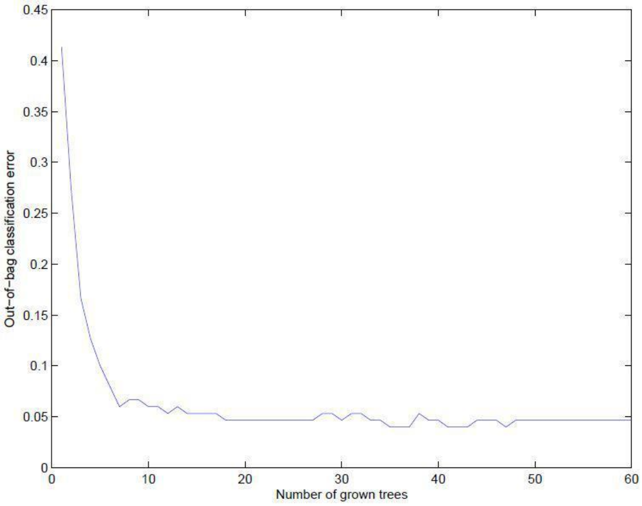


Fig. 10 Random Forest Error Plot

Suppose training data set is represented by T and suppose data set has M features (or attributes or variables).

T= (X_1, Y_1), (X_2, Y_2), ... (X_n, Y_n) and Xi is input vector X_{i1}, X_{i2}, ...X_{iM} and y_i is the label (or output or class). Random Forests algorithm is a classifier based on two methods-bagging and random subspace method. Suppose S variety of trees in forest then first create S datasets of same measurement as original created from random resampling of records in T with-replacement (n times for each dataset). This will endresult in T1, T2, ... TS datasets. Each of these is known as a bootstrap dataset. Due to "with-replacement" each dataset Ti can have replica data records and Ti can be lacking countless records data from unique datasets. This is known as Bagging.

Now, RF creates S trees and uses m (=sqrt(M) or =floor(lnM+1)) random sub elements out of M possible elements to create any tree. This is referred to as random subspace method.

So for every Ti bootstrap dataset will create a tree Ki. If to classify some input records D = x_1, x_2, ... x_M let it pass through each tree and produce S outputs (one for each tree) which can be denoted by Y = y_1, y_2, ... y_s. Final prediction is a majority vote on this set.

After creating the classifiers (S trees), for each (x_i, y_i) in the original training set i.e. T, select all T_k which does not include (x_i, y_i). This subset, pay attention, is a set of bootstrap datasets which does not include a unique record from the original dataset. This set is referred to as out-of-bag examples. There are n such subsets (one for each information record in authentic dataset T). OOB classifier is the aggregation of votes only over T_k such that it does not include T_k such that it does not contain (x_i, y_i).

Out-of-bag estimate for the generalization error is the error rate of the out-of-bag classifier on the training set (compare it with known y_i's).

Advantage of random forest are:

i. No need for pruning trees
ii. Accuracy is generated automatically.
iii. Not very sensitive to outliers in training data.
iv. Setting of parameters are easy.

IV. RESULTS AND DISCUSSIONS

Implemented the proposed system using MATLAB to verify the performance of the system. It is tested over 100 mammogram x-ray image. 75 percent of mammogram images are used for training and 25 percent images are used for testing. In this proposed work, Fuzzy c-means and Morphological Operations are used to get the segmented image. Finally Random Forest classier is used to improve the result. Segmentation provide correct result for large data set.

The accuracy of the given system can be calculated as the ratio between the total number of correctly classified instances and the test set size, given by;

$$\text{Accuracy} = \frac{\text{Instances correctly classified}}{\text{Total Instances}}$$

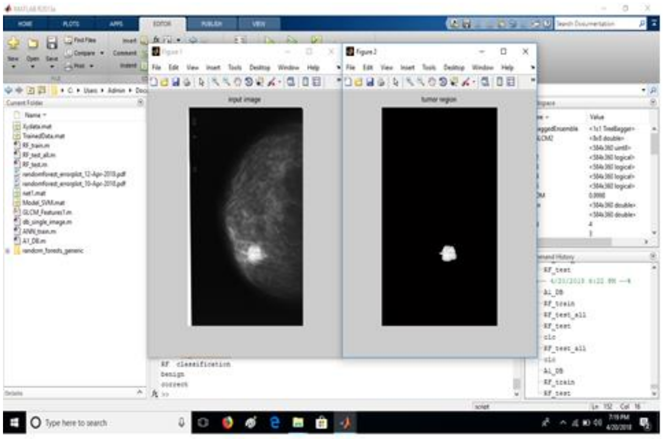
The accuracy of the mammogram segmentation algorithm in this work can be calculated by comparing each segmented mammogram mask in the database (MIAS (Mini mammographic Database)) [16] with its corresponding result that got after segmenting the image with Fuzzy c-means and Random Forest classifier. Proposed system gave 90.47 percent accuracy.

The accuracy of proposed work is compared with some of the existing works [17,18,19] and is given in Table 1.

TABLE 1

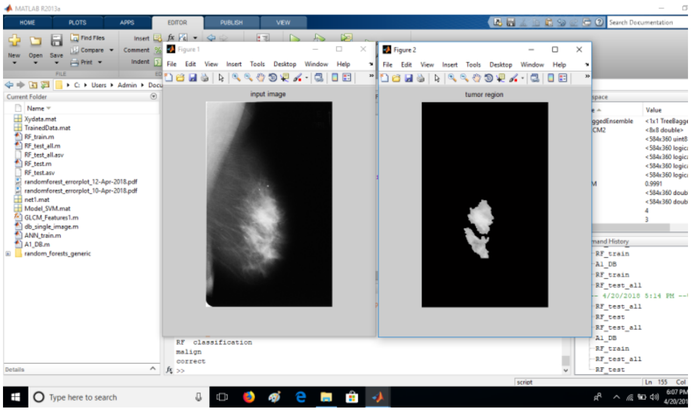
COMPARISON TABLE

Approach	Average Accuracy (%)
K-Means and Fuzzy logic means	85%
Multi Wavelet Techniques	87%
Watershed Segmentation	88%
Proposed Method	90.47%



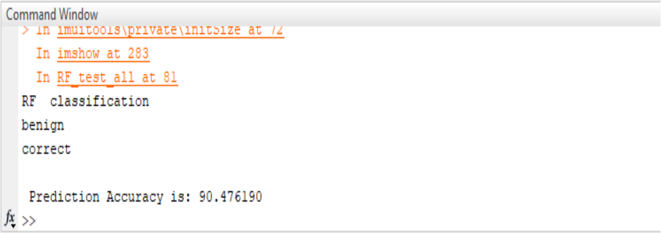
```
RF  classification
benign
correct
>>
```

Fig. 11 Test for a benign image



```
RF  classification
malign
correct
>>
```

Fig. 12 Test for malignant image



Command Window

```
> In imuitools\private\initSize at 72
  In imshow at 283
  In RF_test_all at 81
RF  classification
benign
correct

 Prediction Accuracy is: 90.476190
>>
```

Fig. 13 Accuracy

V. CONCLUSION

Breast cancer is one of the major causes of death among women. Image processing methodology helps doctors to identify cancer patients easily. The study has been proposed to enhance and clarify the focused areas in mammogram images. To solve the problem associated with image segmentation, a hybrid method comprising of two or more methods can be used, which may lead to better accuracy. In the proposed system, an attempt has been made to combine the Fuzzy method along with Random Forest classifier to improve the result and the proposed method gave 90.47% accuracy.

IV. REFERENCES

1. Mohd.Ashique, Ridwan Nayeem, Md.A.Mannan Joadder, Shahrin Ahammad Shetu, "Feature Selection for Breast Cancer Detection from UltrasoundImages,",3rdINTERNATIONALCONFERENCEONINFORMATICS, ELECTRONICS and VISION (2014), 978-14799-5180- 2/14/ 2014 IEEE.
2. Rupinder Singh, Jarnail Singh, Preetkamal Sharma, Sudhir Sharma, " Edge Based Region Growing", IJCTA - July-August 2011.
3. R Beaulah Jeyavathana, Dr. R.Balasubramanian, A. Anbarasa Pandian, "A Survey: Analysis on Pre-processing and Segmentation Techniques for Medical Images", International Journal of Research and Scientific Innovation (IJRSI) - Volume III, Issue VI, June 2016 - ISSN 23212705 .
4. V Kumar, T. Lal, P. Dhuliya and D. Pant, "A study and comparison of different image segmentation algorithms, 2016 2nd International Conference on Advances in Computing, Communication, and Automation (ICACCA) (Fall), Bareilly, 2016, pp. 1-6.
5. Yadollahpour Ali and Shoghi Hamed, "Early Breast Cancer Detection using Mammogram Images: A Review of Image Processing Techniques", BIOSCIENCES BIOTECHNOLOGY RESEARCH ASIA, March 2015. [6] Neha Sharma and Jayanand Manjhi, "Detection of malignant tissue in mammography image using morphology based segmentation technique", International

Journal of Medical Research and Review , March 2016.

6. RRamani, S. Valarmathy and N.Suthanthira Vanitha, "Breast Cancer Detection in Mammograms based on Clustering Techniques- A Survey", International Journal of Computer Applications (0975 8887) Volume 62 No.11, January 2013.
7. Luqman Mahmood Mina, Nor Ashidi Mat Isa "Preprocessing Technique for Mammographic Images ", International Journal of Computer Science and Information Technology Research Vol. 2, Issue 4, pp: (226231),October - December 2014.
8. IlatulFerdouse et al, "Simulation and performance analysis of Adaptive filtering algorithms in noise cancellation". IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 1, January 2011.
9. Waseem Khan, "Image Segmentation Techniques: A Survey ". Journal of Image and Graphics Vol. 1, Issue No. 4, pp.166-170, December 2013.
10. Mariam Biltawi, Nijad Al-Najdawi, Sara Tedmori, " Mammogram Enhancement And Segmentation Methods: Classification, Analysis, And Evaluation". The 13th International Arab Conference on Information Technology ACIT'2012 Dec.10-13.
11. Ahmed, M. N., Yamany, S. M., Mohamed, N., Farag, A. A., Moriarty, " A modified fuzzy c-means algorithm for estimation and segmentation of MRI data. Medical Imaging". IEEE Transactions on, 21(3), 1993. [13] M. Yambal and H. Gupta, "Image Segmentation using Fuzzy C Means Clustering: A survey". International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 7, July 2013.
12. Pradeep n, girisha h, sreepathi b and karibasappa k , "Feature Extraction Of Mammograms ". International Journal of Bioinformatics Research 2012.
13. N. Senthilkumaran and R. Rajesh, "Edge Detection Techniques for Image Segmentation A Survey of Soft Computing Approaches". International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009. [16] Mariam Biltawi, Nijad Al-Najdawi, Sara Tedmori, " Mammogram Enhancement And Segmentation Methods: Classification, Analysis, And Evaluation". The 13th International Arab Conference on Information Technology ACIT'2012 Dec.10-13.
14. H. G. Kaganami and Z. Beij, "Region Based Detection versus Edge Detection". IEEE Transactions on Intelligent information hiding and multimedia signal processing, pp. 1217-1221, 2009.
15. Pradeep n, girisha h, sreepathi b and karibasappa k, "Feature Extraction Of Mammograms", International Journal of Bioinformatics Research 2012.
16. J Suckling et al (1994) "The Mammographic Image Analysis Society Digital Mammogram Database Exerpta Medica", International Congress Series 1069 pp375-378.
17. Teshnehlab M., Aliyari Shoorehdeli M. and Keyvanfard F., "Feature selection and classification of breast MRI lesions based on Multi classi?er", International Symposium on Arti?cial Intelligence and Signal Processing (AISP), 54-58, 2011.
18. Kother Mohideen, Arumuga Perumal, Krishnan and Mohamed Sathik, "Image De noising and Enhancement Using Multi wavelet with Hard Threshold In Digital Mammographic Images", International Arab Journal of e-Technology, 2(1), 2011.
19. Yan Zhang and Xiaoping Cheng, "Medical image segmentation based on watershed and graph theory", Image and Signal Processing (CISP), 3, 1419-1422, 2010