Detection and Classification of Breast Cancer from Digital Mammograms using RF and RF-ELM Algorithm

R. D. Ghongade
Ph. D. Scholar, SGB Amravati University,
Amravati, Maharashtra, India.
rahulghongade@rediffmail.com

D. G. Wakde
Director, P. R. Patil College of Engineering & Technology,
Amravati, , Maharashtra, India
director_prpce@rediffmail.com

Abstract- Neural Network is utilized as a developing analytic tool for the diagnosis of breast cancer. The goal of this research is to determine breast tumor from digital mammograms with a machine learning technique in view of RF and combination of RF-ELM classifier. For digital mammogram images, MIAS database is used. Preprocessing is usually needed to enhance the low quality of the image. The region of interest (ROI) is determined in line with the scale of suspicious region. After the suspicious area is sectioned, features are extracted by texture analysis. GLCM is used as a texture attribute to extract the suspicious area. From all extracted features best features are selected with the help of CBF method. To enhance the exactness of classification, only six features are selected. These features are mean, standard deviation, kurtosis, variance, entropy and correlation coefficient. RF and RF-ELM are used as a classifier. The outcomes of present work show that the CAD system with the usage of RF-ELM classifier may be very powerful and achieves the exceptional results in the prognosis of breast cancer.

Keywords: Breast cancer, CAD, ELM, Feature selection, Digital Mammogram, MIAS, RF-ELM.

I. INTRODUCTION

Breast cancer is extensively analyzed disease in ladies around the world. It is the most common reason for malignancy demise in ladies in less developed regions and second cause in more developed regions after lung cancer. It has been estimated that near 1.6 million cases of breast cancer registered worldwide in 2010 [1][2]. The survey of 2013 indicates that 230,815 women and 2,109 men in the United States were diagnosed with breast cancer. Estimated New Breast Cancer Cases and Deaths by Sex, United States, 2017 are as shown in Table 1. [3] As per the study at NCI's Division of Cancer Epidemiology and Genetics, Breast cancer will grow from 283,000 cases in 2011 to 441,000 in 2030. [4]

Early diagnosis is essential for survival, in particular in developing countries where the diseases are recognized very late. Mammography is the preeminent screening tool that makes use of X-ray to supply an image of the breast to diagnosis breast tumors. Mammography shows the morphological aspects of the breast, such as the anatomical structures and all the breast tissues, namely glandular, fibrous and adipose tissues. In a few instances, the result is probably stressed whilst a mammogram discovers something that looks like cancers but turns out to be benign. On mammograms, dense breast tissue looks white. Breast masses or tumors also look white; hence sometimes dense tissue hides tumors. Indeed, even qualified and experienced radiologists may miss bosom tumors because of the density of bosom [5]. The successful markers of malignancy often applied as part of assessing mammograms are masses and micro-calcifications. Mass detection is a tough and difficult problem than detection of micro-calcifications, that is because of the variant in size and form found in a mammogram and also masses often show poor image contrast [6].

Table I
Statistics of Breast Cancer Cases in U.S., 2017

Estimated New Cases			**Estimated Deaths**		
Male	***Female***	***Total***	***Male***	***Female***	***Total***
2470	252710	255180	460	40610	41070

The CAD system developed here will perceive the abnormality in mammograms and help radiologists to identify the regions of the problem. Hence CAD system for breast cancer has been become out to be a successful supplementary tool in the combat against breast cancers. It improves the detection rate especially in younger women, in which tumors are hidden because of dense breast tissue. [7][8]

This paper proposes a novel technique for breast cancer segmentation and detection from a digital mammogram. The method has been completed with the help of morphological operations and Artificial Neural Networks.

II. LITERATURE REVIEW

A literature evaluates confirmed the current trends in computer-aided diagnosis system for breast cancers with the help of statistical techniques and artificial neural networks. In-Sung Jung *et al.* [9] summarizes the diverse near investigation of neural network algorithms to get the best

978-1-5386-1703-8/17/$31.00 ©2017 IEEE

classification of the bosom disease and discloses how to procure more numerical parameters from the bosom growth image information, with the goal that it can help specialists to conclusion proficiently amongst benign and malignant tumors.

Nakayama *et al* [10] exhibited Bayes discriminant function for recognizing among strange ROIs with a micro calcification bunch and two unique sorts of ordinary ROIs without a micro calcification group. The execution is assessed by utilizing 600 mammograms.

Karahaliou *et al.* [11] proposed a method where they used gray-level and wavelet coefficient texture features of the tissue surrounding MC clusters on mammograms. Probabilistic neural network is used for differentiating malignant from benign with AUC of 0.989.

Zadeh *et al.* [12] use infrared images for the research. The histogram contains statistical information about the texture of the image. Mean, variance, kurtosis, skewness and entropy are five features are extracted using the histogram. Diagnosis is based on the back-propagation combinatorial model of the genetic algorithm and artificial neural network. The consequences of the combinatorial model with 50% sensitivity, 75% specificity and 70% accuracy indicate legitimate accuracy in growth determination.

Dheeba *et al.* [13] proposed another order approach in light of PSOWNN for identification of bosom malignancy in computerized mammograms. The execution of the proposed framework is evaluated by the area under the ROC curve. The outcome of proposed system shows that the area under the ROC curve is 0.96853 with a sensitivity 94.167% of and specificity of 92.105%.

Kamalakannan *et al.* [14] proposed a method where noise from the mammogram is removed with the help of Laplace filter and Gaussian filter. The difference of both filters is calculated for the better clarity of the image. The relevant part is cropped, and Otsu's method is used to determine the threshold value.

Pratiwi *et al.* [15] introduce a CAD system using Radial Basis Function Neural Network for mammograms classification based on Gray-level Co-occurrence Matrix (GLCM) texture based features. The result shows that RBFNN is better than Back-propagation Neural Network (BPNN) in performing breast cancer classification. The accuracy of benign and malignant classification is 94.29% with RBFNN which is 2% higher than BPNN.

Dhungel et al. [16] proposed a framework which utilized four modules to recognize the mass. The main module joins a multi-scale deep belief network (m-DBN) with a Gaussian mixture model classifier for candidate generation. These candidates are the contributions to a moment step containing a course of two phases of profound convolution neural systems that deliver highlights for being utilized by a straight SVM classifier. The third module comprises of a cascade of two stages of random forest for further reduction of false positives, where the input features for the RF classifier is a set of texture and morphological features. In the post processing module that merges regions with high overlap ratio using CCA. This strategy is compelling in the decrease of false positive areas while keeping a high genuine positive detection.

Weiying *et al.* [17] presented an innovative method for the diagnosis of breast cancer based on extreme learning machine. The performance of proposed CAD system compared with SVM and PSO-SVM. This system achieves the good performance with accuracy of 96.02%.

III. METHODOLOGY

In the present work, mammogram images with 1024 x 1024 pixels are imported from MIAS database. To enhance the contrast of the image and to smoothen image, preprocessing is done which will be helpful in further stages. Then segmenting the breast region is carried out in order to find out the suspicious area from breast segments. Later extraction of texture features and texture statistics computation is performed. Some relevant features are selected by Correlation-based feature Selection (CBF) method and these features are used for classification. Fig. 1 shows the proposed approach for detection of abnormality in mammograms.

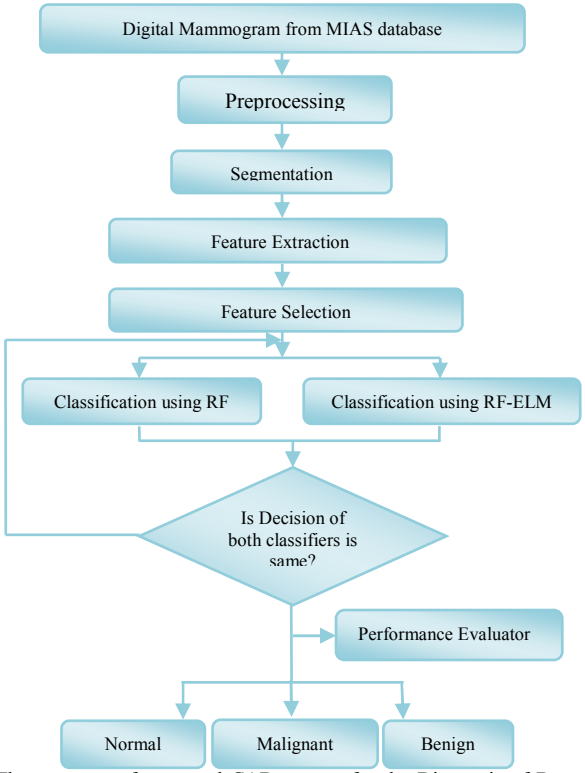

Fig. 1. The structure of proposed CAD system for the Diagnosis of Breast Cancer.

A. Preprocessing

The principle issue for extracting features of the mammographic images is noise, different resolution, quality and low contrast of mammograms. This makes the location of tumor considerably harder. Preprocessing is required to conquer this issue and makes efficient feature extraction of image possible. Firstly, the Gaussian filter is applied for smoothing of images. The Gaussian blur is a type of image-blurring filters that uses a Gaussian function for ascertaining the change to apply to every pixel in the image. It is used to 'blur' image and remove noise. Then adaptive histogram equalization is utilized to upgrade the difference of the grayscale picture by changing the qualities. Gaussian kernel coefficients are sampled from the 2D Gaussian functions as shown in equation (1).

$$G(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (1)$$

Where,

x - Distance from the origin on the horizontal axis,

y - Distance from the origin on the vertical axis and

σ - Standard deviation of the distribution.

B. Segmentation

A critical segmentation process is required that perceive and concentrate the harmful tumors. Region based segmentation is utilized to section masses from its background. Otsu's technique is utilized to naturally perform grouping based image thresholding. The calculation expects that the image contains two classes of pixels taking after bi-modular histogram (forefront pixels and foundation pixels), it then calculates the optimum threshold separating the two classes so that their intra-class variance is minimal. After image segmented (binary mask), it is multiplied with the original image as the normalization process. Thresholding is performed using equation (2).

$$G(x, y) = 1, \text{ for } f(x, y) > T$$
$$= 0, \text{ for } f(x, y) \leq T \quad (2)$$

Where T has chosen the value of Threshold.

C. Feature extraction

The gray level co-occurrence matrix (GLCM) is used as a statistical technique to extract the texture features. These features are contrast, correlation coefficient, energy, homogeneity, mean, standard deviation, entropy, variance, smoothness, kurtosis, skewness and inverse different moment.

On the other hand, Shape features like area, solidity, eccentricity, perimeter and major axis length are extracted.

D. Feature selection

Feature selection is the way toward choosing a subset of important features. It is utilized to diminish the feature space to enhance the exactness of characterization. This likewise limits the calculation time. Out of seventeen features, only six features are selected with the help of Correlation-based feature selection (CBF) method for further process.

It is a correlation based feature selection strategy which is essentially speedier than other subset determination strategies. The CBF assesses a subset of features on the basis of the following hypothesis: "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other". The equation (3) gives the merit of a feature subset 'S' consisting of 'k' features:

$$Merit,\ S_k = \frac{k\,\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \quad (3)$$

Here, $\overline{r_{cf}}$ is the average value of all feature-classification correlations, and

$\overline{r_{ff}}$ is the average value of all feature-feature correlations.

The selected six features are mean, standard deviation, kurtosis, variance, entropy and correlation coefficient.

E. Classification

RF and RF-ELM are used as classifier. RF is an approach proposed by Breiman for classification tasks. It fundamentally originates from the blend of tree-organized classifiers with the haphazardness and vigor gave by stowing and irregular element choice. The grouping is performed by sending a specimen down in each tree and allocating it the name of the terminal hub. Toward the end, the ordinary vote of all trees is represented the classification. Bagging process of RF method of classification is well defined using equation (4).

Let $X = x_1, x_2, \ldots x_n$ be the training set having set of responses $Y = y_1, y_2, \ldots y_n$ for Z times. Then for $z = 1$ to Z, predictions for all such unseen samples which are denoted by $\hat{x}$ and can be defined as

$$\hat{f} = \frac{1}{Z} \sum_{z=1}^{Z} f_z(\hat{x}) \quad (4)$$

RF is very efficient with large datasets and high dimensional data. The architecture of RF is shown in Fig.2.

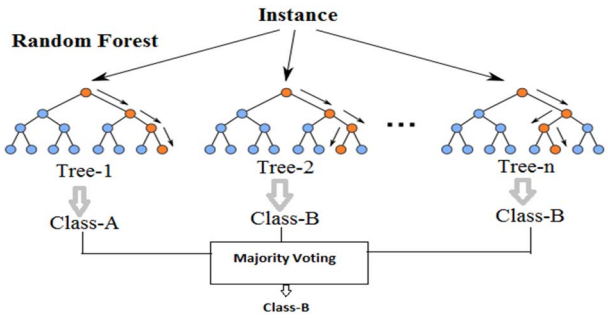

Fig. 2. Architecture of Random Forest

The ELM (Extreme Learning Machine) algorithm learns a model of the form

$$Y = W_2\, \sigma\, (W_1 x)$$

Where $\mathbf{W}_1$ is the matrix of input-to-hidden-layer weights, σ is some activation function, and $\mathbf{W}_2$ is the matrix of hidden-to-output-layer weights.
The algorithm proceeds as follows:

1. Fill $\mathbf{W}_1$ with Gaussian random noise;
2. Estimate $\mathbf{W}_2$ by least-squares fit to a matrix of response variables $\mathbf{Y}$, computed using the pseudo inverse. + , given a design matrix $\mathbf{X}$:

$$W_2 = \sigma\, (W_1 X)^{+} Y$$

RF-ELM is a combination of RF and ELM algorithm which improves the accuracy more than other classifiers. If RF and RF-ELM provide the same classification then classification result obtained else it gets back to the classification again.

F. Performance Evaluation

Confusion matrix, ROC curve with AUC score is the parameters to evaluate the performance of classification algorithm. Confusion matrix helps to get information about both actual and predicted class classification.

The TPR and FPR are used plot the ROC curve. The TPR is used to calculate correctly classified malignant ROIs from all available malignant ROIs. The FPR parameter can calculate incorrectly classified benign ROIs amongst the total number of benign ROIs. At the end Accuracy, Precision, Sensitivity and Specificity parameters are calculated to assess the system performance.

IV. EXPERIMENTS

The experiments are performed using the MIAS database. This database was created to contain two experimental datasets on the same images. In the first dataset, the images are split into two classes: normal or abnormal, and in the second dataset, the images are split into three classes: benign, malign and normal. The abnormal images in this database contain the coordinates and the radius. The first dataset of all extracted features is created, the process of the Gaussian filter and histogram equalization implemented in this research. The RF-ELM classifier is used for classification. The following steps briefly describe the experiment in this research:

From mammogram, the region of interest (ROI) was extracted from images depending on the information contained in the dataset.

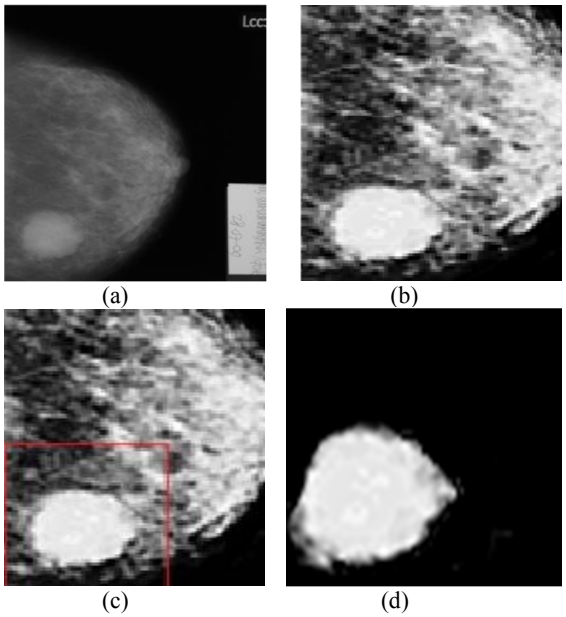

Fig. 3. Detected abnormalities using RF-ELM approach. (a) Original Mammogram Image, (b) Cropped and Enhanced Image, (c) ROI and (d) detected abnormalities.

To remove noise from mammogram images and improve the quality of the region of interest the Gaussian filter and Adaptive histogram equalization are applied.

The features are extracted from the normalized image regions using GLCM. Then from these features, only six features are selected by CBF selection method to improve the accuracy.

RF and RF-ELM algorithms are used for Classification. If both algorithm results are same then it classifies the tumor, else it get back to the classification stage.

Performance is evaluated with the help of Confusion matrix, Accuracy, Sensitivity, Specificity, ROC curve with AUC score. Fig. 3 shows the results at different stages using RF_ELM approach.

V. RESULTS AND DISCUSSION

True Positives (TP), False Positives (FP), True negatives (TN) and False Negatives (FN) are four different possible outcomes of a single prediction for a two class case. Accuracy, Sensitivity, specificity and ROC curve with AUC score are statistical parameters that help to evaluate the performance. Sensitivity measures the proportion of real positives which are properly recognized when the mammogram contains malignancies tissues in it. Specificity quantifies the proportion of negatives which are properly recognized when cancer is not present in the mammogram. The evaluated performance of CAD system using RF and RF-ELM classifier is tabulated in Table 2 and Fig. 4 shows that RF-ELM gives better evaluation results as compare to RF.

Table II
Performance evaluation of CAD system with various statistical parameters.

Method	Accuracy (%)		Sensitivity (%)		Specificity (%)		AUC (microns)	
	Best	*Av*	*Best*	*Av*	*Best*	*Av*	*Best*	*Av*
RF	89	80	90	80	92	81	2	0.9
RF-ELM	98	95	97.9	89	97	91	2.5	1.9

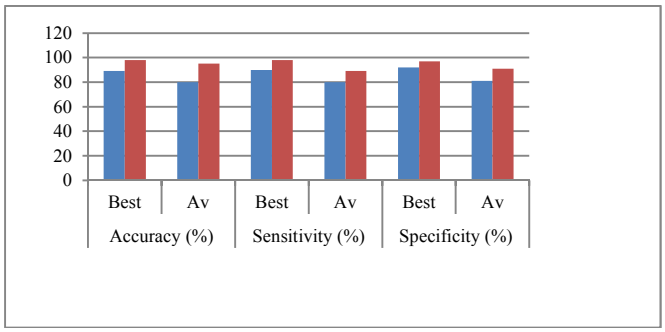

Fig. 4. Plot shows the comparison of RF and RF-ELM performance.

Comparison of ROC (Receiver operating characteristics) curve for RF, RF-ELM and Naive Bayes classifiers are shown in Fig. 5.

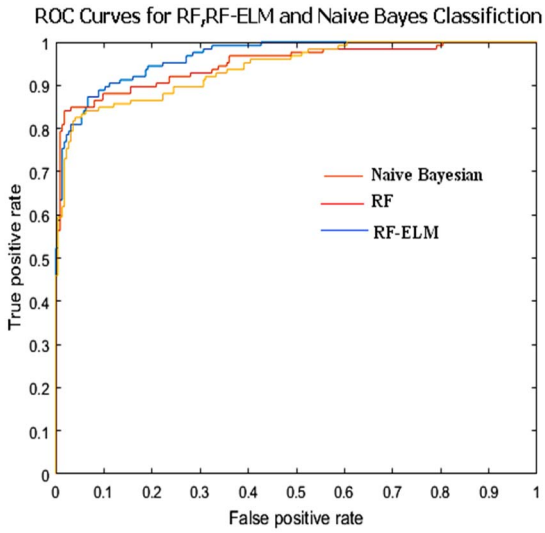

Fig. 5. A comparison of ROC curve for various classifiers.

VI. CONCLUSION

This paper proposed a CAD framework, which classifies tumor from mammograms into normal, benign & malignant. Subsequent to preprocessing of the digital mammogram and the determination of ROI, features are extracted and after that ordered utilizing RF and RF-ELM classifier. Otsu's technique is utilized for Image segmentation to deliver the better outcome.

The outcome demonstrates that RF-ELM classifier gives fundamentally better classification exactness by decreasing the FPs and FNs and it likewise relies on the improvement of feature selection. This finding is astoundingly significant for the radiologist in recognizing the malignancy from digital mammograms.

REFERENCES

[1] Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. "Global cancer statistics. CA Cancer J Clin" 2011; 61(2): pp: 69–90.

[2] Forouzanfar MH, Foreman KJ, Delossantos AM, "Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis" Lancet 2011; 378(9801): pp: 1461–84.

[3] Rebecca L. Siegel MPH, Kimberly D. Miller, Ahmedin Jemal DVM, "Cancer Statistics, 2017". CA: A Cancer Journal for Clinicians"

[4] "Study forecasts new breast cancer cases by 2030", April 23, 2015, By NCI Staff, National Cancer Institute at the National Institutes of Health, USA.

[5] Kolb TM, Lichy J, Newhouse JH. "Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations". Radiology 2002;225(1), pp:165–75.

[6] Cheng HD, Cai X, Chen X, Hu L, Lou X. "Computer-aided detection and classification of microcalcifications in mammograms: a survey". Pattern Recogn 2003;36(12), pp:2967–91.

[7] Cheng HD, Shi XJ, Min R, Hu LM, Cai XP, Du HN. "Approaches for automated detection and classification of masses in mammograms", Pattern Recogn 2006; 39(4), pp:646–68.

[8] Motakis Efthimios, Ivshina Anna V, Kuznetsov Vladimir A. "Data-driven approach to predict survival of cancer patients", IEEE Eng Med Biol Mag 2009;28(4), pp:58–66.

[9] In-Sung Jung, Devinder Thapa, and Gi-Nam Wang, "Neural Network Based Algorithms for Diagnosis
and Classification of Breast Cancer Tumor" CIS 2005, Part I, LNAI 3801, Springer-Verlag Berlin Heidelberg 2005, pp: 107–114

[10] Ryohei Nakayama, Yoshikazu Uchiyama, Koji Yamamoto, Ryoji Watanabe, and Kiyoshi Namba "Computer-Aided Diagnosis Scheme Using a Filter Bank for Detection of Microcalcification Clusters in Mammograms" IEEE transactions on Biomedical Engineering, vol. 53, no. 2, February 2006, pp: 273–283.

[11] Karahaliou Anna N, Boniatis Ioannis S, Skiadopoulos Spyros G, Sakellaropoulos Filippos N, Arikidis Nikolaos S, Likaki Eleni A, et al. "Breast cancer diagnosis: analyzing texture of tissue surrounding microcalcifications", IEEE Trans Inf Technol Biomed 2008;12(6), pp:731–738.

[12] Hossein Ghayoumi Zadeh, Javad Haddadnia, Maryam Hashemian, Kazem Hassanpour. "Diagnosis of Breast Cancer using a Combination of Genetic Algorithm and Artificial Neural Network in Medical Infrared Thermal Imaging" Iranian Journal of Medical Physics, Vol. 9, No. 4, Autumn 2012, pp: 265-274

[13] J. Dheeba, N. Albert Singh, S. Tamil Selvi "Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized

wavelet neural network approach" Journal of Biomedical Informatics 49 (2014), pp:45–52.

[14] Kamalakannan J, Dr. M.Rajashekara Babu, Dr. P. Venkata Krishna, Kansagra Deep Mukeshbhai "Identification of Abnormality from Digital Mammogram to Detect Breast Cancer", 2015 IEEE International Conference on Circuit, Power and Computing Technologies [ICCPCT]; 978-1-4799-7075-9/15.

[15] Mellisa Pratiwi, Alexander, Jeklin Harefa, Sakka Nanda, "Mammograms Classification using Gray-level Co-occurrence Matrix and Radial Basis Function Neural Network," International Conference on Computer Science and Computational Intelligence (ICCSCI 2015), Procedia Computer Science 59 (2015), pp: 83 – 91.

[16] Neeraj Dhungel, Gustavo Carneiro, Andrew P Bradley. "Automated Mass Detection from Mammograms using Deep Learning and Random Forest" 2015 Research Gate conference, October 2015, DOI: 10.1109/DICTA.2015.7371234

[17] Weiying Xie, Yunsong Li, Yide Ma "Breast mass classification in digital mammography based on extreme learning machine", Elsevier Ltd., Neuro-computing (2015), DOI: 10.1016/j.neucom. 2015.08.048