Journal of Engineering and Applied Sciences 12 (Special Issue 12): 9477-9482, 2017
ISSN: 1816-949X
© Medwell Journals, 2017

Random Forest and Extreme Learning Machine Based CAD System for Breast Cancer

[1]R.D. Ghongade and [2]D.G. Wakde
[1]SGB Amravati University, Amravati, Maharashtra, India
[2]P.R. Patil College of Engineering and Technology, Amravati, Maharashtra, India

Abstract: Neural network is utilized as rising diagnosis tool for cancer disease. The goal of this exploration is to determine tumor growth in breast with a machine learning method based on RF, ELM and RF-ELM classifier. MIAS database is utilized for the advanced mammogram images. Pre-processing is for the most part expected to enhance the low nature of image. The ROI is resolved by the measure of suspicious region. After the suspicious area is portioned, features are extracted by texture analysis. GLCM is utilized as a surface credit to extricate the suspicious region. From all extracted features best features are chosen with the assistance of CBF and PCA. RF, ELM and RF-ELM are utilized as classifier. The consequences of present resarch demonstrate that the CAD framework utilizing RF-ELM classifier is exceptionally compelling and accomplishes the best outcome in the finding of breast malignancy.

Key words: Breast cancer, CAD, CBF, ELM, feature selection, mammogram, PCA, RF-ELM

INTRODUCTION

Breast cancer is generally analysed malignancy in ladies worldwide and the one of the main sources of tumour demise among them. It is the most common reason for malignancy demise in ladies in developing areas and second reason in developed areas after lung tumour. It has been assessed that more than 1.6 million instances of bosom tumour enlisted worldwide in 2010 (Jemal *et al*., 2011; Forouzanfar *et al*., 2011). The review of 2013 demonstrates that 230,815 ladies and 2,109 men in the United States were determined to have bosom malignancy.

Estimated new breast cancer cases and deaths by sex in United States in 2017 are 255180 and 41070, respectively (Siegel *et al*., 2017). 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, a more than 50% increase (Anonymous, 2015).

Early conclusion is essential for survival, especially in developing nations where the diseases are analyzed late. Mammography is the transcendent screening apparatus that utilizations X-beam to deliver a image of the breast to analysis bosom disease. Mammography has been more successful in screening a symptomatic lady to decrease mortality. At times result may be befuddled when a mammogram finds something that looks like growth, yet, ends up being generous (not tumour). On mammograms, thick bosom tissue looks white. Bosom masses or tumours additionally look white subsequently some of the time thick tissue conceals tumours. Indeed, even qualified and experienced radiologists may miss bosom growths because of the thickness of bosom (Kolb *et al*., 2002). The capable markers of malignancy routinely used as a piece of evaluating mammograms are masses and micro-calcifications. Mass detection is a troublesome and testing issue than that of micro-calcifications, this is because of variation in size and shape observed in a mammogram and furthermore masses regularly display poor image contrast (Cheng *et al*., 2003).

The CAD framework created here will identify the abnormality in digital mammograms and help radiologists to detect the ranges of concern. Subsequently, CAD framework for breast malignancy has been ended up being a competent supplementary apparatus in the battle against breast tumour. It enhances the identification rate particularly in more youthful ladies where disease masses are shrouded on account of thick bosom tissue (Cheng *et al*., 2006; Motakis *et al*., 2009).

This study proposes a novel approch for breast cancer segmentation and detection from digital mammogram images. The approach has been carried out with the assistance of morphological operations and Artificial Neural Networks (ANNs).

Literature review: A literature review showed the current improvements in CAD systems for breast cancer utilizing factual methodologies and artificial neural networks.

Corresponding Author: D.G. Wakde, P.R. Patil College of Engineering and Technology, Amravati, Maharashtra, India

Karahaliou *et al.* (2008) proposed a technique where they utilized gray-level and wavelet coefficient texture features of the tissue encompassing MC bunches on mammograms. Probabilistic neural system is utilized for separating dangerous from considerate with AUC of 0.989.

Cedeno *et al.* (2011) proposed artificial metaplasticity multilayer perceptron algorithm that gives need in refreshing the weights for the less continuous actuations over the more incessant ones. AMMLP accomplishes a more effective preparing while at the same time keeping up MLP execution. Wisconsin breast cancer database is utilized as a part of present research. The performance is tested using classification accuracy, sensitivity, specificity and confusion matrix. The obtained AMMLP classification accuracy is 99.26%.

Jiji and Marsilin (2012) distinguished the right stage of bosom growth from the tumour profundity. Low level features are extracted. KNN algorithm is utilized for the classification and classification rate accomplished is 93.5% precision. For pattern similarity, Euclidean distance is used. Retrieval performance is assessed by comparing Euclidean distance metric and Mahalanobis distance metric.

Ahmad *et al.* (2013) introduce a computer-aided diagnosis system for breast cancer using a genetic algorithm for simultaneous feature selection and parameter optimization of Artificial Neural Networks (ANN). This algorithm is called GAANN_RP which use wisconsin breast cancer Dataset to produces the best and average, 99.43 and 98.29% correct classification, respectively. The performance is evaluated using three variations of BP training namely the RP, LM and GD.

Nugroho *et al.* (2014) used Multi Layer Perceptron (MLP) as a classifier and CFS for the feature selection towards breast cancer diagnosis on mammograms. Texture features are extracted based on histogram and GLCM.

Dheeba *et al.* (2014) proposed a new classification approach based on PSOWNN for detection of breast cancer in digital mammograms. Performance of the proposed system is evaluated by the area under the ROC curve.

Xie *et al.* (2016) 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 framework accomplishes the great execution with precision of 96.02%

Mohebian *et al.* (2017) presented HPBCR (Hybrid Predictor of Breast Cancer Recurrence), a hybrid technique including statistical features selection, meta-heuristic population-based optimization and ensemble learning to predict breast cancer recurrence in the first 5 years after the diagnosis.

METERIALS AND METHODS

The approach in this research used for the detection of abnormality in mammograms is shown in Fig. 1.

In this research, mammogram images with 1024×1024 pixels are used from MIAS database. To improve the complexity of the images and to smooth the images, pre-processing is done which will be useful in further stages. Segmentation of the breast region is done with a specific end goal to discover the suspicious range from bosom segments. Later extraction of texture features and texture statistics computation is performed.

Some significant features are chosen by CBF and PCA technique and random forest are utilized for classification to decide the masses or non-masses.

Pre-processing: The standard issue for extricating elements of the mammographic images is noise, diverse determination, quality and low contrast of mammograms. This makes detection of tumour significantly harder. Pre-processing is required to overcome this issue and makes effective feature extraction of image conceivable. Mammographic picture is taken from MIAS database for

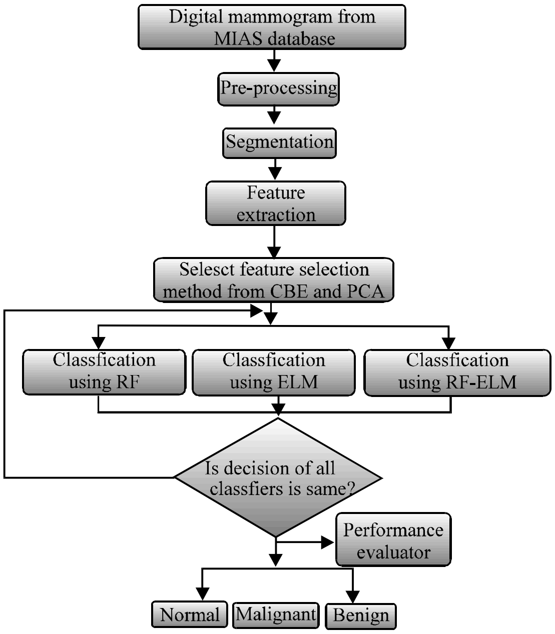
Digital mammogram from MIAS database
Pre-processing
Segmentation
Feature extraction
Selesct feature selection method from CBE and PCA
Classfication using RF
Classfication using ELM
Classfication using RF-ELM
Is decision of all classfiers is same?
Performance evaluator
Normal
Malignant
Benign

Fig. 1: Structure of the proposed CAD system for breast cancer

the pre-processing. 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 computing the change to apply to every pixel in the image. It is utilized to 'blur' image and expel noise. Then adaptive histogram equalization is utilized to upgrade the contrast of gray scale image by transforming the values. Gaussian kernel coefficients are sampled from the 2D Gaussian functions as shown in Eq.1:

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

Where:

x = The distance from the origin in the horizontal axis
y = The distance from the origin in the vertical axis and
σ = The standard deviation of the distribution

Segmentation: A significant segmentation process is required that perceive and concentrate the dangerous tumours. Region based segmentation is used to segment masses from its background. Otsu's technique is utilized to naturally perform grouping based image thresholding. The algorithm assumes that the image contains two classes of pixels following bi-modal histogram (foreground pixels and background pixels), it then computes the ideal limit isolating the two classes, so that, their intra-class change is insignificant. After image segmented (binary mask), it is multiplied with the original image as the normalization process. Thresholding is performed using Eq. 2:

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

where, T has chosen the value of Threshold.

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 (IDM). On the other hand, shape features like area, solidity, eccentricity, perimeter and major axis length are extracted.

Feature selection: Feature selection is the process of selecting a subset of relevant features. It is used to reduce the feature space to improve the accuracy of classification. This also minimizes the computation time. Correlation-Based Feature selection (CBF) is used to select the best features. Out of seventeen features, only seven features are selected for further process.

It is a correlation based feature selection method which is significantly faster than other subset selection methods. Equation 3 gives the merit of a feature subset 'S' consisting of 'k' features:

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

Where:

r_{cf} = The average value of all feature-classification correlations
r_{ff} = The average value of all feature-feature correlations

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

The main use of Principal Component Analysis (PCA) is to reduce the size of the feature space while retaining as much of the information as possible. A way too sees how much information we retain is to look at the explained variance ratio of the principal components. If we define the full variance of a data set as $\sigma = \Sigma_j\lambda_j$ then they explained variance ratio of component j is defined as $r_j = \lambda_j/\sigma$. Selected features to improve the accuracy of classification are mean, standard deviation, smoothness Inverse Different Moment (IDM), entropy, correlation, shape and variance.

Classification: Random Forest (RF) is used as a classifier. RF is an approach proposed by Breiman for classification tasks. It mainly comes from the combination of tree-structured classifiers with the randomness and robustness provided by bagging and random feature selection. The classification is performed by sending a sample down in each tree and assigning it the label of the terminal node. At the end, the normal vote of all trees is accounted for the classification. Bagging process of RF method of classification is defined using mathematical Eq. 4.

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

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

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

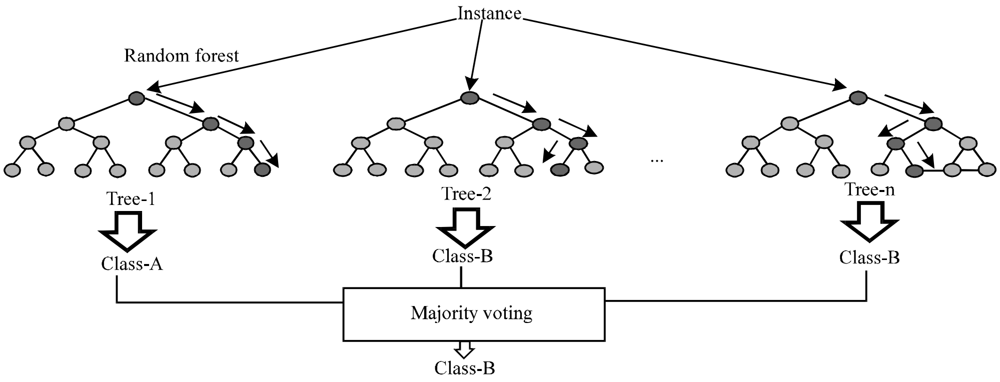

Fig. 2: Architecture of random forest algorithm

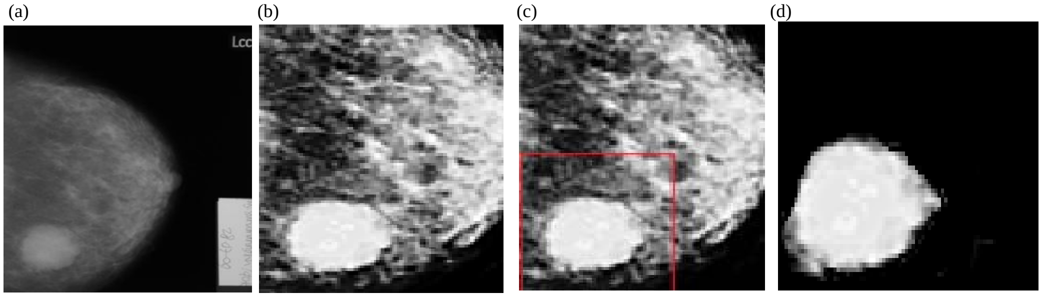

Fig. 3: Detection of abnormalities using RF-ELM approach; a) Original image; b) Cropped and enhanced image; c) ROI and d) Detected abnormalities

ELM algorithm: The ELM training algorithm learns a model of the form:

$$Y = W_2\sigma(W_1x)$$

where, W_1 is the matrix of input-to-hidden-layer weights, σ is some activation function and W_2 is the matrix of hidden-to-output-layer weights. The algorithm proceeds as follows:

1. Fill W_1 with Gaussian random noise
2. Estimate W_2 by least-squares fit to a matrix of response variables Y, computed using the pseudo inverse+given a design matrix X:

$$W_2 = \sigma(W_1,X) + y$$

RF-ELM is a combination of RF and ELM algorithm which improves the accuracy than other classifiers.

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 is 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. Figure 3 shows the detection of abnormalities using RF-ELM approach.

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, ELM and RF-ELM classifier with CBF and PCA feature selection method is tabulated in Table 1 and Table 2, respectively.

Comparison of ROC (Receiver Operating Characteristics) curve for RF, ELM and RF-ELM classifiers is shown in Fig. 4.

Table 1: Performance of the CAD system based on different classifier methods using CBF

	Accuracy (%)		Sensitivity (%)		Specificity (%)		AUC (μ)	
Methods	Best	Av.	Best	Av.	Best	Av.	Best	Av.
RF	89	80.0	90.0	80.7	92	81	2.00	0.90
ELM	92	87.5	92.0	85.0	94	88	1.99	1.12
RF-ELM	98	95.0	97.9	89.0	97	91	2.50	1.90

Table 2: Performance of the CAD system based on different classifier methods using PCA

	Accuracy (%)		Sensitivity (%)		Specificity (%)		AUC (μ)	
Methods	Best	Av.	Best	Av.	Best	Av.	Best	Av.
RF	86.0	82	91.0	81.0	92.0	85	2.19	0.56
ELM	93.0	87	91.2	85.0	91.0	88	1.87	1.56
RF-ELM	96.7	94	97.0	87.4	95.3	90	2.23	1.11

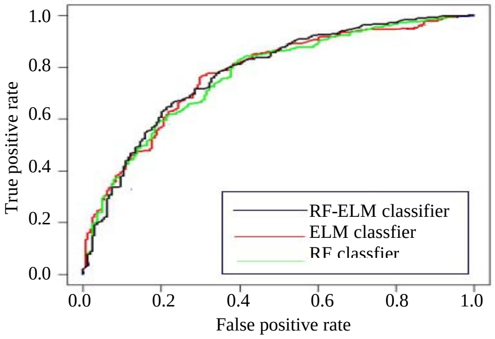

Fig. 4: Comparison of ROC curve for various classifiers using CBF

CONCLUSION

This study proposed a CAD system for breast cancer using RF, ELM and RF-ELM classifier to classify mammograms into normal benign and malignant. The result shows that RF-ELM classifier with CBF feature selection method provides significantly better classification accuracy by reducing the FPs and FNs and it also depends upon the optimization of feature selection. The promising outcomes ought to be because of the segmentation method, the viable feature selection strategy and the incredible classifier. This finding is exceptionally valuable for the radiologist in identifying the malignancy from digital mammograms.

REFERENCES

Ahmad, F., N.A.M. Isa, M.H.M. Noor and Z. Hussain, 2013. Intelligent breast cancer diagnosis using hybrid GA-ANN. Proceedings of the 2013 5th International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN), June 5-7, 2013, IEEE, Madrid, Spain, ISBN:978-1-4799-0587-4, pp: 9-12.

Anonymous, 2015. Study forecasts new breast cancer cases by 2030. National Cancer Institute, National Institutes of Health, USA.

Cedeno, M.A., D.J. Quintanilla and D. Andina, 2011. WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Syst. Appl., 38: 9573-9579.

Cheng, H.D., X. Cai, X. Chen, L. Hu and X. Lou, 2003. Computer-aided detection and classification of microcalcifications in mammograms: A survey. Pattern Recognit., 36: 2967-2991.

Cheng, H.D., X.J. Shi, R. Min, L.M. Hu, X.P. Cai and H.N. Du, 2006. Approaches for automated detection and classification of masses in mammograms. Pattern Recognit., 39: 646-668.

Dheeba, J., N.A. Singh and S.T. Selvi, 2014. Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. J. Biomed. Inf., 49: 45-52.

Forouzanfar, M.H., K.J. Foreman, A.M. Delossantos, R. Lozano and A.D. Lopez *et al*., 2011. Breast and cervical cancer in 187 countries between 1980 and 2010: A systematic analysis. Lancet, 378: 1461-1484.

Jemal, A., F. Bray, M.M. Center, J. Ferlay, E. Ward and D. Forman, 2011. Global cancer statistics. CA: Cancer J. Clin., 61: 69-90.

Jiji, G.W. and J.R. Marsilin, 2012. Automatic diagnose of the stages of breast cancer using intelligent technique. J. Inst. Eng. India Ser. B., 93: 209-215.

Karahaliou, A.N., I.S. Boniatis, S.G. Skiadopoulos, F.N. Sakellaropoulos and N.S. Arikidis *et al*., 2008. Breast cancer diagnosis: Analyzing texture of tissue surrounding microcalcifications. IEEE. Transac. Inf. Technol. Biomed., 12: 731-738.

Kolb, T.M., J. Lichy and J.H. Newhouse, 2002. 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. Radiol., 225: 165-175.

Mohebian, M.R., H.R. Marateb, M. Mansourian, M.A. Mananas and F. Mokarian 2017. A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (HPBCR) using optimized ensemble learning. Comput. Struct. Biotechnol. J., 15: 75-85.

Motakis, E., A.V. Ivshina and V.A. Kuznetsov, 2009. Data-driven approach to predict survival of cancer patients. IEEE. Eng. Med. Biol. Mag., 28: 58-66.

Nugroho, H.A., N. Faisal, I. Soesanti and L. Choridah, 2014. Analysis of digital mammograms for detection of breast cancer. Proceedings of the 2014 International Conference on Computer, Control, Informatics and Its Applications (IC3INA), October 21-23, IEEE, Bandung, Indonesia, ISBN:978-1-4799-4574-0, pp: 25-29.

Siegel, R.L., K.D. Miller, S.A. Fedewa, D.J. Ahnen and R.G. Meester *et al*., 2017. Colorectal cancer statistics, 2017. CA. Cancer J. Clinicians, 67: 177-193.

Xie, W., Y. Li and Y. Ma, 2016. Breast mass classification in digital mammography based on extreme learning machine. Neurocomputing, 173: 930-941.