Original Article

COMPUTER-AIDED MODEL FOR BREAST CANCER DETECTION IN MAMMOGRAMS

ALAA M. ADEL EL-SHAZLI, SHERIN M. YOUSSEF, MARWA ELSHENNAWY

Department of Computer Engineering, Arab Academy for Science and Technology, Alexandria, Egypt
Email: sherin@aast.edu

ABSTRACT

The objective of this research was to introduce a new system for automated detection of breast masses in mammography images. The system will be able to discriminate if the image has a mass or not, as well as benign and malignant masses. The new automated ROI segmentation model, where a profiling model integrated with a new iterative growing region scheme has been proposed. The ROI region segmentation is integrated with both statistical and texture feature extraction and selection to discriminate suspected regions effectively. A classifier model is designed using linear fisher classifier for suspected region identification. To check the system's performance, a large mammogram database has been used for experimental analysis. Sensitivity, specificity, and accuracy have been used as performance measures. In this study, the methods yielded an accuracy of 93% for normal/abnormal classification and a 79% accuracy for bening/malignant classification. The proposed model had an improvement of 8% for normal/abnormal classification, and a 7% improvement for benign/malignant classification over Naga *et al.*, 2001. Moreover, the model improved 8% for normal/abnormal classification over Subashimi *et al.*, 2015. The early diagnosis of this disease has a major role in its treatment. Thus the use of computer systems as a detection tool could be viewed as essential to helping with this disease.

Keywords: Breast Cancer, Classification, Pre-Processing, Segmentation, Sensitivity, Specificity

INTRODUCTION

Breast Cancer is the most significant health problem in the world and early diagnosis has a major role in its treatment. About 1 in 8 U. S. women (about 12%) and 1 in 10 European women (about 10%) will develop invasive breast cancer over the course of her lifetime [1]. It has been shown that early detection and treatment of breast cancer are the most effective methods of reducing mortality [2].

Currently, the most effective method for early detection and screening of breast cancers is mammography [3]. A large amount of mammograms generated by a screening of population must to be diagnosed by relatively few radiologists [4]. Retrospective studies have shown that radiologists can miss the detection of a significant proportion of abnormalities in addition to having high rates of false positives. The estimated sensitivity of radiologists in breast cancer screening is only about 75% [5]. One of the effectively suggested approaches is double reading in order to improve sensitivity and accuracy. In order to improve the accuracy of interpretation, a variety of Computer Assisted Detection (CAD) techniques have been proposed [6]. Developing a CAD algorithm, which uses features extracted from the breast profile region; the region of interest (ROI) is vital. This will cause a reduction in the number of unneeded biopsies to patients with benign masses, and thus reduces healthcare cost and avoids putting the patient in mental and physical stress [7].

A lot of work has been done to detect masses in the mammography images, and different methods were used. Some of the used methods are summarized and explained in this section.

Naga *et al.* [8] have applied Gaussian filter as a preprocessing stage, segmented the breast using multilevel thresholding by extracting the groups of closed contours, and grouping these contours later. The image was classified using Linear Discriminant Analysis (LDA). They have achieved an accuracy of 87% for normal/abnormal

classification, and 74% accuracy for benign/malignant classification. Alireza *et al.* [9] have achieved an average accuracy of 93% by applying the median filter for preprocessing, followed by thresholding techniques for ROI segmentation, and Chebyshev moments for extracting the features for normal/abnormal classification. Amjath *et al.* [10] applied a manual segmentation model, fed the Support Vector Machine (SVM) classifier, features extracted from Discrete Wavelet Transform (DWT), and Discrete Shearlet Transform (DST). This model achieved the highest accuracy for normal/abnormal classification of 92%, and a 80% accuracy for benign/malignant classification when applied a level 3 DST, and calculating each feature in 8 directions for the benign/malignant classification, and when applied a level 4 DST, and calculating each feature in 64 directions for the normal/abnormal classification. Subashini *et al.* [11] have added a very strong preprocessing stage, followed by a manual segmentation of the suspected region, GLCM features were extracted, the Genetic Algorithms (GA) and particle swarm optimization (PSO) are implemented as a feature selection phase, and this was followed by a classification stage using K-Nearest Neighbor classifier, and they achieved an accuracy of 87% for normal/abnormal classification. Indra *et al.* [12] have applied preprocessing stage, detected the breast orientation, and removed the pectoral muscle, using the region-growing algorithm, and achieved an accuracy of 93% for normal/abnormal classification.

Mammograms are medical images that require a preparation stage in order to increase the image quality, remove noise, to make it ready for further processing. The proposed model's main objective is to introduce a new system for automated detection of breast masses in mammography images. The system will be able to discriminate if the image has a mass or not, as well as benign and malignant masses. The block diagram is shown in fig. 1 lists the various steps of the mass detection model, which are explained in details in the following subsections.

Fig. 1: Block diagram of proposed model

MATERIALS AND METHODS

Pre-processing

One of the preliminary stages used for mammogram image enhancement to aid an early identification of suspicious lesions.

Image preprocessing can suppress noise and improve the contrast of the image. There are different types of noises in the mammography images. Firstly, High-intensity noise, such as labels and scanning artifacts. Moreover, tape artifacts, which are markings left by tapes, or other shadows presenting as horizontal running strips. In this stage, the mammograms were preprocessed by the median filtering technique, edge enhancement, erosion and dilation, and contrast enhancement. The median filter was used to remove the blurring effect and speckles from different parts of the image. The speckle noise was removed by replacing it with the local mean value and this is used to preserve edges and features.

In edge enhancement, Gaussian smoothing was used to help with edge smoothing and it works by identifying sharp edge boundaries in the image and increasing the image contrast in the area. This enhances the edge contrast of an image in an attempt to improve its acutance.

Erosion and dilation affects the form, structure or shape of an object. Dilation allows objects to expand, fill holes and contrast disjoint objects. On the other hand, erosion shrinks objects by removing boundaries.

Contrast Limited Adaptive Histogram Equalization (CLAHE) was proposed, which computes several histograms, each corresponding to a distinct section of the image and uses them to redistribute the lightness value of the image. This improves the local contrast and reduces contrast amplification in dense breast tissue.

The difference between the mammogram image before and after the pre-processing stage is shown in fig. 2.
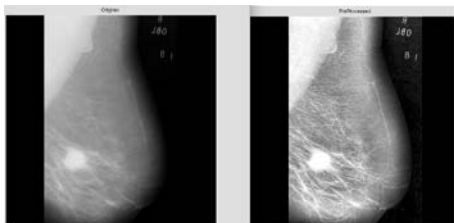

Fig. 2: Image pre-processing stage, a) Original Image b) Pre-processed Image

Proposed ROI segmentation

In computer vision, image segmentation is the process of partitioning a digital image into multiple segments. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze and extract the ROI (Region of Interest) [13]. According to medical experts, the area where the breast lumps appear is the breast profile excluding the background and the pectoral muscle, and this is the ROI to be segmented. The proposed segmentation of ROI (Region Of Interest) is illustrated in 3 steps.

Breast orientation

In the beginning, the foreground breast was separated from the background noise and artifact using thresholding, which was implemented using Otsu's method [14]. Otsu's threshold was calculated using the following algorithm:

1. Compute histogram and probabilities of each intensity level

2. Set up initial class weights and class means

3. Step through all possible thresholds t=1 ... maximum intensity

a. Update class weights and class mean

b. Compute intra-class variance

4. The desired threshold corresponds to the minimum intra-class variance

The breast orientation in each mammogram image needs to be determined. In order to identify the breast profile orientation (left or right) using an automated procedure, the binary image is cropped left to right and then cropped top to bottom, such that the breast profile touches all four borders (left, right, top and bottom) of the image. The breast profile is afterward classified as right or left oriented.

Profiling

In this stage, some extra noises are being removed, including part of the pectoral muscle. A projection profile is a histogram of the number of black pixel values accumulated along rows taken through the image. This is the second step in the proposed segmentation model. After the breast orientation is detected, the nipple and the 2 farthest points in the breast are determined, and the image is profiled in order to remove all the extra noises.

Seeded region growing algorithm

Region growing algorithm is a segmentation method, which is used to segment the pectoral muscle from the breast. It's a pixel-based image segmentation method since it involves the selection of initial seed points. The Algorithm examines neighboring pixels of initial seed points and determines whether the pixel neighbors should be added to the region or not. The process is iterated on, in the same manner as general data clustering algorithms. The seed point was chosen in the pectoral muscle depending on the breast orientation. Based on inspection of all acquired mammogram images a threshold range of 30±5 is identified as the optimum threshold range satisfying all mammogram images reliably to segment the pectoral muscle from the breast profile. The neighborhood pixels of a seed point are examined and pixels within the accepted range will be grouped together. The Pixel is chosen depending on the relative difference in the multiplicative tolerance level region growing:

$$\frac{f(m,n) - \mu R_c}{\mu R_c} \leq T \ \text{........} \ (1)$$

Where $f(m,n)$ is the gray level of the pixel being examined and μR_c is the original seed value. The multiplicative tolerance level determines the maximum gray level deviation allowed within a region [15].

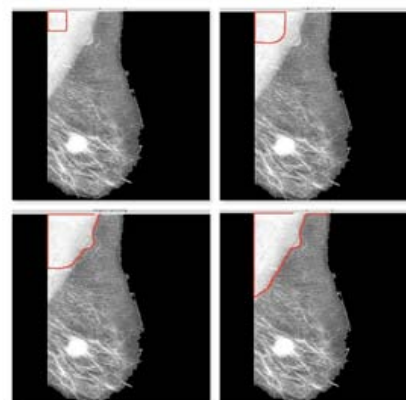Fig. 3 shows different an implementation of the seeded region growing algorithm.

Fig. 3: Seeded region growing algorithm implementation

An example of a preprocessed image before and after the proposed segmentation model is shown in fig. 4.
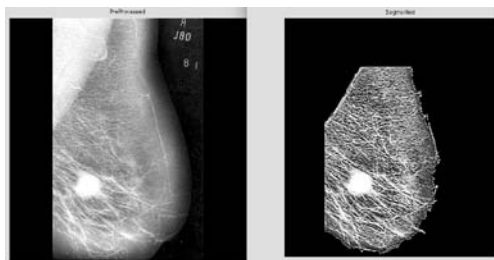
Fig. 4: a) Pre-processed image, b) Segmented image

Feature extraction

The third stage is feature extraction, and those features are calculated on the segmented ROI. Different features are extracted in order to differentiate the Mass/Non-Mass images, and the Benign/Malignant. Certain features were taken into consideration to measure the uniformity, closeness of the distribution, and the local variations. These features are useful in the discrimination between the Mass/NonMass breast. While in the benign/Malignant case, the smoothness and symmetry measure are important in the classification since the benign mass has smooth borders and is more uniform than malignant masses, which have irregular borders.

There is a wide range of features, and only some of them are significant. Using the redundant or large amount of them may degrade the performance of the classifier [16]. 6 features were extracted, 3 are texture features, and the other 3 are intensity features. The implemented intensity features are contrast, standard deviation, and variance. While the texture features that were extracted using gray level co-occurrence matrices (GLCM) were constructed in 4 directions ($0°, 45°, 90°$, and $135°$) with distance 1 are a correlation, homogeneity, and entropy.

Feature selection

Feature selection is the process of selecting the set of feature that results in the best performance measure. There are different approaches for feature selection stage, and in this paper, the Stepwise feature selection method was implemented. Stepwise feature selection is one of the most used methods to reduce the number of features fed to the classifier in order to obtain the best results. The Forward selection approach is implemented, which involves starting with no variables in the model, testing the addition

of each variable using a chosen model comparison criterion, adding the variable (if any) that improves the model the most, and repeating this process until none improves the model [17].

Classification

In our model, after selecting the features, those features will be input to the classifier to classify the image to Mass/NonMass or Benign/Malignant. Classifier such as linear discriminant analysis (LDA) performed well in the classification stage. Linear discriminant analysis (LDA) was implemented in the classification stage and results were obtained and compared.

LDA

LDA is a traditional method for classification. The main idea of this method is to construct the decision boundaries directly by optimizing the error criterion to separate the classes of objects [18]. If there are n classes, and linear discriminant analysis classifies the observations as the following n linear function:

$$g_i(x) = \mu_i C^{-1} x_k^T - \frac{1}{2}\mu_i C^{-1}\mu_i^T + \ln(p_i), 1 \le i \le n \ (4.2)$$

Where x is the features of all data, and each row is denoted by k. μ_i is the mean features in group i, which is an average of x_i. C is the pooled covariance matrix and the p_i is the prior probability vector for each class. The values of C and μ are determined through the analysis of a training set. Once the values of and are determined, they can be used o classify the new observations. In the normal ad abnormal, and benign and malignant classification.

RESULTS AND DISCUSSION

The model was applied to the 322 images (115 Mass/207 NonMass) for the Mass/NonMass classification. While for the Benign/Malignant, 115 images (63 Benign/52 Malignant) were used from the mini-MIAS database (Mammographic Image Analysis Society). The images are gray scaled and of size 1024 x 1024 pixels each. The model was implemented using MATLAB.

The evaluation was carried out in two stages of mass classification system namely a) mass/nonmass classification b)benign/malignant classification.

Firstly, the image is enhanced and preprocessed before the segmentation step. Secondly, the ROI is extracted followed by feature extraction and selection stage. Finally, LDA classifier implemented to classify the image as either normal/abnormal or benign/malignant.

Table 1: Normal/Abnormal classification compared with previous work

Normal/Abnormal			
	Sensitivity	Specificity	Accuracy
Proposed Model	97%	90%	93%
Naga R. *et al.* [8]	93%	69%	87%
Alireza Talebpour *et al.* [9]	92%	94%	93%
Subashini *et al.* [11]			87%
Amjath Ali *et al.* [10]	94%	89%	92%
Indra Kanta *et al.* [12]	94%	93%	93%

In table 1, results of previous work mentioned earlier in the normal/abnormal classification is compared to the proposed model. While in table 2, the performance of the proposed model is compared with other work for the benign/malignant classification.

Table 2: Benign/Malignant classification compared with previous work

Benign/Malignant			
	Sensitivity	Specificity	Accuracy
Proposed Model	86%	73%	79%
Naga R. *et al.* [8]	100%	63%	74%
Amjath Ali *et al.* [10]	68%	95%	80%

CONCLUSION

The estimated sensitivity of radiologists in breast cancer screening is limited, but the performance would be improved if they were prompted with the possible locations of abnormalities. Computer-

aided models for breast cancer detection can provide such help, and they are important and necessary for breast cancer control.

The paper has introduced a new system that integrated an iterative ROI segmentation model with intensity and texture feature

extraction and selection. The proposed model had an improvement of 8% for normal/abnormal classification, and a 7% improvement for benign/malignant classification over Naga *et al.* [8]. Moreover, an improvement of 8% for normal/abnormal classification over Subashimi *et al.* [11].

CONFLICT OF INTERESTS

Declared none

REFERENCES

1. American Cancer Society. "Breast Cancer Facts and Figures"; 2015. Available from: http://www.cancer.org/research/ cancer facts statistics/breast-cancer-facts-figures. [Last accessed on 10 Mar 2016].
2. Michael Brady, Ralph Highnam. Mammographic image analysis; 1999. p. 5620-9.
3. K Bovis, S Singh, J Fieldsend, C Pinder. Identification of masses in digital mammograms with MLP and RBF nets. In: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks Com; 2000. p. 342–7.
4. ML Giger, RM Nishikawa, M Kupinski, U Bick, M Zhang, RA Schmidt, *et al*. Computerized detection of breast lesions in digitized mammograms and results with a clinically-implemented intelligent workstation", in computer assisted radiology and surgery. HU Lemke, K Inamura, MW Vannier. eds. Elsevier, Berlin, Germany; 1997. p. 325–30.
5. RE Bird. Professional quality assurance for mammography screening programs. J Radiol 1990;175:587-605.
6. K Thangavel, M Karnan, R Sivakumar, A Mohidee. Automatic detection of microcalcification in the mammograms-a review. Int J Graphics Vision Image Processing 2005;5:23–53.
7. MJ Islam, M Ahmadi, MA Sid-Ahmed. An efficient automatic mass classification method in digitized mammograms using the artificial neural network. Int J Artificial Intelligence Appl 2010;1:1–13.
8. Naga R Mudigonda, Rangaraj M Rangayyan. Detection of breast masses in mammograms by density slicing and texture flow-field analysis. IEEE Transactions Med Imaging 2001; 20:1215-27.
9. Alireza Alireza, Dooman, Hamid. Automated abnormal mass detection in the mammogram images using chebyshev moments. Res J Appl Sci Eng Technol 2013;5:513-8.
10. Amjath Ali J. Mass classification in digital mammograms are based on discrete shearlet transform. J Comput Sci Syst Biol 2013;9:726-32.
11. Subashini, Bhavani. Detection and classification of masses in mammograms using a hybrid GA-PSO-KNN approach. Int J Adv Res Trends Eng Technol 2015;2:36-40.
12. Indra Sanjay, Prof Samir. Automated digital mammogram segmentation for detection of abnormal masses using binary homogeneity enhancement algorithm. Indian J Computer Sci Eng 2011;2:416-27.
13. Wikipedia. Image segmentation; 2016. Available from: https://en.wikipedia.org/wiki/Image_segmentation. [Last accessed on 10 Mar 2016].
14. Andrew Greensted. Otsu thresholding; 2010. Available from: http://www.labbookpages.co.uk/software/imgProc/otsuThreshold. [Last accessed on 10 Mar 2016].
15. Rangayyan R. Biomedical image analysis. CRC Press: Biomedical Image Analysis; 2005.
16. F Eddaoudi, F Regragui, A Mahmoudi, N Lamouri. Masses detection using SVM classifier based on textures analysis. Appl Mathematical Sci 2011;5:367-79.
17. Mark, Jonathan, Goldberg, Michael A. Multiple regression analysis and mass assessment: a review of the issues. Appraisal J 2001;56:89–109.
18. HD Cheng, XJ Shi, R Min, LM Hu, XP Cai, HN Du. Approaches for automated detection and classification of masses in mammograms. Pattern Recognition J 2005;39:646-68.

How to cite this article