An Approach for Breast Cancer Mass Detection in Mammograms

Walaa Gouda, Mazen M. Selim, T. Elshishtawy

Computer System Engineering Department, Shoubra Faculty of Engineering, Benha Univ., Cairo, Egypt.

Abstract

Breast cancer is one of the major causes of death among women all over the world. An improvement of early detection and diagnosis techniques is very important for women's quality of life. Computer-Aided Detection (CAD) systems have been used for aiding radiologists in their decision in order to solve the limitations of human observers. This paper presents a methodology for mass detection in digital mammograms. This methodology begins with segmenting Regions of Interest (ROIs) using morphological operations and automatic thresholding. Features are extracted from the ROIs and Principal Component Analysis (PCA) is applied for reducing the features dimensionality. Finally, the methodology performs classification through Neural Networks (NNs). The proposed system was tested on several mammographic images extracted from DDSM database. Results showed that the proposed methodology provided more accuracy than other compared techniques.

Keywords: Computer-Aided Detection, mammogram, Regions of Interest, Principal Component Analysis, DDSM.

1. Introduction

Breast cancer is the second most cancer diagnosed among women and the second most cancer deaths in the world. It can be treated by early discovery which can significantly reduce breast cancer mortality. Mammography is at present the best available technique for early detection of breast cancer [1]. The most common breast abnormalities that may indicate breast cancer are masses and calcifications. Masses appear in the mammogram as bright regions of different sizes, margins (circumscribed, micro lobular, obscured, indistinct, and spiculated), shapes (round, oval, lobular, irregular), and gray-level intensities and contrasts that depend on their surrounding tissues. These masses are called tumors and can be either cancerous "malignant" or non-cancerous "benign". Round and oval shaped masses with smooth and circumscribed margins usually indicate benign changes. On the other hand, a malignant mass usually has a spiculated, rough and blurry boundary. They can't be recognized from the surrounding parenchyma because their features can be similar to the normal inhomogeneous breast tissues [2; 1]

Missed detections may be attributed to several factors including poor image quality, subtle nature of radiographic findings, eye fatigue, or oversight. This makes the automatic

mass detection and classification a challenging task for both radiologists and CAD systems [2]. The principal difficulty in this task is the lack of a single algorithm that produces good results for all images. To address this problem; several image processing techniques have been employed. These techniques have been shown to be useful as a second opinion to radiologist for breast cancer detection on mammograms [3]. The general methodology for diagnosis of breast masses is shown in figure 1:

Figure 1 General methodology for diagnosis

The proposed system intends to classify breast tissues in mammography into mass and non-mass groups. The system starts by applying image enhancement techniques to improve images brightness so that features can be easily located and recognized. Segmentation is further performed to extract ROIs. Texture, intensity, and geometric features are then extracted from ROIs and PCA is applied for dimensionality reduction. Finally, classification is performed using backpropagation NN classifier.

This paper is organized as follows; section 2 presents recent works about detection and diagnosis of masses in mammogram images. Section 3 presents the proposed system. Section 4 shows the experimental results. Finally, section 5 gives the concluding remarks.

2. Related Work

Recent work aims to develop computer aided breast cancer detection and diagnosis techniques. The research made by *Liu, et el.* [4] made a classification of masses with level set segmentation and multiple kernel learning. Morphological features were extracted from the boundary of segmented regions. Linear discriminant analysis, support vector machine (SVM) and multiple kernel learning were investigated for the final classification. Their method achieved an accuracy of 76%. Another method developed by *Yu Zhang, et el.* [5] presented a novel segmentation method for identifying mass regions in mammograms. For each ROI, an enhancement function was applied proceeded with a filters function to reduce noise. Next, energy features based on the co-occurrence matrix of pixels were computed. These energy features were used to extract the contour of the mass using an edge-based segmentation technique. While *Mariusz Bajger, et el.* [6] presented an automatic method for the detection of mammographic masses. This method utilized statistical region merging for segmentation

(SRM) and linear discriminant analysis (LDA) for classification. Their results showed that the area under the Receiver Operating Characteristic (ROC) curve value for classifying each region was 0.96. Another technique proposed by *Bong-ryul Lee, et el.* [7] performed mass segmentation by applying region growing and morphological operations. The sensitivity was 78% at 4 FP/image.

Boujelben, et el. [8] proposed another approach to extract convexity and angular features based on boundary analysis. Their approach used Multilayer Perception (MLP) and k-Nearest Neighbors (kNN) classifiers to distinguish between the pathological records and the healthy ones. The results showed 94.2% sensitivity and 97.9% specificity. *L.d.O. Martins, et el*. [9] developed a methodology for masses detection on digitized mammograms using the K-means algorithm for image segmentation. Co-occurrence matrixes were used to describe the textures of segmented structures. The classification was performed using SVM which separates features into two groups, using shape and texture descriptors. Their method showed 85% accuracy.

Rahmati, et el. [10] presented a region-based active contour approach to segment masses in digital mammograms. The algorithm used a Maximum Likelihood approach based on the calculation of the statistics of the inner and the outer regions. The results demonstrated an average segmentation accuracy of 81% for 100 test images. *L.O. Martins et el*. [11] presented a mass detection method using growing neural gas algorithm to perform segmentation. For each segmented region, shape measures were computed in order to discard bad mass candidates. Texture measures on the other hand were obtained from Ripley's K function and a SVM classifier were used for classification. Their method provided an accuracy rate of 89.30%. *GAO, et el.* [12] proposed another mass detection scheme based on the SVM and the relevance feedback. The sensitivity of the SVM classifier rose to 90.6% and the false positive was equal to 3.6 marks per image.

We may observe that there is a need for methodologies that provide support to automatic detection of lesions in mammogram images with little or no specialist participation. Such objective is a great challenge for the segmentation methods because of the dependability on the characteristics of objects.

3. The proposed system

The proposed system will follow the same phases as in figure 1 and will be discussed in the following subsections.

3.1 Image Preprocessing

Image enhancement techniques are used to improve an image, improvement is sometimes defined objectively e.g., increase the signal-to-noise ratio, and sometimes subjectively e.g., make certain features easier to see by modifying the colors or intensities. Intensity adjustment is an image enhancement technique that maps an image's intensity values to a new range [13]. So that image enhancement has to be applied to mammographic images in order to reduce the effect of noise and improve the accuracy of detecting early signs of breast cancer. In this work, the mean and average filters were used to adjust and enhance the image brightness, color and contrast to optimum levels. Figure 2 shows the original image and the enhanced image.

(a) (b)

Figure 2: (a) Original image (b) Enhanced image

3.2 Mass Segmentation

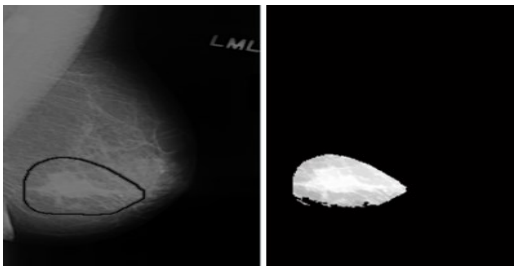
In order to detect suspicious regions, several techniques have been used. Global thresholding is used as a primary step for segmentation. Global thresholding is based on the global information, such as histogram. It selects a single threshold value from the histogram of the entire image. After a global thresholding value is determined, the objects can be separated from the background.

The histogram usage shows that, regions with an abnormality impose extra peaks while a healthy region has only a single peak, so the fact that masses usually have greater intensity than the surrounding tissue can be used for finding global threshold value. By

locating a threshold value, the regions with abnormalities can be segmented. Because masses are often superimposed on the tissue of the same intensity level global thresholding were used as a primary step. The output of the global thresholding is mainly used as an input to the next step in most of systems [14; 15; 1].

In addition, morphological operations were used to suppress structures that are lighter than their surroundings and that were connected to the image borders. It also works as a tool for extracting or modifying structure of objects within an image [15]. Basic morphological operators, such as dilation and erosion, which are the basic operations of morphology as all other operations are built from a combination of these two, are particularly useful for the analysis of binary images, although they can be extended for use with gray scale images. Dilation adds pixels to the boundaries of objects in an image, while erosion removes pixels on object boundaries. The number of pixels added or removed from the objects in an image depends on the size and shape of the structuring element used to process the image [16; 15; 17].

Furthermore, Automatic thresholding using Otsu's method [18] is further implemented where it performs histogram shape-based image thresholding or, the reduction of a gray level image to a binary image. The algorithm assumes that the image to be thresholded contains two classes of pixels (e.g. foreground and background) then calculates the optimum threshold separating those two classes so that their combined is minimal**.** The original image and a sample output are shown in figure 3.



(a) (b)

Figure 3: (a) Original image (b) Detected suspicious regions

3.3 Features Extraction

The features extraction is the key step in mass detection since the performance of CAD depends more on features selection than classification. The feature space is very large and complex due to the wide variety of normal tissues and abnormalities. Further, some features are more significant than others. In our work, several features are extracted from

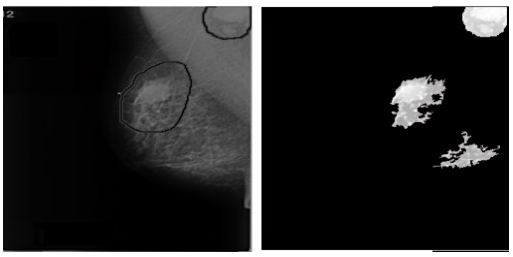
ROIs and features extracted from the gray level characteristics, (intensity), shape, and texture of the lesion and the surrounding tissue can usually be expressed as a mathematical description, and are helpful for a classifier to distinguish masses as malignant or benign like average intensity (mean), average contrast (standard deviation), convexity, area, skewness, energy, entropy, variance, etc., but, it is very difficult to predict which feature or features combinations will achieve better classification rate [14; 1].

3.4 Feature Selection

One often faces with the task of selecting an optimized subset of features from a large number of available features. PCA [19; 13] is used to select the most important ones for the classification of mass as benign or malignant, it has three effects, it orthogonalizes the components of the input vectors i.e. they are uncorrelated with each other, it orders the resulting orthogonal components (principal components) so that those with the largest variation come first, and eliminates those components that contribute the least to the variation in the data set.

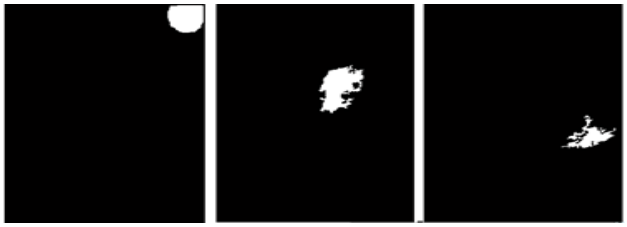
The success of a classification depends largely on the features selected and their role in the model rather than redundant features which should be removed to improve the classifier performance. Generally, different features combinations will result in different performance. In addition, relatively few features used in a classifier can keep the classification performance robust [14].

Our proposed system uses 36 features (5 intensity features, 10 geometric features and 21 texture features) that are extracted from the suspicious detected regions. The extraction of some texture characteristics from structures is done using co-occurrence matrix, where co-occurrence matrix is a tabulation of how often different combinations of pixel brightness values, gray levels, occur in a pixel pair in an image. For images containing more than one ROI as shown in figure 4, Each ROI is segmented as shown in figure 5 and described by the same number of features.



(a) (b)

Figure 4 (a) Original Image, (b) 3 detected objects (ROIs)



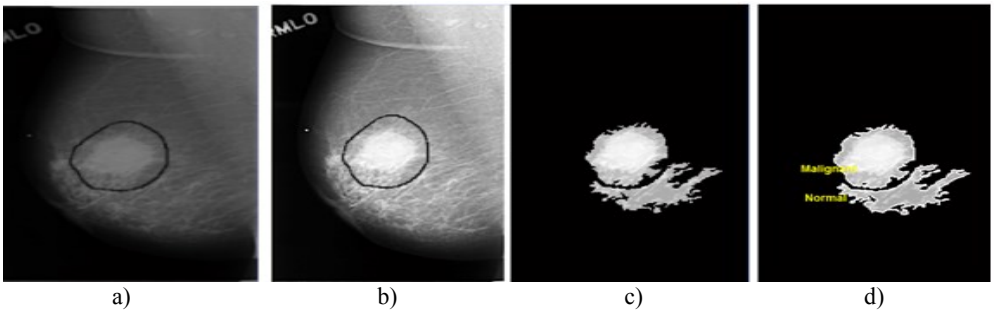
(a) (b) (c)

Figure 5 (a), (b) and (c) are the three detected objects each is separated in a new image

Due to the large dimensionality of the input vector, PCA is used for reduction by performing a covariance analysis between features without much loss of information. It generates a new set of variables, called principal components, such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information. The number of selected principal components is chosen as a compromise between training efficiency (few PCA components) and accuracy (a large number of PCA components) [19; 13].

3.5 Classification using Neural Network

Neural networks (NNs) are one of the major classification approaches. The proposed system used Feed Forward Back Propagation (FFBP) network [13] for classifications. In our work, the samples data is divided into 3 groups: training, testing and cross validation. Once the network is trained the weights are then frozen. Once trained, the testing set is fed into the network and the network output is compared with the desired output. A cross validation is used during the training process to prevent the NN from over fitting. Figure 6 shows the effect of the proposed methodology on a sample image from DDSM.



a) b) c) d)

Figure 6 a) Original Image, b) Enhanced Image, c) Segmented Object and d) Classification Result

4 Experimental Results

The proposed system was tested using a set of images, which are selected from Digital Database for Screening Mammography (DDSM) database [20]. Cases with mass lesions are selected through the reports that only included the BI-RADS descriptors for mass margin and mass shape. Our test set consists of 715 cases (including 241 cases with malignant, 222 cases with benign masses and 250 cases with no mass) that are selected out of 2262 cases based on the BI-RADS criteria.

After detecting ROIs and extracting the required features, PCA is used to reduce their dimensionality; the output is then passed to the classifier. The number of features used isn't known, as it depends on the classifier parameters which are the weight, number of hidden layers and number of neurons per each layer. In this work, the weights are chosen randomly.

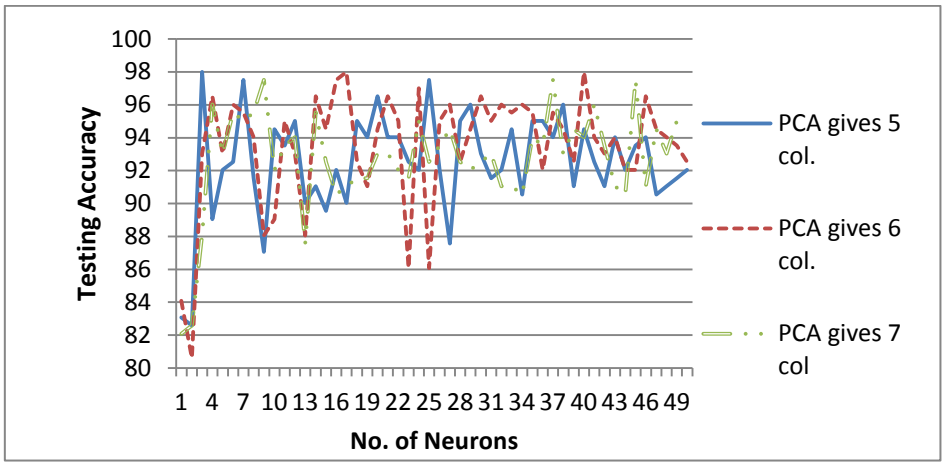


Figure 7 results from FFBP at different PCA

Figure 7 shows the best results from using FFBP as a classifier. It shows testing accuracy against the spread or radius. Three curves are drawn corresponding to the different PCA values as by changing the total variance, the number of selected principal components which represents number of columns will change.

Table 1 best obtained results

No. of Neurons	Testing Accuracy	TP	FP	TN	FN	Sensitivity	FPR
3	98.00995025	37	2	160	2	94.87179487	1.23456
78	98.00995025	39	0	158	4	90.69767442	0

Table 1 shows the best results obtained when the input size is reduced by PCA to 5 columns or features, while using one hidden layer and using Gradient descent with momentum and adaptive learning rate backpropagation. Table 2 shows the result of the proposed system compared to others.

Table 2 DDSM Comparison

Reference	Sensitivity
F. Zou, et el. **[21]**	**82.6%**
GAO, et el. **[12]**	**90.6%**
L.d.O. Martins, et el. **[9]**	**86%**
L.O. Martins, et el. **[11]**	**89.30%**
Jing, et el. **[22]**	**94%**
Bong-ryul Lee, et el. **[7]**	**78%**
Boujelben, et el. **[8]**	**94.2%**
Liu, et el. **[4]**	**76%**
Proposed Methodology	**94.87%**

The use of geometrical, intensity and texture measures to characterize the segmented objects presents some advantages in relation to other approaches. Usually, just the geometric or shape information isn't enough to completely describe the mass, since a great part of its characterization comes from its texture and intensity. Besides, the use of texture measures discards most of the objects that represent healthy tissues. In this work, the result obtained from training NN using each type of feature alone, also using the combination of each two types of features and also the combination of the three types of features is shown in figure 8.

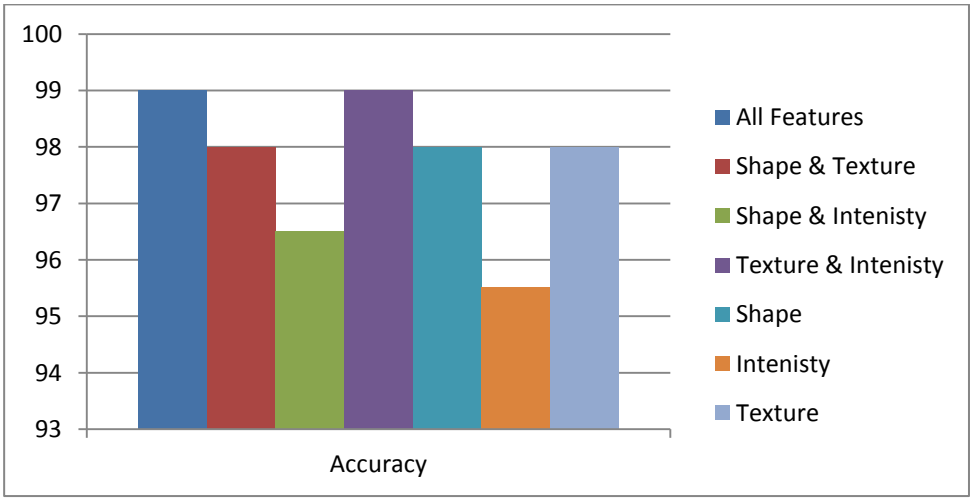


Figure 8 the obtained testing accuracy with changing used features

Thus, through the analysis of the results, it is possible to conclude that the proposed methodology proved to be effective in detecting masses in screening mammograms. The presented results allow us to infer that the use of the local thresholding, morphological operations and global thresholding (Otsu's method) in the task of segmenting screening mammograms provides a good rate of correct segmentation of mass structures. Similarly, the use of the co-occurrence matrix allows describing textures, intensity and shape descriptors in an efficient way, thus contributing to the correct recognition of segmented structures, also it is noticed that the usage of shape features in our proposed system aren't effective as the performance isn't affected by it. Furthermore, the FFBP classification provides good generalization, also contributing to the effectiveness of the methodology during the classification of segmented structures.

5 Conclusion:

The use of computational tools to aid detection and diagnosis of breast masses has grown and gained increasing acceptance in recent years, as a kind of second readers of medical images. These tools have been contributing to increase the early detection rates for breast cancer. This paper presented a methodology for detection of masses in digital screening mammograms, which can also be used in the development of a CAD tool. The results indicate that the use of these techniques in the detection of masses is promising; since it achieves good rates of accuracy, also indicate that the usage of PCA in selecting features gives good results. Further researches can be done in the development of a CAD system capable of assisting health professionals in the painstaking task of tracing mammograms in search of mass abnormalities.

References

1. *A Survey of Image Processing Algorithms in Digital Mammography.* **Jelena Bozek, Mario Mustra, Kresimir Delac, Mislav Grgic.** Germany : s.n., 2009, Recent Advances in Multimedia Signal Processing and Communications, pp. 631 – 657.

2. *Computer-Aided Detection and Classification of Masses in Digitized Mammograms.* **Islam, Mohammed Jahirul.** 2009. RESEARCH CENTRE FOR INTEGRATED MICROSYSTEMS - UNIVERSITY OF WINDSOR.

3. **Heang-Ping, Chan.** *Development of an Advanced Computer-Aided Diagnosis System for breast cancer detection.* Michigan : s.n., 2006.

4. *Mass Classification in Mammography with Morphological Features and Multiple Kernel Learning.* **Liu, X., Z.Feng and J. Liu.** 2011, Bioinformatics and Biomedical Engineering, pp. 1 - 4.

5. **Yu Zhang, Noriko Tomuro, Jacob Furst and Daniela Stan Raicu.** A Contour-based Mass Segmentation in Mammograms. *ScientificCommons.* [Online] 2010.

6. *Mammographic Mass Detection with Statistical Region Merging.* **Mariusz Bajger, Fei Ma, Simon Williams & Murk Bottema.** 2010, Digital Image Computing: Techniques and Applications, pp. 27-32.

7. *Automated recommendation of initial mass positions for mass segmentation in digital mammograms.* **Bong-ryul Lee, Jong-doo Lee and Myeong-jin Lee.** 2010, Electronics and Information Engineering (ICEIE), 2010 International Conference On , pp. V2-202 .

8. *Feature Extraction from Contours Shape for Tumor Analyzing in Mammographic Images.* **Boujelben, A. Chaabani, A.C. Tmar, H.and Abid, M.** 2010, Digital Image Computing: Techniques and Applications, 2009. DICTA '09. .

9. *Detection of Masses in Digital Mammograms using K-means and Support Vector Machine.* **L.d.O. Martins, G.B. Junior,A.o.C. Silva, A.C.de Paiva & M. Gattass.** 2009, Electronic Letters on Computer Vision and Image Analysis, pp. 39-50.

10. *Maximum Likelihood Active Contours Specialized for Mammography Segmentation.* **Rahmati, P. and Ayatollahi, A.** 2009, Biomedical Engineering and Informatics, 2009. BMEI '09. 2nd International Conference on .

11. *Detection of Breast Masses in Mammogram Images Using Growing Neural Gas Algorithm and Ripley's K Function.* **L.O. Martins, A. C. Silva, A. C. Paiva and M. Gattass.** 2009, Journal of Signal Processing Systems, pp. 77-90.

12. *Mass detection algorithm based on support vector machine and relevance feedback.* **GAO, Ying WANG & Xinbo.** 3, 2008, Higher Education Press and Springer-Verlag, Vol. 3, pp. 267–273.

13. *The MathWorks Tutorial.* s.l. : The MathWorks, Inc., 2008.

14. *Approaches for automated detection and classification of masses in mammograms.* **H.D. Cheng, X.J. Shi, R. Min, L.M. Hu, X.P. Cai and H.N. Du.** 2006, Pattern Recognition, pp. 646 – 668.

15. **Woods, Rafael C. Gonzalez & Richard E.** *Digital Image Processing (3rd Edition).* Upper Saddle River, NJ, USA : Prentice Hall, 2008.

16. **McAndrew, Alasdair.** *An Introduction to Digital Image Processing with Matlab, semester 1.* Victoria University of Technology : School of Computer Science and Mathematics, 2004.

17. **Dougherty, Geoff.** *Digital Image Processing for Medical Applications.* California State University, Channel Islands : CAMBRIDGE UNIVERSITY PRESS, 2009.

18. *A Threshold Selection Method from Gray-Level Histograms.* **Otsu, N.** 1979, IEEE Transactions on Systems, Man, and Cybernetics, pp. 62-66.

19. **Jolliffe, I. T.** *Principal Component Analysis.* s.l. : 2nd edition, Springer, 2002.

20. *The Digital Database for Screening Mammography.* **Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore,and W. Philip Kegelmeyer.** 2001, Proceedings of the Fifth International Workshop on Digital Mammography, M.J. Yaffe, ed., Medical Physics Publishing, pp. 212-218.

21. *Gradient Vector Flow Field and Mass Region Extraction in Digital Mammograms.* **F. Zou, Y. Zheng, Z. Zhou & K. Agyepong.** 2008, 21st IEEE International Symposium on Computer Based Medical Systems, pp. 41-43.

22. *Mass Detection in Digital Mammograms Using Twin Support Vector Machine-based CAD System* . **Jing, Xiong Si & Lu.** 2009, WASE International Conference on Information Engineering, pp. 240-243.