DEVELOPMENT OF A COMPUTER-AIDED CLASSIFICATION SYSTEM FOR CANCER DETECTION FROM DIGITAL MAMMOGRAMS

Mohamed A. Alolfe[1], Abo-Bakr M. Youssef[1], Yasser M. Kadah[1], and Ahmed S. Mohamed[1]
[1] System & Biomedical Engineering Department, Cairo University, Giza, Egypt
E-mail: al_olfe2001@k-space.org

Abstract-Mammogram–breast x-ray is considered the most effective, low cost, and reliable method in early detection of breast cancer. Although general rules for the differentiation between benign and malignant breast lesion exist, only 15 to 30% of masses referred for surgical biopsy are actually malignant. Computer-Aided Classification system was used to help in diagnosing abnormalities faster than traditional screening program without the drawback attribute to human factors. In this work, an approach is proposed to develop a computer-aided classification system for cancer detection from digital mammograms. The proposed system consists of three major steps. The first step is region of interest (ROI) extraction of 256×256 pixels size. The second step is the feature extraction; we used a set of 88 features and we found that 78 of these feature are capable of differentiating between normal and cancerous breast tissues in order to minimize the classification error. The third step is the classification process; we used the technique of the k-Nearest Neighbor (k-NN) to classify between normal and cancerous tissues. The proposed system was shown to have the large potential for cancer detection from digital mammograms.

1. INTRODUCTION

Breast cancer is a leading cause of fatality among all cancers for women. However, the etiologies of breast cancer are unknown and no single dominant cause has emerged. Still, there is no known way of preventing breast cancer but early detection allows treatment before it is spread to other parts of the body. Currently, X-ray mammography is the single most effective, low-cost, and highly sensitive technique for detecting small lesions resulting in at least a 30% reduction in breast cancer deaths [1]. However, the sensitivity of mammography is highly challenged by the presence of dense breast parenchyma, which deteriorates both detection and characterization tasks [2].

It may not be feasible to routinely perform a second reading by a radiologist due to financial, technical, and logistical restraints. Therefore, efforts were made to develop a computer-aided detection (CAD) system. [3],[4] CAD can be defined as a diagnosis made to improve radiologists' performance by indicating the sites of potential abnormalities, to reduce the number of missed lesions, and/or by providing quantitative analysis of specific regions in an image to improve diagnosis. CAD systems typically operate as automated "second-opinion" or "double reading" systems [5].

Many techniques have been used to detect masses in the mammograms. Youssry *et al*. [6] used a technique that depends mainly on the difference between normal and cancerous histograms and used four features for the classification process through a neural network classifier. The four features are statistical ones which are the mean and the first three moments. Preprocessing techniques were used such as histogram equalization and segmentation. Kobatake *et al*. [7] presented a Computer-Aided Diagnostic (CAD) system for the detection malignant tumors on digital mammograms, and used 9 features to identify malignant tumors. Yu *et al*. [8] presented a CAD system for the automatic detection of clustered microcalcifications through two steps. The first one is to segment potential microcalcification pixels by using wavelet and gray level statistical features and to connect them into potential individual microcalcification objects. The second step is to check these potential objects by using 31 statistical features. Neural network classifiers were used. Results are satisfactory but not

highly guaranteed because the learning set was used in the testing set. Verma *et al*. [9] presented technique that based on fuzzy-neural and feature extraction, and used 14 features for the proposed method. A fuzzy technique in conjunction with three features was used to detect a microcalcification pattern and a neural network to classify it into benign or malignant. Hagihara *et al*. [10] presented a CAD for breast cancers by improvement of classifiers, used 5 features related to the concurrency matrix, 3 features related to the density histogram, and one feature related to the shape of the extracted region. Furuya *et al*. [11] improvement of performance to discriminate malignant tumors from normal tissue on mammograms by feature selection and evaluation of features selection criteria, and used four types features, first-order statistics features, second-order statistic (co-occurrence) features, density features, and shape features. Fogela *et al*. [12] used the patient age as a feature besides radiographic features to train artificial neural networks to detect breast cancer. Brake *et al*. [13] studied the scale effect on the detection process by using single scale and multi-scale detection algorithms of masses in digital mammograms. The implementation of such techniques had to use different sets of features which used to differentiate between normal and cancer breast tissue from digital mammograms.

In this paper, the development of a computer-aided classification system for cancer detection from digital mammograms is presented. A new computer-aided classification system used most effective sets of features that can differentiate between normal and cancer breast tissue. The proposed system consists of three major steps: The first step is ROI extraction of 256×256 pixels size. The second step is the feature extraction, where a set of 88 features are used and that only 78 of these features are capable of differentiating between normal and cancerous breast tissues are found. The third step is the classification process; the technique of the k-Nearest Neighbor (k-NN) is used to classify between normal and cancerous tissues.

2. MATERIALS & METHODS

This study is done through two main phases; the learning phase and the testing phase. Through the learning phase, the system how to differentiate between normal and cancerous cases is learned by using normal and cancerous images. In the testing phase, the performance of the system is test by entering a test image to compute the correctness degree of the system decision. The Figure (1) shows a schematic diagram for the proposed system.

2.1. Mammogram database:

The mammogram images used in this paper are provided by the University of South Florida, the digital database for screening mammography (DDSM) [14]. The dataset consists of digitized mammogram images, composed of both oblique and cranio-caudal views. Each mammogram shows one or more tumor mass marked by expert radiologists. The position of individual masses is marked. The location of the abnormalities in form of its boundary provided as chain code where the first two values are the starting column and row of the lesion boundary while other numbers correspond to a specific direction on the X and Y coordinates. The images are digitized from films using the Lumysis scanner with 12 bits depth.

2.2. Selection of ROI:

Using the contour supplied by the DDSM for each mammogram, the ROI of size 256×256 pixels is extracted with mass centered in the window, and divided into two sets: the learning set and the testing set. The learning set is composed of 88 cancerous images and 88 normal images while the testing set contained 57 cancerous images and 32 normal images. The normal images are taken from the same image that has cancerous regions.

2.3. Feature extraction:

A typical mammogram contains a vast amount of heterogeneous information that depicts different tissues, vessels, ducts, chest skin, breast edge, the film, and the X-ray machine characteristics. In order to build a robust diagnostic system towards correctly classifying normal and abnormal regions of mammograms, we have to present all the available information that exists in mammograms to the diagnostic system so that it can easily discriminate between the normal and the abnormal tissue. However, the use of all the heterogeneous information, results to high dimensioned feature vectors that degrade the diagnostic accuracy of the utilized systems significantly as well as increase their computational complexity. Therefore, reliable feature vectors should be considered that reduce the amount of irrelevant information thus producing robust Mammographic descriptors of

compact size. In our approach, we examined a set of 88 features were applied to the ROI using a window of size 64 pixels with 64 pixels shift, i.e. no overlap.

The features extracted in this study divided into four categories: first order statistics features, second order statistics (gray level co-occurrence matrix) features, shape features, and fractal dimension features.

1) *First order statistics features:* provides different statistical properties of the intensity histogram of an image [15]. They depend only on individual pixel values and not on the interaction or co-occurrence of neighboring pixel values. In this study, first order textural features were calculated: mean value of gray levels, standard deviation of gray levels, kurtosis, skewness, variance, maximum of gray level, range of gray level, entropy, second moment, and percentile.

2) *Second order statistics (gray level co-occurrence matrix) features:* The gray level co-occurrence matrix (GLCM) is a well-established robust statistical tool for extracting second order texture information from images [16], [17]. The GLCM characterizes the spatial distribution of gray levels in the selected ROI subregion. An element at location (i,j) of the GLCM signifies the joint probability density of the occurrence of gray levels i and j in a specified orientation θ and specified distance d from each other. Thus, for different θ and d values, different GLCMs are generated. In this study, four GLCMs corresponding to four different directions (θ=0°, 45°, 90° and 135°) and one distance (d=1 pixel), were computed for each selected ROI subregion. Sixteen features were derived from each GLCM. Specifically, the features studied are: energy, contrast, homogeneity, correlation, first order difference moment, entropy of co-occurrence, maximum of co-occurrence, shade, prominence, second order inverse difference moment, information correlation 2, sum of squares, sum average, sum entropy, difference entropy. Four values were obtained for each feature corresponding to the four matrices.

3) *Shape features:* provide information's on the shape of ROI. Eight features were extracted, spreadness; this feature shows the degree of spread of the shape around the centered intuitively, i.e. measure the circularity of ROI [7], and seven invariant moments; these features a set of moments is invariant to translation, rotation, and scale change [15].

4) *Fractal dimension features:* A fractal is an irregular geometric object with an infinite nesting of structure at all scale. Some of the most important properties of fractals are self-similarity, chaos, and non-integer fractal dimension (FD). The FD offers a quantitative measure of self-similarity and scaling. The fractal dimension can be defined as the exponent of the number of self-similar pieces (N) with magnification factor ($1/r$) into which a figure may be broken. The equation for FD is as (1):

$$FD = \frac{\ln(N)}{\ln\left(1/r\right)} \tag{1}$$

In this study two methods which had taken to estimate the fractal dimension feature: The piecewise modified box-counting (PMBC) and the piecewise triangular prism surface area (PTPSA) methods. In PMBC method, the image of size $M \times M$ pixels is scaled down to a size $r \times r$ where $M/2 \geq r > 1$ and r is an integer. Consider the image $i(x, y)$ as a 2D plane and the pixel intensity p as the height above a plane. Thus, image $i(x, y)$ is partitioned into grids of size $r \times r$ and on each grid there is a column of boxes of size $r \times r \times p$. Assume that the maximum and minimum gray levels of the image $i(x, y)$ in *(i, j)-th* grid fall in box number k and l respectively. Then $n_r(i, j) = k - l + 1$ is the contribution of N_r in the *(i, j)-th* grid. The contributions from all the grids using the equation (2):

$$N_r = \sum_{i,j} n_r(i, j) \tag{2}$$

Where N_r is computed for different values of the square of size r. The FD of an image is calculated from the slope of the linear regression line obtained when the horizontal axis and the vertical axis are taken as log $1/r$ and log N_r. And in the PTPSA method used the grayscale elevation values at the corners of a box at points (*A, B, C*, and *D*), and the average value of the corners as center elevation value at point (*E*) forms four triangular (ABE, BCE, CDE, and DAE). By repeating this calculation for different box size r, the logarithm of surface areas of the top triangular surfaces versus logarithm of the box size is calculated to obtain the slope (FD) [18].

2.4. Feature selection:

The purpose of this step is to get the features that have the ability of differentiation between normality and cancer to be used in the classification process. In other words, the discrimination power of the features is test. The input to this test is two sets of values for each feature. One set represents the normal case and the other set represents the cancerous case. We assume that each set follows a t distribution. The t-test checks the amount of overlapping between the two distributions. If there is no overlapping, then this feature has the ability of differentiation. But in nature, it is not easy to find complete independent distributions without overlapping. So, we determine a significance level to consider the two sets come from two different distributions. We chose this significance level to be 5 %. It means that the probability of incorrectly considering two independent distributions is 0.05 while the truth is that the two sets come from the same distributions. The test computes a value called the p-value which is the probability of observing one sample from the first set in the second distribution. If the p-value is less than the significance level, then these two sets come from two different distributions and this feature can differentiate [19].

To prepare the two sets of each feature, the feature matrix resulted from the step of features extraction is used. For each feature, we transfer the matrix of each image to a vector. Thus, we have for each feature a number of vectors equal to the numbers of the sample normal and cancerous images. These vectors are concatenated under each other to form the normal cluster and the cancerous cluster, and these two sets are the input to the t-test step.

The previous process was done for the 88 features to test their discrimination power to avoid using non-classifying features to reduce the classification error.

2.5. Classification:

The classification process is divided into the learning phase and the testing phase. In the learning phase, known data are given and the feature parameters are calculated by the processing which precedes classification. Separately, the data on a candidate region which has already been decided as a tumor or as normal are given, and the classifier is trained. We used the learning set for this phase which consists of 88 cancerous ROI and 88 normal ROI. In the testing phase, unknown data are given and the classification is performed using the classifier after learning. Breast cancer image diagnosis assistance is the task in the testing phase. We used a testing set for this phase which consisted of 57 cancerous ROI and 32 normal ROI. The Voting K-Nearest Neighbor (K-NN) classifier is used [20].

The Voting k-Nearest Neighbor (k-NN) classifier is nonparametric technique, it assigns a test sample to the class of the majority of its K-neighbors; that is, assuming that the number of voting neighbors is $k=k_1+k_2+k_3$ (where k_i is the number of samples from class i in the k-sample neighborhood of the test sample), the test sample is assigned to class m if $k_m = \max \{k_i, i=1, 2, 3\}$[20].

The features of the sample images forming each cluster are not concatenated under each other. Instead, they are left separately through the cluster. For the test image, we calculate the features vector of size M×1. Then, we get the distance between this vector and every sample image in the two clusters. After that, we sort these distances in ascending order. With the choice of k, we assign the test sample to its class. The value of k must be odd. If $k = 1$, the first distance is the smallest one and we classify the test sample to be from the cluster having the learning sample of the minimum distance. With $k = 3$, the test sample is classified to be belonging to the cluster that has 2 or 3 distances from the minimum 3 distances in the ascending vector. Through this study, we compared the results of using $k = 1$, $k = 3$, and $k=5$.

The cluster is formed from some vectors; each one is of size 78×1. The number of these vectors equal the number of sample images used in learning the system. So, the normal cluster is composed from 88 vectors of size 78×1. The cancerous cluster is composed from 88 vectors of size 78×1.

3. RESULTS & DISCUSSIONS

3.1. Feature Extraction and selection:

The previously mentioned 88 features using a window size of 64 pixels and a window shift of 64 pixels i.e. no overlap is applied. Features are tested using a hypothesis test to decide whether or not this feature can discriminate between normal and abnormal tissues using a significance level of 0.05. The hypothesis indicated that only 10 features (kurtosis, spreadness, 6^{th} moment, 7^{th} moment, information correlation2 at angle 0^o, information correlation2 at angle 45^o, information correlation2 at angle 90^o, and information correlation2 at angle 135^o, difference entropy at angle 45^o, and sum entropy at angle 45^o) can't discriminate between the two clusters because their p-value is larger than the significance level of 0.05.

3.2. Classification:

We measured, quantitatively, the detection performance of the classifiers by computing the sensitivity and specificity on the data. Sensitivity is the conditional probability of detecting cancer while there is really cancer in the image. Specificity is the conditional probability of detecting normal breast while the true state of the breast is normal.

In the terms of the false-negative rate and the false-positive rate:
Sensitivity = 1- false-negative rate.
Specificity = 1- false-positive rate.

False-negative rate: the probability that the classification result indicates a normal breast while the true diagnosis is indeed a breast disease (i.e. positive). This case should be completely avoided since it represents a danger to the patient.

False-positive rate: the probability that the classification result indicates a breast disease while the true diagnosis is indeed a normal breast (i.e. negative). This case can be tolerated, but should be as infrequent as possible.

Table (1) shows the results of the voting k-NN classifier with varying the value of *k* to take 1, 3, and 5.
By testing the learning set and using the *k* value of 1, the system detected 88 images from 88 cancerous images and detected 88 images from 88 normal images. This gives a sensitivity of 100 % and a specificity of 100 %. By using the *k* value of 3, the system detected 63 images from 88 cancerous images and detected 68 images from 88 normal images. This gives a sensitivity of 71.59% and a specificity of 77.27 %. By using the *k* value of 5, the system detected 65 images from 88 cancerous images and detected 67 images from 88 normal images. This gives a sensitivity of 75% and a specificity of 76.14 %.

Table (1): Results for voting K-NN classifier

Parameter	Voting k-Nearest Neighbor (k-NN) classifier					
	K=1		K=3		K=5	
	Learning set	Testing set	Learning set	Testing set	Learning set	Testing set
Sensitivity	100%	71.93%	71.59%	70.18%	75%	68.42%
Specificity	100%	75%	77.27%	50%	76.14%	65.63%

By testing the testing set and using the *k* value of 1, the system detected 41 images from 57 cancerous images and detected 24 images from 32 normal images. This gives a sensitivity of 71.93 % and a specificity of 75 %. By using the *k* value of 3, the system detected 40 images from 57 cancerous images and detected 16 images from 32 normal images. This gives a sensitivity of 70.18% and a specificity of 50%. By using the *k* value of 5, the system detected 39 images from 57 cancerous images and detected 21 images from 32 normal images. This gives a sensitivity of 68.42% and a specificity of 65.63%.

Comparing the results obtained from the K-NN classifier; for learning set, sensitivity and specificity results with *k*=1 is much better than results with *k*=3, and *k*=5. For testing set, sensitivity results with *k*=1 is much better than results with *k*=3, and *k*=5. But specificity results with *k*=3 is much better than results with *k*=1, and *k*=5.

Comparing the results obtained from the K-NN classifier in this study with the results obtained from other previous study [21], with k=1, the sensitivity and specificity results are much better than previous study, and with k=3, the sensitivity results are much better than previous study, but the specificity results is less better than previous study.

These results are not so much satisfactory. This returns to many reasons. The first reason comes from the great variability in the database mammograms. The cancer values and the normality values change extensively which leads to more overlapping between the normal cluster space and the cancerous cluster space. The second reason is the small number of used cases in learning the system which does not cover the entire space of each cluster.

4. CONCLUSIONS

In this study, a computer-aided classification system for mass detection in the digitized mammograms of the breast is presented. This system depends on selecting some features and using them in the classification process. The 10 features can not differentiate between normality and cancer after testing their discrimination power is proved. Also, the k-Nearest Neighbor (k-NN) classifier with *k*=1 gave the best results; a sensitivity of 71.93 % and a specificity of 75%.

REFERENCES

[1] H. Cheng, Y. M. Lui, and R. I. Freimanis, "A Novel Approach to Microcalcification Detection Using Fuzzy Logic Technique," IEEE transactions on medical imaging, vol. 17, no. 3, June 1998.
[2] P. M. Sampat, M. K. Markey, and A. C. Bovik, *Computer-aided detection and diagnosis in mammography*, Handbook of Image and Video Processing, 2nd ed., A. C. Bovik Ed. Academic Press, pp.1195-1217, 2005.
[3] F. Winsberg, M. Elkin, J. Macy, V. Bordaz, and W. Weymouth, "Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis", Radiology, vol. 89, pp. 211–5, 1967.
[4] C. Marx, A. Malich, M. Facius, U. Grebenstein, D. Sauner, S. O. R. Pfleiderer, and W. A. Kaiser, "Are unnecessary follow-up procedures induced by computer-aided diagnosis (CAD) in mammography? Comparison of Mammographic diagnosis with and without use of CAD," European Journal of Radiology, vol. 51, pp. 66–72, 2004.
[5] P. Sajda, C. Spence and J. Pearson, "Learning Contextual Relationships in Mammograms Using a Hierarchical Pyramid Neural Network," IEEE transactions on medical imaging, vol. 21, no. 3, march 2002. [6] N. Youssry, F. Abou-Chadi, and A. M. El-Sayad, "A neural network approach for mass detection in digitized mammograms," ACBME, 2002.
[7] H. Kobatake, M. Murakami, H. Takeo, and S. Nawano, "Computerized Detection of Malignant Tumors on Digital Mammograms," IEEE transactions on medical imaging, vol. 18, no. 5, may 1999.
[8] S. Yu, and L. Guan, "A CAD System for the Automatic Detection of Clustered Microcalcifications in Digitized Mammogram Films," IEEE transactions on medical imaging, vol. 19, no. 2, February 2000.
[9] B. Verma, and J. Zakos, "A Computer-Aided Diagnosis System for Digital Mammograms Based on Fuzzy-Neural and Feature Extraction Techniques," IEEE transactions on information technology in biomedicine, vol.5, no. 1, march 2001.
[10] Y. Hagihara, Y. Hagihara, and J. Wei, "Enhancement of CAD System for Breast Cancers by Improvement of Classifiers," Systems and Computers in Japan, vol. 36, no. 9, 2005.
[11] S. Furuya, T. wei, Y. hagihara, A. Shimizu, H. Kabotake, and S. Nawano, "Improvement of performance to discriminate malignant tumors from normal tissue on mammograms by feature selection and evaluation of features selection criteria," Systems and Computers in Japan, vol. 35, no. 7, 2004.
[12] D. B. Fogela, E. C. Wasson b, E. M. Boughtonc, V. W. Pm-to, " A step toward computer-assisted mammography using evolutionary programming and neural networks," Cancer Letters, vol. 119, no. 93-07, 1997.
[13] G. M. Brake, and N. Karssemeijer, "Single and multiscale detection of masses in digital mammograms," IEEE transactions on medical imaging, vol. 18, pp. 628-639, July 1999.

[14] http://marathon.csee.usf.edu/Mammography/Database.html
[15] R. C. Gonzalez, and R. E. Woods, *Digital Image processing*, Prentice-Hall. Inc., New Jersey, 2002, pp. 76-142.
[16] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification", IEEE Trans. System Man. Cybernetics, vol. SMC-3, pp. 610–621, 1973.
[17] R. F. Walker, P. Jackway, and I. D. Longstaff, "Improving co- occurrence matrix feature discrimination", in Proc. 3rd Conference on Digital Image Computing: Techniques and Applications (DICTA'95), pp. 643-648, Brisbane, Australia, 1995.
[18] C. T. Leondes, *Medical Imaging Systems Technology Analysis and Computational Methods*, World Scientific Inc., New Jersey, 2005, pp. 63-85.
[19] C. T. LE*, Introductory Biostatistics*, John Wiley & Sons Publication, April 2003.
[20] Y. M. Kadah, A. A. farag, A. M. badawy, and A. M. Youssef, "Classification algorithm for quantitative tissue characterization of diffuse liver disease from ultrasound," IEEE transactions on medical imaging, vol. 15, no. 4, August 1996.
[21] I. M. Ibrahim, A. A. Yassen, A. F. Qurany, G. E. Essam, M. A. Hefnawy, M. A. Yacoub, Y. M. Kadah, "Computer-Aided Diagnostic System for Mass Detection in Digitized Mammograms," Proc. Third Cairo Int. Biomed. ENG. Conf. (CIBEC 06), Giza, Egypt, 2006, pp. 1-4.

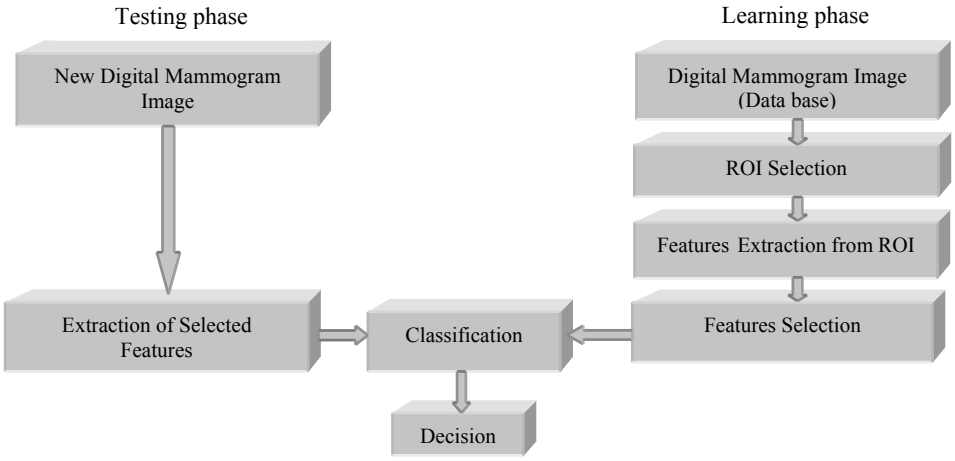

Figure (1): A schematic diagram for the computer-aided classification system.