Computer-Aided Early Detection Diagnosis System of Breast Cancer with Fuzzy Clustering Means Approach

Ayman E. Khedr
Faculty of Computers and Information, Helwan University
Email: ayman_khedr@helwan.edu.eg

Ashraf Darwish
Faculty of Science, Helwan University
Email: ashraf.darwish.eg@ieee.org,

Atef Z. Ghalwash
Faculty of Computers and Information, Helwan University
Email: atef_ghalwash@yahoo.com

Mohammed A. Osman
Faculty of Science, Helwan University
Email: m_mohameen@yahoo.com

ABSTRACT

Early detection and diagnosis of breast cancer represent the key for breast cancer control and can increase the success of treatment. This paper investigates a new Computer-Aided Diagnostic (CAD) system for the diagnosis process of benign and malignant breast tumors from digital mammography. X-ray mammograms are considered the most effective and reliable method in early detection of breast cancer. In this paper the breast tumor is segmented from medical image using Fuzzy Clustering Means (FCM) and the features for mammogram images are extracted. The results of this paper showed that these features are used to train the classifier to classify tumors. The effectiveness of this paper is examined using classification accuracy, sensitivity and specificity and the practical part of the proposed system distinguishes tumors with high accuracy.

Keywords—Breast cancer, Neural network classifier, CAD, FCM.

1. BACKGROUND AND RELATED WORK

Breast cancer is one of the major causes of death among women, especially in developed countries. The early detection of this disease can reduce the rate of death in women. CAD can help the radiologist in detecting the abnormalities in an efficient way. The mammograms can be used in the detection of breast cancer according to the World Health Organization’s International Agency for Research on Cancer.

Mammographic images are X-ray images of breast region [1]. Computer-assisted breast tumor classification which is based on the image analysis techniques provides more useful information. The conventional method for the breast tumor classification consists of three steps process. The first step involves the segmentation of breast tumor from the image. The second step is feature extraction and the third one is the classification process using a classifier. The goal of this study is to increase the diagnostic accuracy of image processing for optimum classification between benign and malignant abnormalities in digital mammograms.

Image enhancement module is a vital part as a preprocessing for any image processing technique. Image processing techniques like morphological operations and threshold techniques are applied in this study to enhance the mammogram images for the computerized detection of breast cancer.

Segmentation of medical images is an important step. This study employs a fuzzy segmentation algorithm for segmenting the mammogram images. Texture based features are extracted from the segmented images. These features are fed to the classifier for classification process. The binary classification accuracy of the developed system is measured using the Receiver Operating Characteristic (ROC) analysis with performance measures such as sensitivity, specificity and accuracy.

This paper proposes a new technique based on fuzzy algorithm and ANN to diagnosis breast cancer from digital mammograms.

Research in areas of Computer-Aided Diagnostic (CAD) systems developed within a decade. In early studies investigators outlined many approaches and limitations of Computer-Aided Diagnosis (CAD) in mammography. Winsberg, Elkin, Macy, Brodaz, and Weymouth [2] described a method that compares density between left and right breasts. An algorithm was developed by Kimme, O'Loughlin, and Sklansky [3] to detect abnormal breast regions. They calculated seven features for breast images and compared them corresponding to regions of the left and right breasts. Smith, Wagner, Guenther, and Solmon [4] introduced a measure to distinguish between malignant and benign cancer. Hand, Semmlow, Ackerman, and Alcorn [5] constructed fourteen parameters of three basic textural features, intensity, roughness and directionality to...
detect malignant areas on xeromammograms. They achieved sensitivity of 87%. Now recent studies characterized by greater use of image processing, feature analysis and artificial intelligence methods. Varela, Tahoces, Mendez, Souto and Vidal [6] applied iris filter and means of adaptive threshold to segment images and extract feature to train neural network classifier. System results sensitivity 88% and 94% at 1.02 false positives per image. Kumar and MONI [7] applied Fuzzy Clustering Means (FCM) to extract tumor from Computed Tomography (CT), textural information obtained using curvelet transform. Consequently after classification 94.3% accuracy obtained. This paper presents a new method for breast cancer early detection as describe in Fig. 1.

Fig. 1. Overview of the Proposed framework for classification
This paper is organized as following. Section I focuses on introduction and related work. Section II introduces the process of diagnosis throughout some phases of the proposed system. Section III shows results and discussion of the research. Conclusion presented in section IV.

2. THE PROPOSED DIAGNOSIS SYSTEM PHASES

This section introduces the phases of the proposed system in this paper.

2.1 Mammogram Database

Mammograms and data in this study are provided by the University of South Florid, the Digital Database for Screening Mammography (DDSM). It is available at (http://marathon.csee.usf.edu/Mammography/Database.htm). Database contains approximately 2500 studies. Each study includes two mammogram images for each breast. mammogram shows one or more clusters of tumors determined by expert radiologists. Each cluster of tumor is given in a contour surrounding the mass.

2.2 Tumor Extraction

Breast tumor is segmented as shown in Fig. 2 and Fig. 3 using Fuzzy Clustering Means (FCM) [8]. It is based on the minimization of the objective function. Partitioning by fuzzy is carried out through an iterative optimization of the membership function based on the similarity between the data and the center of a cluster. Fuzzy Clustering Means (FCM) assigns different degrees of membership to each point. The membership of a point is thus shared among various clusters. This creates the concept of a fuzzy boundaries which differs from the traditional concept of well-defined boundaries. Thus, Fuzzy Clustering Means (FCM) varies the threshold between clusters through an iterative process. As a result, the threshold is determined appropriately for every slice and the tumor region can be successfully extracted. $J_m(U,v)$ is the object function and $u_i$ is the membership function, are defined using the equations (1), (2)

$$J_m(U,v) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m (d_{ik})^2$$

(1)

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} (d_{jk})^{2(m-1)}}$$

(2)

d2k is the distance between the $k_{th}$ data (pixel value) and the center of the $i_{th}$ cluster and $V_i$ denotes the center value of the $i_{th}$ cluster, which are defined by equations (3) and (4) as follows:

$$d^2_{ik} = \|X_k - V_i\|$$

(3)

$$V_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m}$$

(4)

Where, $x_k$ is the intensity of the $k_{th}$ pixel, $n$ is the number of data (pixels), $c$ is the number of clusters, and $m$ is the exponent weight. The pixels in the background (low intensity) are included in the first cluster. The second cluster includes pixels in the tumor region (medium intensity) and the pixels in the breast region other than tumor (high intensity) are included in the third cluster. The tumor region is output for further analysis.

2.3 Feature Extraction

Feature extraction is very important stage in pattern classification. There are several types of features extracted from the mammograms. To build a system for the diagnosis process of benign and malignant breast tumors, we must get all available information existing in mammograms. But not all features can differentiate between benign and malignant tumors, so we used features that can do. In the proposed method the segmented tumor images used are of size 256×256. A set of 10 features were calculated.

1- Standard deviation: It measures how values spread out in a dataset with respect to the mean.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

(5)

2- Variance: It measures the dispersion of a set of data points around their mean value.

$$Var = s^2$$

(6)
3- Mean: It represents the average gray level in the window.
\[ x = \frac{\sum x}{n} \]  
(7)

4- Skewness: It is a measurement of the asymmetry of the data around the sample mean. If skewness is negative, the data are spread out more to the left of the mean than to the right. If skewness is positive, the data are spread out more to the right.
\[ y = \frac{E(x - \mu)^3}{\sigma^3} \]  
(8)

5- Kurtosis: It is a measurement of how outlier-prone a distribution is.
\[ K = \frac{E(x - \mu)^4}{\sigma^4} \]  
(9)

6- Entropy: A statistical measure of randomness that can be used to characterize the texture of the image.
\[ S_E = -\sum_{i=0}^{G-1} P(i, j) \log_2 \{P(i, j)\} \]  
(10)

7- Contrast: It measures the local variations in the gray-level co-occurrence matrix.
\[ CON = \sum_{i,j\in G} (i-j)^2 \cdot co(i, j) \]  
(11)

8- Energy: It provides the sum of squared elements in the gray-level co-occurrence matrix (GLCM), also known as uniformity or the angular second moment.
\[ ASM = \sum_{i,j\in G} [co(i, j)]^2 \]  
(12)

9- Homogeneity: It measures the closeness of distribution of elements in the Gray-Level Co-occurrence Matrix (GLCM) to the GLCM diagonal.
\[ HOM = \sum_{i,j} \frac{P(i, j)}{1 + |i-j|} \]  
(13)

10- Correlation: It measures the joint probability occurrence of the specified pixel pairs.
\[ COR = \sum_{i,j=0}^{G-1} P(i, j)(i - \mu_i)(j - \mu_j) / \sigma_i \sigma_j \]  
(14)

In neural network and other data mining approaches texture features values obtained need to be represented in a normalized scale. Features values are scaled (normalized) in range between 0 and 1. Feature normalization is performed using the following expression.
\[ Nf(x) = \frac{f(x) - \min(f(x))}{\max(f(x)) - \min(f(x))} \]  
(15)

Where f(x) represents the feature and \( \min(f(x)) \) and \( \max(f(x)) \) represents the minimum and maximum values corresponding to the feature f(x).

2.4 Classification
Classification is the process of learning a model that maps each attribute set to one of the predefined class labels. A classification technique (or classifier) is a systematic approach to building classification models from an input data set. In this study, neural network is the classification technique. Classification techniques are most suited for predicting or describing data sets with binary or nominal categories. Neural network employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of input data. The model generated should both fit the input data well and correctly predict the class labels of records it has never seen before. Fig. 4 shows the general approach for solving classification problems.

Fig. 4. General approach for building a classification model
Tumor classification is carried out by using a neural network classifier. Neural networks have proven themselves as the best tool for tumor classification [9]. Neural classification consists of two processes: training and testing. A training set consisting of records whose class labels are known must be provided. The training set is used to build a classification model, which is subsequently applied to test set. Testing set consists of records with unknown class labels. The accuracy of the classification depends on the efficiency of the training process.

A pattern recognition network is a feed forward network determined by activation function such as sigmoid in hidden and output layers. A neural network is a set of connected input, hidden and output units in which each connection has a weight associated with it [10]. Fig. 5 shows neural network with one hidden layer and one output layer. In this study, the input layer has 10 nodes and the hidden layer has 10 nodes and the output layer has one node. The neural network trained by adjusting the weights so as to be able to predict the correct class. The output layer produce either 1 for normal or 0 for cancer.

![Sample Neural Network with One Hidden Layer](image)

Fig. 5: A sample neural network with one hidden layer

### 2.5 Performance Evaluation

A number of different measures are commonly used to evaluate the performance of the proposed method. These measures including accuracy, sensitivity and specificity. Sensitivity is the ratio of tumors which were marked and classified as tumor. Sensitivity = True Positive/True Positive+ False Negative. Specificity is the ratio of tumors which were not marked and also not classified as tumor. Specificity = True Negative/ True Negative+ False Positive. Accuracy measures the quality of binary classification. Confusion matrix defined as in table I.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>True Positives (TP)</td>
</tr>
<tr>
<td>Negative</td>
<td>False Negatives (FN)</td>
</tr>
</tbody>
</table>

### 3. RESULTS & DISCUSSION

In order to evaluate this study, conduct experiments on a set of 100 images obtained from DDSM database. The proposed method trained with 65 images (40 malignant, 25 benign) and tested with 35 images (20 malignant, 15 benign). The confusion matrix for classification is shown in Fig. 6. Table II show the computed sensitivity, specificity and accuracy for the proposed method. The obtained classification accuracy is (97.1%) whereas sensitivity and specificity are 100% and 93.3%. The overall accuracy for benign is 100% and for cancer is 95.2%.

![Confusion Matrix for Testing Result](image)

Table 2: Performance Measures

<table>
<thead>
<tr>
<th>Tested Cases</th>
<th>specificity</th>
<th>sensitivity</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>35 case</td>
<td>93.3%</td>
<td>100%</td>
<td>97.1%</td>
</tr>
</tbody>
</table>

### 4. CONCLUSION

In this study, the main objective is to provide a Computer-Aided Diagnosis (CAD) system for the diagnosis process of benign and malignant breast tumors from digital mammography through using Fuzzy Clustering Means (FCM) and Artificial Neural Network (ANN). The method employed in this study has given better performance. The maximum accuracy rate for tumor classification is (97.1%). The performance can be increased more by increasing the number of samples. For future work, features combined with statistical moment features to improve the results in classification of mammogram images. The proposed system can be extended for medical diseases diagnosis.

### REFERENCES


