Journal of Theoretical and Applied Information Technology
30th November 2011. Vol. 33 No.2
© 2005 - 2011 JATIT & LLS. All rights reserved·
ISSN: 1992-8645www.jatit.org E-ISSN: 1817-3195

COMPARATIVE STUDY ON FEATURE EXTRACTION METHOD FOR BREAST CANCER CLASSIFICATION

[1]R.NITHYA, [2]B.SANTHI

[1]Asstt Prof., School of Computing, SASTRA University, Thanjavur, Tamilnadu, India-613402

[2] Prof., School of computing, SASTRA University, Thanjavur, Tamilnadu, India -613402

E-mail: nithya.r@mca.sastra.edu,shanthi@cse.sastra.edu

ABSTRACT

This paper presents an evaluation and comparison of the performance of three different feature extraction methods for classification of normal and abnormal patterns in mammogram. Three different feature extraction methods used here are intensity histogram, GLCM (Grey Level Co-occurrence Matrix) and intensity based features. A supervised classifier system based on neural network is used. The performance of the each feature extraction method is evaluated on Digital Database for Screening Mammography (DDSM) breast cancer database. The experimental results suggest that GLCM method outperformed the other two methods.

Keywords: *Artificial Neural Network (ANN), Breast Cancer, GLCM, Histogram, Intensity, Feature*

1. INTRODUCTION

Breast cancer is the most common disease in women in many countries. Breast image analysis can be performed using mammography, magnetic resonance, thermography and ultrasound images [1]. Mammography is highly accurate and low cost detection method. Most breast abnormality is detected as a mass on the breast through biopsy/digital mammography. Screening mammography is widely used for early detection of breast cancer. Biopsy is invasive procedure and makes patient discomfort [2]. Digital mammography is proven as efficient tool to detect breast cancer before clinical symptoms appear. Digital mammography is currently considered as standard procedure for breast cancer diagnosis [3]. Various artificial intelligence techniques such as artificial neural network and fuzzy logic are used for classification problems in the area of medical diagnosis. Image feature extraction is important step in mammogram classification. These features are extracted using image processing techniques. Several features are extracted from digital mammograms including texture feature, position feature and shape feature etc. Textures are one of the important features used for many applications. Texture features have been widely used in mammogram classification. The texture featuresare ability to distinguish between normal and abnormal pattern. Texture is an alteration and variation of surface of the image. In general, texture can be characterized as the space distribution of gray levels in a neighborhood. Texture feature have been proven to be useful in differentiating normal and abnormal pattern. Extracted texture features provide information about textural characteristics of the image. Different classifier used in biomedical imaging applications including neural network, support vector machine and fuzzy classifier. Neural network have been widely used for breast cancer diagnosis. There are two types of texture measure: first order and second order [4]. In the first order, texture measures are statistics calculated from an individual pixel and do not consider pixel neighbor relationships. In the second order, measures consider the relationship between neighbor pixels. The intensity histogram and intensity features are first order texture calculation. The GLCM is a second order texture calculation. Texture features has been extracted and used as parameter to enhance the classification result. This paper presents a comparison among three types of texture features used in mammogram classification. A texture is a method of capturing pattern in the image. These features are calculated using statistical measures such as entropy, contrast and uniformity etc. Automatic classification into normal and abnormal pattern is based on the texture features extracted from the mammograms. A computer aided diagnosis system is helping radiologists to more accurate detection of breast cancer.

The paper is structured as follows. In section 2 related works are discussed. Section 3 deals with the proposed methodology. In section 4 performance measures are explained in detail. Section 5 is the experimental results, followed by conclusions at section 6.

2. RELATED WORK

In the literature, various numbers of techniques are described to detect and classify the presence of breast cancer in digital mammograms. A lot of research has been done on the textural analysis on mammographic images. Cancer classification using GLCM features and they obtained sensitivity and specificity of more than 90%. IndraKantaMaitra et al, [1] used GLCM features to identification of abnormal masses and their study included mammograms from the MIAS database. H.S.Sheshadri et al, [7] presented a mammogram breast tissue classification using intensity histogram features.Their study included 350 mammograms from the MIAS database.H.B.Kekre et al, [8] proposed a mammogram segmentation using texture features. Their study included mammograms from the MIAS database. Hamid Soltanian-Zadeh et al, [5] presented a comparison of texture and shape features for microcalcification classification.A.MohdKhuzi et al, [3] used GLCM texture features to identification of masses in digital mammogram. Their study included 100 mammograms from the MIAS (Mammogram Image Analysis Society) database. U.RajendraAcharyaet al, [6] used a neural network to breast B.N.Prathibha et al, [9] used a kernel discriminant analysis for mammogram classification using texture features. They conducted the experiments on MIAS database.

3. METHODOLOGY

The flowchart for proposed mammogram classification is shown in figure 1.

3.1 Database

To evaluate the proposed method, Digital Database for Screening Mammography database is used for the experiment. The DDSM cancer dataset was obtained from a university of south Florida. Images are available online at the http://marathon.csee.usf.edu/Mammography/DDSM.

3.2. Feature extraction method

Feature extraction is a method of capturing visual content of an image. The objective of feature extraction process is to represent raw image in its reduced form to facilitate decision making process such as pattern classification. A variety of technique used for texture feature extraction such as intensity histogram, co-occurrence matrix and intensity based features. Texture features are extracted from the mammograms. Feature extraction step is important step to get high classification rate. A set of features are extracted in order to allow a classifier to distinguish between normal and abnormal pattern. The abnormality can be identified on the basis of textural appearance. Extracted features are used in neural classifier to train it for the recognition of particular class either normal or abnormal. The ability of the classifier to assign the unknown object to the correct class is dependent on the extracted features.

3.2.1. Intensity histogram features

Histogram is a graph showing the number of pixels in an image at each different intensity value found in that image. For an 8-bit gray scale image, there are 256 intensity values are possible. The intensity histogram features are first order statistics. The histogram is plotted from the image and from the histogram a four features are extracted that can discriminate between the two classes of mammogram. Fourfeatures such smoothness, uniformity, third moment and entropy is calculated using intensity histogram graph. The histogram graph is constructed by counting the number of pixels at each intensity value[10]. Table 1 provides equation and explanation of the four features. In this equation, G is the number of intensity levels, m is the mean, σ2 is the variance and nth moment of the mean is calculated by

$$\mu_n = \sum_{i=0}^{G-1} (z_i - m)^n \, p(z_i)$$

$m = \sum_{i=0}^{G-1} z_i \ p(z_i)\sigma = \sqrt{\mu_2(z)} = \sqrt{\sigma^2}$, where z_i be a random variable indicating intensity and let p (z_i), i=0, 1, 2, G-1[7].

3.2.2. GLCM features

The texture features are extracted using GLCM. The GLCM represents second order statistics based on neighboring pixels. The GLCM is a two dimensional array which takes into account the specific position of a pixel relative to other pixels [3]. The GLCM is a tabulation of how often different combination of pixel brightness values occur in an image. This GLCM matrices are constructed at a distance of d=1, 2, 3, 4 and for direction of data given as 0^0, 45^0, 90^0, 180^0. P (i, j) represents the probability that two pixels with a specified separation have greylevels i and j [11-13]. The texture descriptors derived from GLCM are cluster shade, contrast, energy and sum of square variance. Table 2 provides equation and explanation for four features. In this equation, G is the number of grey level used. μ is the mean value of P. σ_i,$\sigma_j$are standard deviation, where

$\mu=\sum_{i,j=0}^{G-1} i \, P(i,j)$

$$p_i(i) = \sum_{j=0}^{G-1} P(i,j) \, p_j(j) = \sum_{i=0}^{G-1} P(i,j)$$

$\mu_i=\sum_{i=0}^{G-1} i \, p_i(i)$ $\mu_j=\sum_{j=0}^{G-1} j \, p_j \, (j))$

$\sigma_i{}^2=\sum_{i=0}^{G-1}(i-\mu_i)^2 p_i(i) \, \sigma_j{}^2=\sum_{j=0}^{G-1}(j-\mu_j)^2 p_j \, (j))$

3.2.3. Intensity based features

Pixel intensities are simplest available feature useful for pattern recognition. Intensity features are first order statistics depends only on individual pixel values. The intensity and its variation inside the mammograms can be measured by features like: median, mode, standard deviation and variance. The features include median, mode, standard deviation and variance are calculated using mean dataset and explanation given in table 3. In this explanation,meandataset is calculated as the average intensity of every column in the mammogram. If size of the mammogram is m x n, then the total number of mean is n. Sample features for the three feature extraction method is shown in table 4.

3.3. Classification

The proposed method used a three layer artificial neural network and sigmoid activation function in hidden and output layers. The schematic representation of neural network with ‘n’ inputs, ’m’ hidden units and one output unit [14].The extracted features are considered as input to the neural classifier. A neural network is a set of connected input/output units in which each connection has a weight associated with it [15]. The neural network trained by adjusting the weights so as to be able to predict the correct class. The desired output was specified as 1 for normal and 0 for abnormal. The input features are normalized between 0 and 1. Theclassification process is divided into the training phase and the testing phase. In the training phase known data are given. In the testing phase, unknown data are given and the classification is performed using the classifier after training. The accuracy of the classification depends on the efficiency of the training.

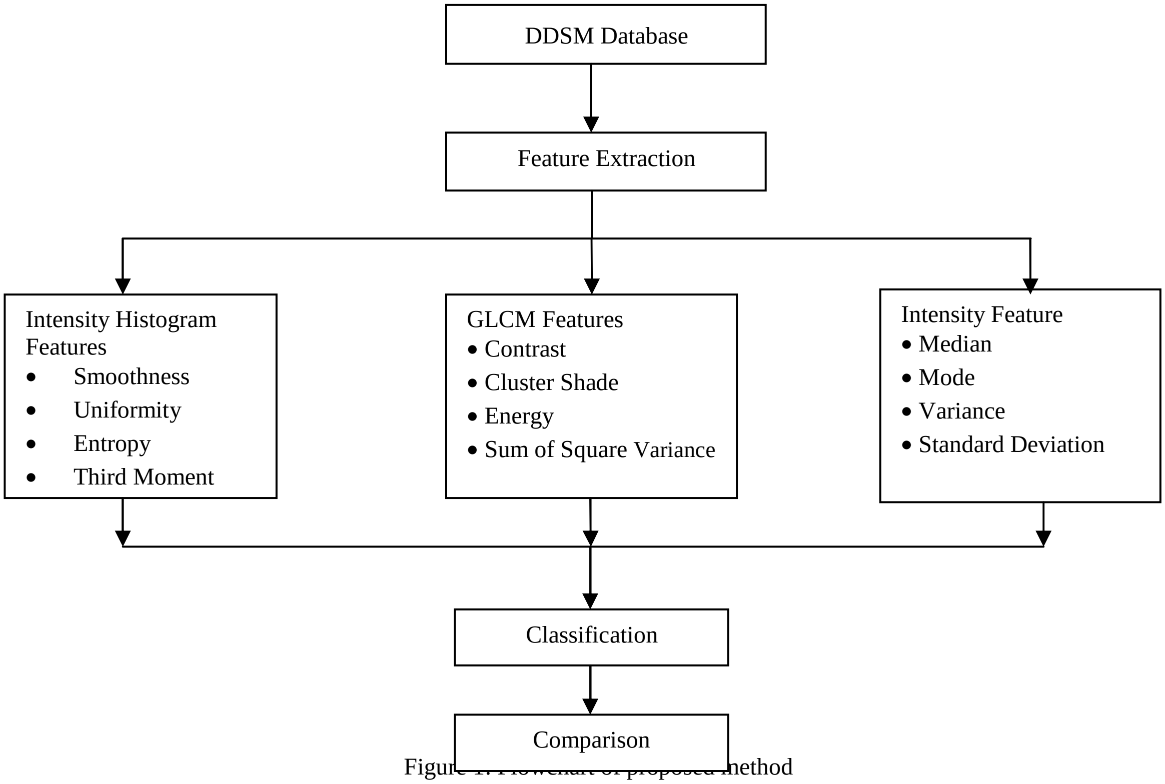


Figure 1: Flowchart of proposed method

Moment	Expression	Measure of texture
Smoothness	$R = 1 - \frac{1}{1+\sigma^2}$	Smoothness of intensity in a histogram.
Third moment	$\mu_3 = \sum_{i=0}^{G-1} (z_i - m)^3\, p(z_i)$	Skewness of a histogram.
Uniformity	$U = \sum_{i=0}^{G-1} p^2\ (z_i)$	Uniformity of intensity in a histogram.
Entropy	$e = -\sum_{i=0}^{G-1} p(z_i) \log_2 p(z_i)$	A measure of randomness.

Features	Explanation	Formula
contrast	Intensity contrast between a pixel and its neighbor	$\sum_{i,j=0}^{G-1} (i-j)^2 P(i,j)$
Cluster shade	Cluster shade is a measure of skewness of the matrix. When cluster shade is high image is not symmetry.	$\sum_{i,j=0}^{G-1} (i + j - \sigma_I - \sigma_J)^3 P(i,j)$
Energy	Energy is also known as uniformity of ASM (angular second moment) which is the sum of squared elements from the GLCM.	$\sum_{i,j=0}^{G-1} P(i,j)\ 2$
Sum of square variance	This feature puts relatively high weights on the elements that differ from the average value of P (i, j).	$\sum_{i,j=0}^{G-1} P(i,j)\ (i-\mu)^2$

Feature	Description
Median	Mean dataset are arranged in ascending order and then middle value is taken as median.
Mode	The mode of mean dataset is the value that occurs most often in mean dataset.
Variance	The variability of values in the mean dataset. $\sigma^2=\frac{1}{n-1}\sum_{i=1}^{n}(\text{mean}(i)\text{-M})$, where $\text{M}=\frac{1}{n}\sum_{i=1}^{n}\text{mean}(i)$
Standard deviation	It is the square root of the variance. $\text{SD}=\sqrt{\sigma^2}$

Table 1: Features of intensity histogram
Table 2: Features of GLCM

Table 3: Features of intensity

Feature Type	Feature	Mammogram1 (normal)	Mammogram2 (normal)	Mammogram3 (abnormal)	Mammogram4 (abnormal)
GLCM	Contrast	0.7890	0.8840	0.1060	0.3290
	Energy	0.1890	0.1510	0.3640	0.3230
	Cluster Shade	-2.9110	-2.314	9.3150	3.3720
	Sum of Square Variance	9.4439	9.4709	5.3483	6.9627
Intensity Histogram	Smoothness	0.0186	0.0139	0.0358	0.0273
	Third Moment	-0.0770	-0.2168	0.2628	0.7844
	Uniformity	0.0705	0.0302	0.1872	0.0504
	Entropy	5.5923	6.1934	4.7095	5.6207
Intensity	Median	74.1944	56.5381	19.4078	41.2549
	Mode	73	58	3	2
	Variance	143.2266	178.5121	520.3214	1230.3743
	Standard Deviation	11.968	13.361	22.811	35.077

Table 4: Feature values for a normal and abnormal mammogram

4. MEASURES FOR PERFORMANCE EVALUATION

A number of different measures are commonly used to evaluate the performance of the proposed method. These measures including classification accuracy (AC) and Mathews Correlation Co-efficient (MCC) are calculated from confusion matrix. The confusion matrix describes actual and predicted classes of the proposed method and shown in table 5.

Actual	Predicted	
	Positive	Negative
Positive	TP(True Positive)	FP(False Positive)
Negative	FN (False Negative)	TN (True Negative)

TP- correct classification of abnormal.
FP- incorrect classification of abnormal.
TN- correct classification of normal.
FN- incorrect classification of normal.

Table 5: Confusion matrix

$$\text{AC} = \frac{(\text{TP+TN})}{(\text{TP+FP+TN+FN})}$$

Accuracy assesses the effectiveness of the classifier.

$$MCC = \frac{(TPXTN-FPXFN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

MCC is used to measure the quality of binary classification. The MCC can be calculated from the confusion matrix using the formula. It returns a value from -1(inverse prediction) to +1(perfect prediction) [16].

5. EXPERIMENTAL RESULTS

The effectiveness of the three texture feature extraction methods are trained and tested using neural

classifier. The dataset used for this experiment is composed of 250 mammograms from the DDSM database which includes 125 normal and 125 abnormal,80% (200 out of 250) set of images are used for training and 20% (50 out of 250) used for testing. The effectiveness of the three different feature extraction methods will be evaluated and compared. Three experiments are conducted. In each experiment, the architecture of the neural network, training and testing samples are same. In the experiment 1, intensity histogram features are extracted and its classification done using neural classifier. In the experiment 2, GLCM features are extracted and its classification. In the experiment 3, intensity based features are extracted and its classification. The results shows that intensity histogram based neural network is giving 92% classification rate, intensity based neural network is giving 96% classification rate and GLCM based neural network is giving 98% classification rate. The confusion matrix for three different feature extraction method presented in table 6 to 8. The performance measures are calculated individually for the three different feature extraction methods are shown in table 9 andfigure 2 (a) and (b).

Actual	Predicted	
	Cancer (Positive)	Normal (Negative)
Cancer(Positive)	24(TP)	1(FP)
Normal(Negative)	3(FN)	22(TN)

Table 6: Confusion matrix for intensity histogram features

Actual	Predicted	
	Cancer (Positive)	Normal (Negative)
Cancer(Positive)	23(TP)	2(FP)
Normal(Negative)	0(FN)	25(TN)

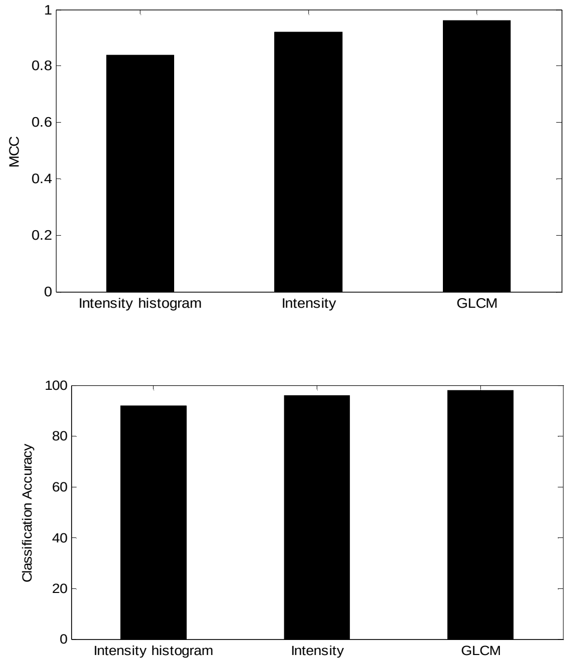
Table 7: Confusion matrix for intensity based features

Actual	Predicted	
	Cancer (Positive)	Normal (Negative)
Cancer(Positive)	24(TP)	1(FP)
Normal(Negative)	0(FN)	25(TN)

Table 8: Confusion matrix for GLCM features

Measures	Intensity histogram	GLCM	Intensity
AC (%)	92	98	96
MCC (-1 to +1)	0.84	0.96	0.92

Table 9: Evaluation results



2(a)
2(b)

Figure 2: (a) and (b) Performance measure comparison

6. CONCLUSION

This paper examined the three types of texture feature extraction method. The results are proving that GLCM features based neural network is giving higher classification rate of 98%. The GLCM gives a better performance when compared with intensity histogram and intensity features. In future, classification performance of the several classifiers will also be compared to find out the optimum classification procedure.

REFERENCES:

[1] IndraKantaMaitra, Sanjay Nag and Samir Kumar Bandyopadhyay," Identification of abnormal masses in digital mammography images", International Journal of Computer Graphics, Vol.2, No.1, 2011.

[2] Ali Keles, AyturkKeles and UgurYavuz,"Expert system based on neuro-fuzzy rules for diagnosis breast cancer", Experts Systems with Applications, pp.5719-5726, 2011.

[3] A.MohdKhuzi, R.Besar, Wan Zaki and NN.Ahmad,"Identification of masses in digital mammogram using gray level co-occurrences matrices", Biomedical Imaging and Intervention Journal, 2009.

[4] A.Karahaliou, S.Skiadopoulos, I.Boniatis, P.Sakellaropoulos, E.Likaki, G.Panayiotakis and L.Costaridou,"Texture analysis of tissue surrounding microcalcifications on mammograms for breast cancer diagnosis", The British journal of Radiology, 80, pp.648-656, 2007.

[5] Hamid Soltanian-Zadeh,FarshidRafee-Rad and SiamakPourabdollah-Nejad,"Comparison of multiwavelet,wavelet,Haralick and shape features for microcalcification in mammograms", Pattern Recognition,37,pp.1973-1986,2004.

[6] RajendraAcharya, Yang and Kaw,"Computer-based identification of breast cancer using digitized mammograms", Journal of Medical Systems, 32, pp.499-507, 2008.

[7] H.S.Sheshadri, A.Kandaswamy,"Experimental investigation on breast tissue classification based on statistical feature extraction of mammograms", Computerized Medical Imaging and Graphics, No.31, pp.46-48, 2007.

[8] H.B.Kekre and SayleeGharge, "Texture based segmentation using statistical properties for mammographic images", International Journal of Advanced Computer Science and it Applications, Vol.1, No.5, 2010.

[9] B.N.Prathibha and V.Sadasivam,"A kernel discriminant analysis in mammogram classification using with texture features in wavelet domain", International Journal on Computational Intelligence, Vol.1, Issue.1, 2010.

[10] I.Christoyianni, A.Kourtras, E.Dermatas and G.Kokkinakis,"Computer aided diagnosis of breast cancer in digitized mammograms", Computerized Medical Imaging and Graphics, 26, pp.309-319, 2002.

[11] A.M.Khuzi, R.Besar and W.M.D. Wan Zaki, "Texture features selection for masses detection in digital mammogram", 4th Kuala Lumbur International Conference on Biomedical Engineering, proceedings, pp.629-632, 2008.

[12] Viet Dzung Nguyen, DucThuan Nguyen, TienDzung Nguyen and Van Thanh Pham, "An automated method to segment and classify masses in mammograms", International Journal of Computer and Information Engineering, 52, 2009.

[13] R.Nithya and B.Santhi,"Classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer", International Journal of Computer Applications", Vol.28, No.6, 2011.

[14] J.Padmavathi, "A comparative study on breast cancer prediction using RBF and MLP", International Journal of Scientific and Engineering Research", Vol.2, Issue.1, 2011.

[15] JinchangRen, Dong Wang and Jianmin Jiang," Effective recognition of MCCs in mammograms using an improved neural classifier", Engineering Applications of Artificial Intelligence, 24, pp.638-645, 2011.

[16] AlirezaOsareh and BitaShadgar, "A computer aided diagnosis system for breast cancer", International Journal of Computer Science Issues, Vol.8, Issue-2, 2011.