Journal of Medical Imaging and Health Informatics
Vol. 4, 291–296, 2014

An Efficient Biomedical Imaging Technique for Automatic Detection of Abnormalities in Digital Mammograms

Sheeba Jenifer[1,*], S. Parasuraman[1], and Amudha Kadirvel[2]

[1]*School of Engineering, Monash University Malaysia, Bandar Sunway, 46150, Selangor, Malaysia*
[2]*Jefferey Cheah School of Medicine and Health Sciences, Monash University Malaysia Bandar Sunway, 46150, Selangor, Malaysia*

Many biomedical imaging systems provide significant contribution in aiding radiologists in interpreting digital mammograms and thus enhancing early detection of breast cancer. The motivation of this paper is to study the efficiency of existing systems and propose an efficient biomedical imaging technique which can detect abnormalities in digital mammograms. This proposed technique involves several image-processing stages combined with image preprocessing, optimum thresholding, extracting haralick features and classifying the extracted features. Preprocessing images involve 2D median filtering with Contrast-Limited Adaptive Histogram Equalization (CLAHE) techniques and optimum thresholding using Otsu's method. Feature extraction involves extracting 6 textural features using Gray Level Co-occurrence Matrices (GLCM) for four spatial orientations: 0°, 45°, 90° and 135°. Radial basis function neural (RBFN) networks were used to classify the mammograms in to two types: benign or malignant. The experimental results were tested using MIAS database of digital mammograms. The overall accuracy of the proposed technique is 99.70%. The results proved that the proposed technique outperformed other existing techniques in the aspects of sensitivity, specificity and accuracy.

Keywords: Image Analysis, Breast Cancer, Mammogram, Gray Level Co-Occurrence Matrices, Radial Basis Function Networks.

1. INTRODUCTION

Globally, about 22.9% of women suffer from breast cancer.[1] In 2008, a total of 458,503 deaths were due to breast cancer.[1] American Cancer Society stated that by 2012, nearly 40,000 women were expected to die of breast cancer in the United States alone. Survival rates can be improved if cancer is detected before the appearance of any symptoms. Mammography is the most effective technique for early detection of breast cancer. However, they are difficult images to be interpreted even by expert radiologists. Computer-aided detection (CAD) using biomedical imaging techniques can help radiologists in this difficult job of interpreting mammograms and thus can improve early detection of breast cancer. Various kinds of computer-aided methods for interpreting digital mammograms have been investigated to improve the accuracy of detection of breast cancer. Naga et al.[2] proposed a computer-aided diagnostic system that classified the mammograms into benign and malignant masses by computing gradient-based and texture-based features. Mutaz et al.[3] used GLCM to investigate digital mammograms which achieved 91.67% sensitivity and 84.17% specificity. Michael et al.[4] proposed a three-step process to detect malignant tumors in digital mammograms. Babu et al.[5] used shock filters to remove noise. Marker Controlled Watershed algorithms were used for segmentation purposes. Evolutionary-Extreme Learning Machine (E-ELM) is used to classify normal and abnormal mammograms. The system produces an accuracy of 93.33%. Following this introduction is the proposed technique of this study. Later, conclusions are made based on the results.

2. THE PROPOSED TECHNIQUE

The proposed technique is made up of four main blocks: Preprocessing block, Segmentation block, Feature Extraction block and Classification block. Figure 1 shows the overall block diagram of the proposed technique. Firstly, the system reads digitized mammograms from mini-MIAS database.[6] These images are fed to the Preprocessing block where the images are filtered and enhanced using 2D median filtering and CLAHE techniques. The preprocessed images are then fed to the Segmentation block where the regions of interest are separated from the breast regions using optimum thresholding techniques based on Otsu's method.

*Author to whom correspondence should be addressed.

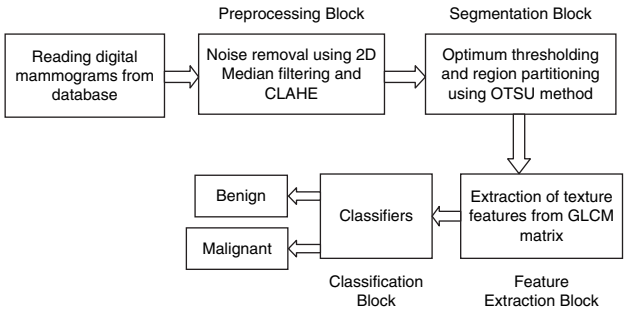

Fig. 1. Overall block diagram of the proposed technique.

The final output of the segmentation process is then fed to the Feature Extraction block. Thirteen textural features were extracted using GLCM for four spatial orientations: 0°, 45°, 90° and 135°. Selected features are then fed to the final block for Classification of mammograms in to two types: benign or malignant.

2.1. Digital Mammogram Database

The case sample used in this proposed work consists of 322 mammographic images taken from mini-MIAS database of mammograms.[6] Images are digitized at 200 micron pixel edge and 8-bit gray-level quantization with 256 different gray levels (0–255). They are padded in order to obtain all images with a size of 1024×1024 pixels. All images are in portable gray map (pgm) format.[6]

2.2. Preprocessing

Our work involves two steps for pre-processing: (i) Noise removal using 2D Median filtering (ii) Image enhancement using CLAHE. In median filtering, the neighboring pixels are ranked according to intensity and the median value becomes the new value for the central pixel.[7] The CLAHE technique is implemented for image enhancement. CLAHE limits the amplification by clipping the histogram at a predefined value before computing the probability density function. Normally, Cumulative Distribution Function (CDF) is used for distribution.[8] The probability density function of a Cumulative Distribution is

$$F(x, \beta) = 1 - e^{\wedge}(-x/(\beta)), \quad \text{for } x \geq 0; \quad \beta > 0 \qquad (1)$$

In this work, an Exponential Distribution Function (EDF)[8] is used. The probability density function of an Exponential Distribution is

$$f(x, \lambda) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0 \qquad (2)$$

where x is a random variable and $\lambda = 1/\beta$ is called rate parameter of the distribution. So as λ gets larger, the thing in the process that we are waiting for to happen, tends to happen more quickly. Figure 2 shows the output of various stages of pre-processing.

Table I. Notations.

Symbols	Definition
I	Image
$x(i, j)$	The pixel of the image I at position (i, j).
G	Number of gray levels.
t	Threshold which partitions the image into two regions, R_{low} and R_{high}.
$p(x)$	Probability of gray level x in the image I.
$\mu_{low}(t)$, $\mu_{high}(t)$	Gray level means of regions R_{low} and R_{high}
$\sigma^2_{low}(t)$, $\sigma^2_{high}(t)$	Gray level variances of regions R_{low} and R_{high}
$\sigma^2_I(t)$	In-class variance
$\sigma^2_B(t)$	Between-class variance
$\eta(t)$	Objective function
P	Gray-level co-occurrence matrix
M_x, M_y	Mean of P_x, P_y
S_x, S_y	Standard deviations of P_x, P_y
HX, HY	Entropies of P_x, P_y

2.3. Optimum Thresholding

The Otsu's method is used for optimum thresholding and region partitioning of the breast region.[9] This method automatically reduces the gray level image to a binary image. The algorithm thresholds the image into two classes of pixels: foreground and background. Then an optimum threshold is calculated so that their combined spread (intra-class variance) is minimal. Table I lists the notations used in the mathematical process.

2.3.1. Algorithm

The threshold level T_1 for the breast region is determined as[9, 10]

$$T_1 = I_{max} + 2.5\sigma_{bg} \qquad (3)$$

where I_{max} is the pixel intensity of the highest peak count in the histogram and σ_{bg} is a standard deviation of all pixel values less than I_{max}. T_1 partitions breast region R_B from background. Let $x(i, j)$ be the pixel of breast region R_B of an image I. At position (i, j), for a selected threshold 't', if $x(i, j) \leq t$, then $(i, j) \epsilon R_{low}$ (Fat Region), otherwise, $(i, j) \epsilon R_{high}$ (High-intensity region). Probability functions of regions R_{low} and R_{high} are given

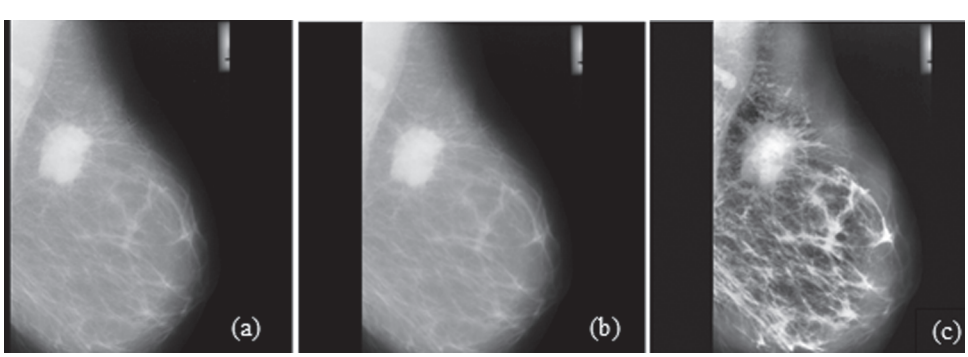

Fig. 2. Steps of preprocessing (a) An example mammographic image from MIAS database (b) Filtered image after noise removal using 2D Median filtering (c) Enhanced image using CLAHE.

as follows:

$$w_{\text{low}}(t) = \sum_{x=0}^{t} p(x) \tag{4}$$

$$w_{\text{high}}(t) = \sum_{x=t+1}^{G-1} p(x) \tag{5}$$

where $p(x)$ denotes the probability of gray level x in the image I. Gray level Means and Variances of regions R_{low} and R_{high} are[9, 10]

$$\mu_{\text{low}}(t) = \sum_{x=0}^{t} \frac{xp(x)}{w_{\text{low}}(t)} \tag{6}$$

$$\mu_{\text{high}}(t) = \sum_{x=t+1}^{G-1} \frac{xp(x)}{w_{\text{high}}(t)} \tag{7}$$

$$\sigma^2_{\text{low}}(t) = \sum_{x=0}^{t} \frac{(x-\mu_{\text{low}}(t))^2 p(x)}{w_{\text{low}}(t)} \tag{8}$$

$$\sigma^2_{\text{high}}(t) = \sum_{x=t+1}^{G-1} \frac{(x-\mu_{\text{high}}(t))^2 p(x)}{w_{\text{high}}(t)} \tag{9}$$

The in-class variance $\sigma_I^2(t)$ of these two regions is described as,

$$\sigma_I^2(t) = w_{\text{low}}(t)\sigma^2_{\text{low}}(t) + w_{\text{high}}(t)\sigma^2_{\text{high}}(t) \tag{10}$$

and the between-class variance $\sigma_B^2(t)$ of these two regions is described as

$$\sigma_B^2(t) = w_{\text{low}}(t)w_{\text{high}}(t)(\mu_{\text{low}}(t) - \mu_{\text{high}}(t))^2 \tag{11}$$

The objective function $\eta(t)$ is obtained as follows:

$$\eta(t) = \frac{\sigma_B^2(t)}{(\sigma_I^2(t) + \sigma_B^2(t))} \tag{12}$$

The optimum threshold T_{opt} that maximizes $\eta(t)$ is given as,[9, 10]

$$T_{\text{opt}} = \underset{0 \le t \le G-1}{\text{ArgMax}}(t) \tag{13}$$

We partition the breast region R_B into two regions: Fat region R_{low} and High—intensity region R_{High}, using threshold t. Then, we recursively partition R_{High} into two regions: Fatty and glandular region R_{H0} with gray level and Dense region R_{H1} using threshold T_{opt}. Figure 3 shows the various stages of region partitioning. Figure 4 shows the flow chart of the Otsu algorithm which is used for finding the optimum threshold.

2.4. Feature Extraction Using GLCM

For a given image I, of size $N \times N$, the GLCM 'P' can be defined as[11, 12]

$$P(i,j) = \sum_{p=1}^{N}\sum_{q=1}^{N} \begin{cases} 1, & \text{if } I(p,q) = i \quad \text{and} \\ & \quad I(p+d_x, q+d_y) = j \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

Here the offset (d_x, d_y) specifies the distance between the pixel of neighbor (p, q) and its neighbor. The offset d_x and d_y depends on the direction for four spatial orientations: 0°, 45°, 90° and 135°. The 13 statistical haralick texture features used in this work include Energy, Contrast, Correlation, Sum of squares: Variance, Inverse difference moment, Sum average, Sum variance, Sum entropy, Entropy, Difference variance, Difference Entropy, Information measure of correlation 1 and Information measure of correlation 2.[11, 12] These features have been used in the diagnosis of micro-calcifications in mammography.

Haralick measures are based on analyzing the second order statistics of the gray level histogram of the given Region.[11, 12]

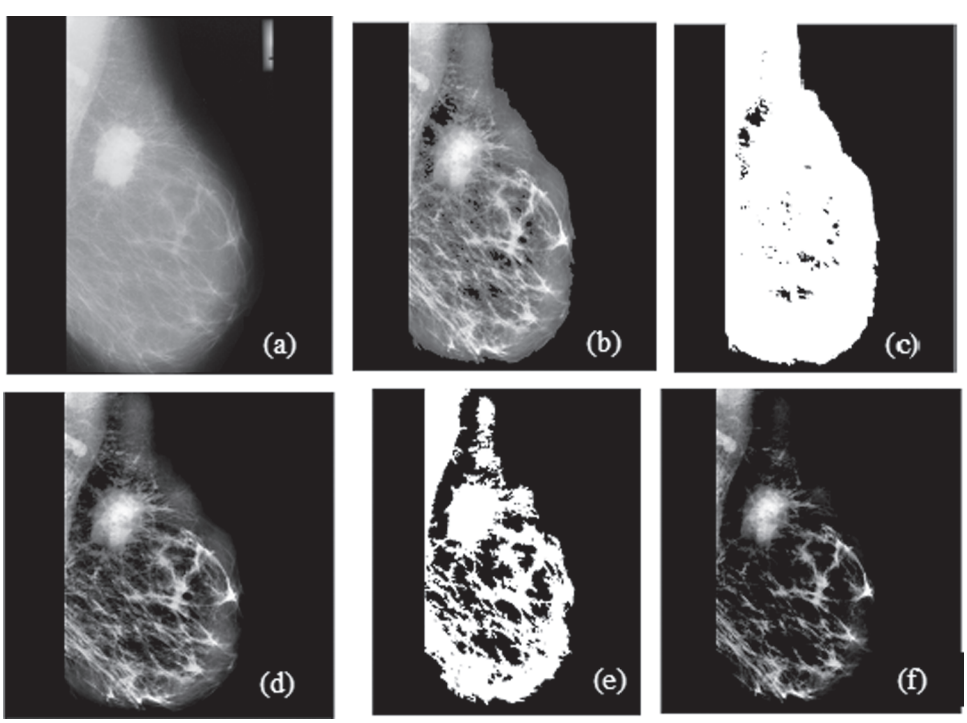

Fig. 3. Sample results of region partitioning (a) Preprocessed Mammographic image (refer Fig. 2) from the database (b) The extracted breast region R_B from the preprocessed mammographic image (c) Breast region R_B divided into two regions: R_{low}, R_{High} (d) The segmented breast region (e) Recursively partitioning R_{High} into two regions: R_{H0}, R_{H1} (f) Final segmented output.

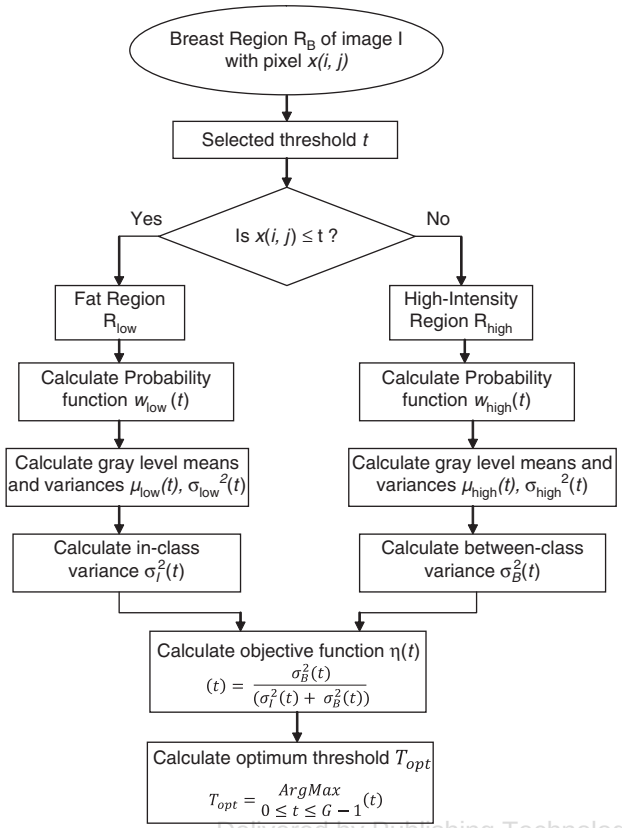

Fig. 4. Flow chart of the Otsu algorithm to calculate the optimum threshold.

This process is done by forming a set of GLCMs. The notations used for equations are given in Table I:

$$P_x(i) = \sum_{j=0}^{G-1} P(i, j) \tag{15}$$

$$P_y(i) = \sum_{j=0}^{G-1} P(i, j) \tag{16}$$

$$M_x = \sum_{i=0}^{G-1} i \sum_{j=0}^{G-1} P(i, j) = \sum_{i=0}^{G-1} iP_x(i) \tag{17}$$

$$S_x = \left(\sum_{i=0}^{G-1} P_x(i)(i - M_x)^2 \right)^{1/2} \tag{18}$$

$$M_y = \sum_{i=0}^{G-1} i \sum_{j=0}^{G-1} P(i, j) = \sum_{i=0}^{G-1} iP_y(i) \tag{19}$$

$$S_y = \left(\sum_{i=0}^{G-1} P_y(i)(i - M_y)^2 \right)^{1/2} \tag{20}$$

$$HXY = - \sum_{i, j=0}^{G-1} p(i, j) \log_2 p(i, j) \tag{21}$$

$$HXY1 = \sum_{i, j=0}^{G-1} p(i, j) \log_2 (p_x(i) p_y(i)) \tag{22}$$

$$HXY2 = - \sum_{i, j=0}^{G-1} p_x(i) p_y(i) \log_2 (p_x(i) p_y(i)) \tag{23}$$

Energy: Energy is the sum of squared elements in the GLCM[11, 12]

$$\text{Energy} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i, j)^2 \tag{24}$$

Contrast: Contrast of an aids in image definition. The image is clearer for larger contrast ratio and vice versa.[11]

$$\text{Contrast} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i, j)(i - j)^2 \tag{25}$$

Correlation: Correlation is defined as a measure of gray tone linear dependencies in the image[11]

$$\text{Correlation} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{P(i, j) X(iXj) - (M_x X M_y)}{S_x S_y} \tag{26}$$

Sum of Squares: Variance[11]

$$\text{Sum of Squares:Variance} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i, j)(i - M)^2 \tag{27}$$

Inverse Difference Moment: The inverse difference moment measures homogeneity of texture.[11]

$$\text{Inverse Difference Moment} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{P(i, j)}{1 + |i - j|^2} \tag{28}$$

Sum Average: Sum Average is the average of the normalized grey–tone image in the spatial domain.[11]

$$\text{Sum Average (sa)} = \sum_{i=0}^{2G-2} iP_{x+y} \tag{29}$$

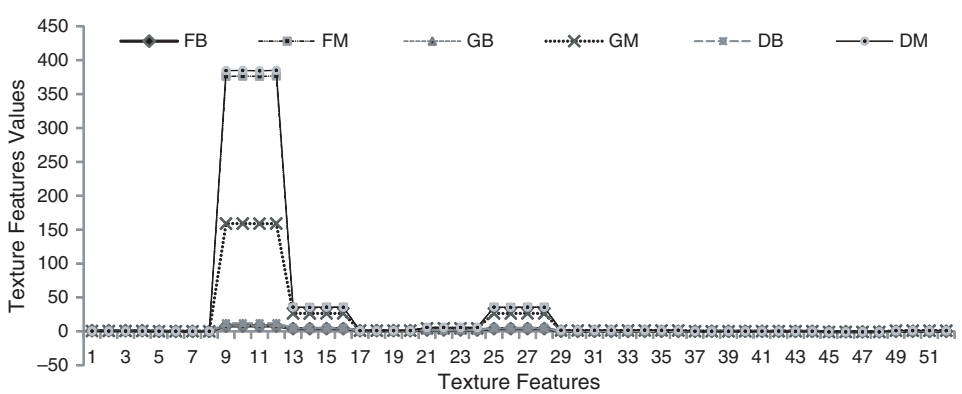

Fig. 5. Overall plot of 52 textural features (13 features in 4 directions) extracted from selected images.

Sum Variance: Sum Variance is the variance of the normalized grey—tone image in the spatial domain.[11]

$$\text{Sum Variance} = \sum_{k=2}^{2G}(k - sa)^2 p_{x+y}(k) \tag{30}$$

Sum Entropy: The Sum Entropy is a measure of randomness within an image.[11]

$$\text{Sum Entropy} = -\sum_{i=0}^{2G-2} P_{x+y}(i)\log(P_{x+y}(i)) \tag{31}$$

Entropy: Entropy measures the randomness of an image. It achieves a maximum value when most elements in GLCM are very small. A complex image produces a high entropy value.[11]

$$\text{Entropy} = -\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} P(i,j)\log(P(i,j)) \tag{32}$$

Difference Variance: The Difference Variance is an image variation in normalized co-occurrence matrix.[11]

$$\text{Difference Variance} = -\sum_{i=0}^{G-1}(i - sa)^2 p_{x+y}(i) \tag{33}$$

Difference Entropy: The Difference Entropy is also an indication of the amount of randomness in an image.[11]

$$\text{Difference Entropy} = -\sum_{i=0}^{G-1} p_{x-y}(i)\log(p_{x-y}(i)) \tag{34}$$

Information Measure of Correlation 1:[11]

$$\text{Information Measure of Correlation} = \frac{HXY - HXY1}{\max\{HX, HY\}} \tag{35}$$

Information Measure of Correlation 2:[11]

$$\text{Information Measure of Correlation 2} = \sqrt{(1 - \exp[-2(HXY2 - HXY)])} \tag{36}$$

3. RESULTS AND CONCLUSION

The 13 textural features extracted from 6 different types of digital mammograms are shown in Figure 5.

Figure 5 shows the results that are tabulated for the 6 chosen types of digital mammograms: Fatty Benign (FB), Fatty Malignant (FM), Glandular Benign (GB), Glandular Malignant (GM), Dense Benign (DB) and Dense Malignant (DM). The plot shows that out of 13, only 6 textural features clearly differentiate Benign and Malignant cases. Those 6 textural features are Correlation, Sum of Squares: Variance, Inverse Difference, Sum Average, Sum Variance and Sum Entropy. Figures 6(a)–(f) shows the plot of the 6 individual textural features calculated in 4 different directional angles 0°, 45°, 90° and 135°. Figure 6(a) plots correlation values of 6 types of digital mammograms. Correlation values of malignant images are in the range of 158.8 to 385.1 whereas the benign images are in the range of 8 to 12. Angle 135° gives highest value for all the malignant cases. Thus, the correlation of digital mammograms varied with the type of images. So, it's a good indicator to classify the different type of mammograms. Figure 6(b) plots Sum of Squares: Variance values for all the mammograms. The variance values of malignant images range between 26.3 to 35.6 whereas benign values range between 3.9 to 5.4. Again this feature proves to be a good parameter to classify benign and malignant type of images. Figure 6(e) shows the values of Sum Variance for all the types of images. The malignant cases are high (i.e.,) in the range of 26.3 to 35.6 and benign cases are low, in the range of 3.9 to 5.4. This is yet another feature which clearly shows higher values for malignant type of digital mammograms and benign type of mammograms. Overall in all the cases tested above, we can summaries that the texture feature values are higher for Fatty, Glandular and Dense malignant type and lower for Fatty, Glandular and Dense benign type of digital mammograms.

These results were compared in terms of sensitivity, specificity and accuracy with different existing methodologies used for

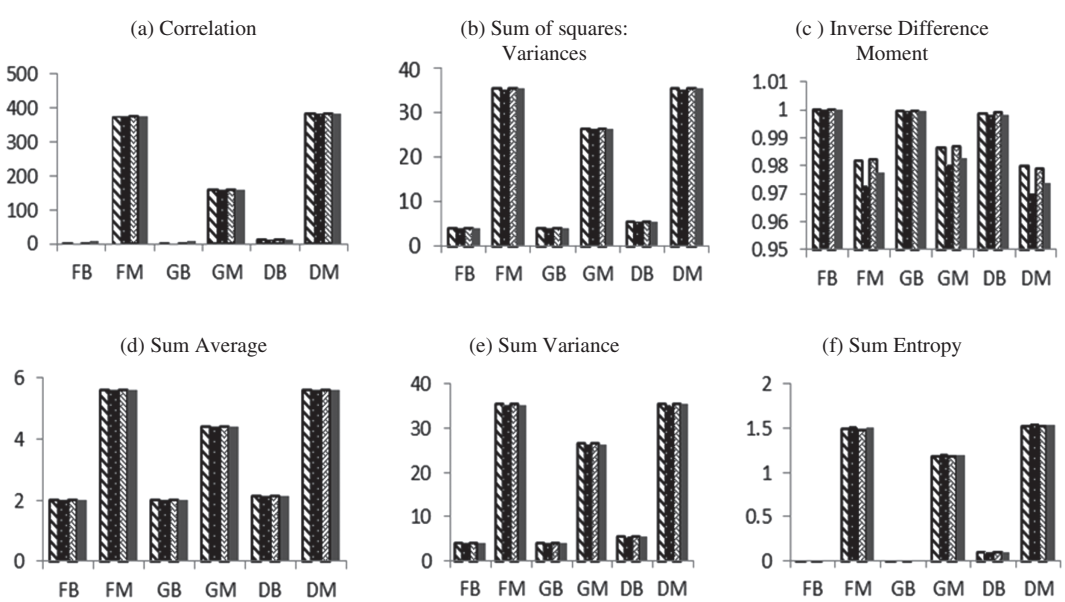

Fig. 6. (a) to (f) Plot shows digital mammograms of the type Fatty Benign (FB), Fatty Malignant (FM), Glandular Benign (GB), Glandular Malignant (GM), Dense Benign (DB) and Dense Malignant (DM) versus the 6 prominent texture features extracted out of 13 haralick feature values. The plot shows that the textural values of malignant type of breast mammograms are mostly higher than the textural values of benign type of breast mammograms. This clearly shows that textural features are very good features for classification of breast cancer using digital mammograms.

Table II. Comparison of methodologies for detection of abnormalities in digital mammogram images.

Methods	Sensitivity (%)	Specificity (%)	Accuracy (%)
Al Mutaz et al.[3]	91.67	84.17	–
Andre Nunes et al.[13]	83.24	84.14	83.94
Wener et al.[14]	80	85.68	84.62
Wan Mimi et al.[15] Method 1	74	–	–
Wan Mimi et al.[15] Method 2	92	–	–
Wan Mimi et al.[15] Method 3	99	–	–
E. Y. K. Ng et al.[16]	81.2	88.2	88.8
Proposed method:	**100**	**98**	**99.70**

detection of abnormalities in digital mammograms and given in the Table II. Sensitivity is the measure of the proportion of actual positives which are correctly identified as such. In our study, sensitivity is the percentage of malignant cases correctly classified. Specificity measures the proportion of negatives which are correctly identified as such. In our study, specificity is the percentage of benign cases correctly classified. In the field of artificial intelligence, a confusion matrix is a specific table layout that allows visualization of the performance of an algorithm. In the Confusion Matrix, for each cell in the matrix there are fields such as True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN). With respect to the number of detection, Classification using RBFN network produces only 1 error (0 FP and 1 FN). The value of our three statistical measures can be derived from the confusion matrix as follows:

$$\text{Sensitivity} = \frac{TP}{(TP+FN)}, \quad \text{Specificity} = \frac{TN}{(FP+TN)}$$

$$\text{and} \quad \text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

Several biomedical imaging techniques were investigated by means of the proposed techniques over a set of digital mammograms from MIAS database. All results were validated with the clinical data given along with the digital mammograms in the database. The results of the proposed system are compared with methodologies of some of the existing systems. Table II shows that the highest level of sensitivity 100%, specificity 98% and accuracy of 99.7% is attained by the proposed method. This indicates that the proposed methodology using RBFN network is effective in prediction of breast cancer to a large extent. This methodology has potential for further developments of real—time breast cancer diagnosis in providing a second opinion to radiologists.

References and Notes

1. World Health Organization, World Cancer Report 2008, WHO, Geneva, Switzerland **(2008)**.
2. Naga R. Mudigonda, Rangaraj M. Rangayyan, and J. E. L. Desautels, Gradient and texture analysis for the classification. *IEEE Transactions on Medical Imaging* 19, 1032 **(2000)**.
3. Al Mutaz M. Abdalla, S. Dress, and N. Zaki, Detection of masses in digital mammogram using second order statistics and artificial neural network. *International Journal of Computer Science and Information Technology* 3, 176 **(2011)**.
4. M. A. Yacoub, A. S. Mohamed, and Y. M. Kadah, A CAD system for the detection of malignant tumors in digital mammograms, *Proc. Cairo International Biomedical Engineering Conference*, Egypt **(2006)**.
5. Babu, Sukumar, and Anandan, Quantitative analysis of digitized mammograms using nonsubsampled contourlets and evolutionary extreme learning machine. *J. Med. Imaging Health Infor.* 3, 206 **(2013)**.
6. C. R. M. Boggis, I. Hutt, S. Astley, D. Betal, N. Cerneaz, D. R. Dance, S.-L. Kok, J. Parker, I. Ricketts, J. Savage, E. Stamatakis, P. Taylor, N. Karrsemeijer, and A. Clark, The mammographic image analysis society digital mammogram database. *Exerpta Medica.* 1069, 375 **(1994)**.
7. T. Arodźa, M. Kurdziel a, T. J. Popielab, E. O. D. Sevrec, and D. A. Yuenc, Detection of clustered microcalcifications in small field digital mammography. *Computer Methods and Programs in Biomedicine* 81, 56 **(2006)**.
8. J. Voss, An Introduction to Statistical Computing: A Simulation-Based Approach, John Wiley & Sons **(2013)**.
9. N. Otsu, A threshold selection method from gray-level histograms. *IEEE Trans System Man Cybernet* SMC-9, 62 **(1979)**.
10. Y. J. Lee, J. M. Park, and H. W. Park, Mammographic mass detection by adaptive thresholding and region growing. *International Journal of Imaging systems and Technology* 11, 340 **(2000)**.
11. A. Eleyan and H. Demirel, Co-occurrence matrix and its statistical features as a new approach for recognition. *Journal of Electrical Engineering and Computer Science* 19, 97 **(2011)**.
12. Haralick, Shanmugam, and Dinstein, Textural features for image classification. *IEEE Trans. Sys. Man. Cybernet* SMS-3, 610 **(1973)**.
13. A. P. Nunes, A. C. Silva, and A. C. De Paiva, Detection of masses in mammographic images using geometry, Simpson's Diversity Index and SVM. *International Journal of Signal and Imaging Systems Engineering* 3, 40 **(2010)**.
14. W. B. Sampaio, E. M. Diniz, A. C. Silva, A. C. de Paiva, and M. Gattass, Detection of masses in mammogram images using CNN, geostatistic functions and SVM. *Computers in Biology and Medicine* 41, 653 **(2011)**.
15. W. M. Diyana and R. Besar, Automated methods in clustered microcalcificationsdetection module of a CAD system. *Journal of Mechanics in Medicine and Biology* 3, 247 **(2003)**.
16. E. Y. K. Ng and E. C. Kee, Integrative computer-aided diagnostic with breast thermogram. *Journal of Mechanics in Medicine and Biology* 7, 1 **(2007)**.

Received: 6 September 2013. Accepted: 3 November 2013.