Automatic Breast Segmentation and Cancer Detection via SVM in Mammograms

Abdul Qayyum
Pakistan Institute of
Engineering and Applied Sciences
Islamabad, Pakistan
Email: muhdabdulqayyum@gmail.com

A. Basit
Pakistan Institute of
Nuclear Science and Technology,
Islamabad, Pakistan
Email: abdulbasit1975@gmail.com

***Abstract*—Abstract Automatic detection of breast cancer in mammograms is a challenging task in Computer Aided Diagnosis (CAD) techniques. This paper presents a simple methodology for breast cancer detection in digital mammograms. Proposed methodology consists of three major steps, i.e. segmentation of breast region, removal of pectoral muscle and classification of breast muscle into normal and abnormal tissues. Segmentation of breast muscle was performed by employing Otsus segmentation technique, afterwards removal of pectoral muscle is carried out by canny edge detection and straight line approximation technique. In next step, Gray Level Co-occurrence Matrices (GLCM) was created form which several features were extracted. At the end, SVM classifier was trained to classify breast region into normal and abnormal tissues. Proposed methodology was validated on Mini-MIAS database and results were compared with previously proposed techniques, which shows that proposed technique can be reliably apply for breast cancer detection.**

***Keywords*—CAD, Breast cancer, Mammograms, Mini-MIAS database, Otsus segmentation, GLCM.**

I. Introduction

Breast cancer has emerged as a leading incidence in Pakistan among many Asian countries [1]. It is reported that every 9th women in Pakistan is susceptible to suffer from breast cancer [1]. It is observed that breast cancer cases are more frequently reported in developed countries but mortality rate by breast cancer is more common in developing countries [1], [2], [3]. It was estimated that about 700,000 women in the world are diagnosed for breast cancer out of 300,000 die each year [4]. Since breast cancer can be caused by various unknown reasons, therefore complete prevention from breast cancer is not possible [5]. However, early stage detection of breast cancer can greatly reduce the mortality rate in women [6].

A mammogram screening is the most common and widely used technique for early detection of breast cancer [7]. It is considered to be the most reliable and cost effective method for detection of breast cancer [7]. In mammographic technique, a specialized low dose x-ray imaging modality is used to obtain a gray scale picture of breast region known as mammograms. Digital mammograms provide better dynamic contrast of breast tissues than screen film mammograms, and are widely utilized in CAD systems. A CAD system performs computerized mammographic analysis on digital mammograms to locate breast cancer. Now a days, many radiologist uses the results of CAD systems as a 2nd opinion before making a final decision. In general CAD techniques can be divided into two major stages as segmentation stage and computer aided cancer detection stage [8]. In segmentation step, researchers perform segmentation on specific part of mammograms to extract breast region by removing noise, labels, markers and other artifacts. After complete extraction of breast region in mammogram, the next step involves removal of pectoral muscle from breast region. In second stage of CAD techniques, several texture features are extracted from normal and abnormal breast tissues [9] and classifiers are trained via machine learning techniques in order to perform breast cancer detection in mammograms. Current study aims to introduce a simple methodology for segmentation of breast region, removal of pectoral muscle and detection of breast cancer in mammograms. Entire methodology is validated on Mini Mammographic Imaging Analysis Society (Mini-MIAS) database. The paper is organized as follows, section II explains previous techniques on automatic segmentation of breast region, removal of pectoral muscle and detection of breast cancer in mammograms. Section III discusses the proposed technique for automatic breast cancer detection. Section IV describe the results obtain with proposed technique and Section V compares the results of current methodology with previously developed methodologies.

II. Literature Review

Nagi et al. [10] proposed an automatic technique for segmentation of breast region from digital mammogram. In this technique, a mammogram was thresholded at a fixed gray scale level of 18 to obtain binary mask. Afterwards the artifacts, labels and markers in a binary mask are filtered such that resultant binary mask contains only breast region. This technique carries a flaw, because mammograms usually contain low intensity pixels (less than 18 gray scale) near skin air interface, therefore, segmentation of breast region by fix gray scale intensity level may over segment the breast region. Nanayakkara et al. [11] proposed an automatic technique for breast region segmentation in mammograms. This technique employs a modified region growing technique known as fast marching technique for segmentation of breast muscle. Mustra et al. [12] proposed a robust and automatic technique for segmentation of breast region in which mammograms are first aligned and are then thresholded by threshold values obtained by k-means clustering in which total 10 clusters were

978-1-5090-3552-6/16/$31.00

formed. Afterwards, morphological operations were performed on binary mask to extract breast region. Similarly, several techniques has been employed for detection of pectoral muscle in mammograms. The methodology of Sreedevi et al. [5] employed a gray level thresholding with canny edge detection technique for detection and removal of pectoral muscle from mammograms. In this technique, a pectoral muscle is identified using gray level thresholding and its boundary is detected by canny edge detection technique. This technique carries a flaw, because in some mammograms, boundary of pectoral muscle is not entirely visible. Therefore in such cases gray level thresholding cannot be used to remove pectoral muscle accurately. Mustra et al. [12] removed pectoral muscle by using standard edge detection technique and cubic polynomial estimation of muscle curvature. In this technique, 10 random points are selected from visible boundary of pectoral muscle which are then used for polynomial fitting of muscle boundary. Afterwards, cubic polynomial is used to estimate remaining invisible pectoral muscle boundary. A region description based methodology was proposed by Ojo et al. [13] which employs area splitting and merging techniques to remove pectoral muscle. In this methodology, a straight line approximation was performed to estimate the pectoral muscle boundary. Since the pectoral muscle boundary is not usually straight, therefore this approach either over segment or under segment the pectoral muscle. Similarly, various techniques have also been proposed for detection of breast cancer in digital mammograms. Nithya et al. [14] presented a performance comparison of three different feature extraction techniques for detection of breast cancer. A supervised neural network classifier was used to evaluate performance of Gray Level Co-occurrence Matrix (GLCM) features, intensity histogram features and intensity based features for detection of breast cancer in mammograms. Tai et al. [9] extracted GLCM and Optical Density Co-occurrence Matrix features. These features are classified by Linear Discriminant Analysis for detection of breast cancer. AlQoud et al. [15] used Gabor features and Local Binary Patterns features from normal and abnormal breast regions. A supervised Artificial Neural Network (ANN) based classification was performed to classify normal and abnormal breast tissues in mammograms.

In proposed methodology, Otsus thresholding technique is used for extraction of breast region in mammograms, subsequently combination of canny edge detection and straight line approximation was used to estimate pectoral muscle boundary and at the end, LBF were extracted from normal and abnormal breast tissues and SVM was trained to classify normal and abnormal breast tissues.

III. Methodology

The proposed methodology of our study is shown using block diagram in Fig. 1. First of all median filtering is performed to suppress noise in a mammogram. Later on Otsus technique is applied to threshold a mammogram to generate a binary mask. Generated binary mask in previous step contains breast region along with labels, and other artifacts. In order to remove labels and artifacts, connected component labeling was performed. Obtained binary mask was then used to extract breast region from original mammogram. In fourth step, location of pectoral muscle was identified. Afterwards, combination of canny edge detection and straight line approximation is used to detect pectoral muscle boundary which was then used to remove pectoral muscle. Furthermore, several features were extracted from GLCM of normal and abnormal breast tissues and at the end SVM was trained to detect breast cancer in mammogram. These steps are explained in more details in the following section.

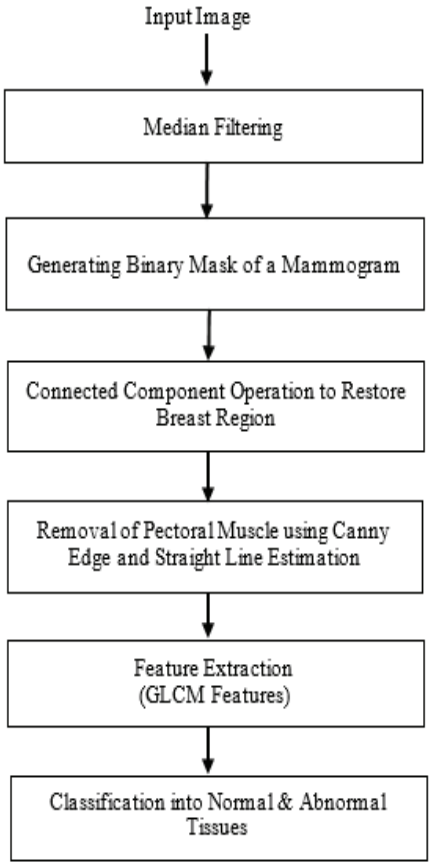

Fig. 1. Block diagram of proposed methodology

A. Median Filtering

Noise filtration is an important step in processing digital mammograms for CAD techniques. Most of mammograms usually contains low intensity noise near the skin-air interface of breast region. Sometimes, scratch artifacts are also present in mammograms. These noise and artifacts carries high frequency contents. Therefore, it is recommended to perform low pass filtering to suppress noise and artifacts in digital mammograms. Therefore, a non-linear median filter of window size $3 * 3$ was selected for noise suppression. A median filter run through every pixel of a mammogram, it takes a window of specified size from the neighborhood of each pixel say $I_{(x,y)}$ in image, sorts these pixels in ascending order and finally replaces the median value in sorted array with image pixel value $I_{(x,y)}$. The superiority of non-linear median filter over linear mean filter is that, it preserves the edges in an image and do not distribute noise content over its neighborhood pixels.

B. Extraction of Breast Region

1) Generation of Binary Mask: Mammograms contain undesirable labels, markers and other artifacts. In order to extract breast region, a mammogram was thresholded using Otsus method. Otsus technique finds a global threshold value to segment foreground pixels from its back ground. The threshold value obtained by Otsus method is said to be an optimum threshold value because it maximizes interclass variance and minimizes intra-class variance. In proposed technique, Otsus method was used to divide an entire mammogram into 16 distinct classes which has between class variance $C_1, C_2, C_3 \ldots C_{16}$ and defined as,

$$\sigma_{BW}{}^2 = \sum_{t=1}^{16} P_t (m_t - m_G)^2 \tag{1}$$

where,
$\sigma_{BW}{}^2$ is between class variance,
P_t is the cumulative probability that pixel belongs to t^{th} class,
m_t is mean pixel intensity of t^{th} class,
m_G is the mean pixel intensity of entire mammogram.
Equation (1) is optimized as,

$$t_1{}^*, t_2{}^*, \ldots, t_{16}{}^* = argmax_t \sigma_{BW}{}^2 (t_1, t_2, \ldots, t_{16})$$

Afterwards, the lowest value of threshold t_1^* obtained with Otsus method was used to binaries a mammogram. The output of this step is shown in Fig. 2.

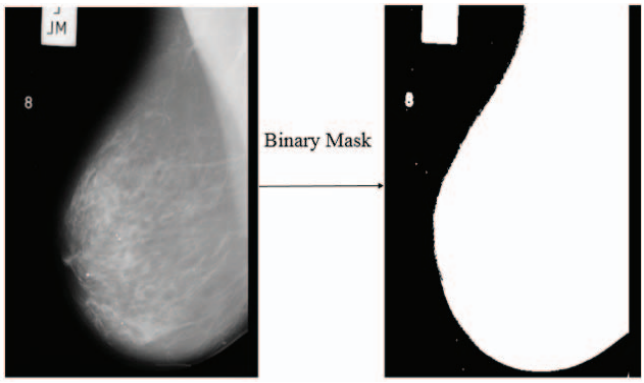

Fig. 2. Generation of binary mask by otsu's thresholding method

2) Connected Component Labeling: This section describes the restoration of breast region from generated binary mask by employing connected component labeling (CCL). Since the breast region is the largest connected component in the binary mask as shown in Fig. 2. Therefore our target is to extract the largest connected component from binary mask. For this purpose CCL was performed on the binary mask and afterwards, the largest connected component was extracted as shown in Fig. 3. Now obtained binary mask contains only a breast region. This mask is now dilated using morphological operators to smooth the boundary of breast region. Afterwards, this mask is multiplied element wise with noise suppressed mammogram to extract the breast region in a mammogram. Fig. 3, illustrates the steps for complete detection of breast region in a mammogram.

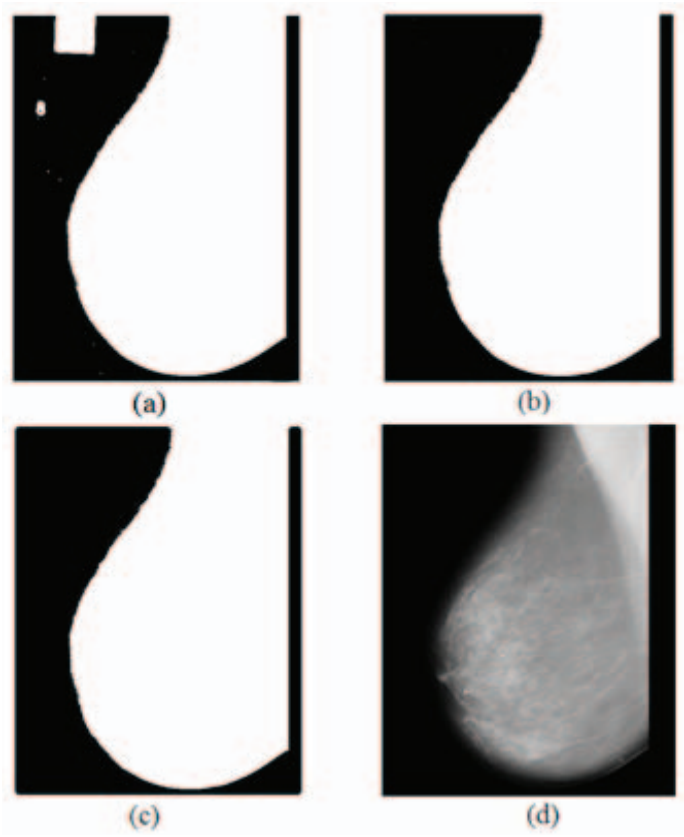

Fig. 3. (a) Binary mask contains label and marker. (b) Connected component labeling is performed to extract largest breast region. (c) Binary mask. (b) is dilated. (d) Orignal image is Multiplied with (c) to extract breast region in mammogram.

C. Removal of Pectoral Muscle

Pectoral Muscle appears as a brightest region located on either top right or top left in MLO view of a mammogram. A pectoral muscle has brightness level comparable with the cancerous tissues. Therefore removal of pectoral muscle is very important for accurate detection of breast cancer. However, removal of pectoral muscle is not an easy task because, in some mammograms, boundary of pectoral muscle is surrounded by bright fibroglandular tissues. Therefore it may be partially visible in mammograms. Our proposed method employs canny edge detection along with straight line approximation for complete detection of pectoral muscle boundary. In first step, all left aligned mammograms in database are flipped towards right, so that pectoral muscle is shifted towards top left corner. Afterwards, canny edge detection algorithm was employed to detect the boundary of pectoral muscle as shown in Fig. 4(b). In order to estimate remaining pectoral muscle boundary, a straight line is passed through the end points $A_{(x_1,y_1)}$ and $B_{(x_2,y_2)}$ respectively. Equation of line passing through points A and B is given by,

$$y - y_1 = \frac{(y_2 - y_1)}{(x_2 - x_1)} * (x - x_1) \tag{2}$$

Fig. 4(b), shows the detected pectoral muscle boundary using canny edge detection and Fig. shows 4(c), a straight

line used to estimate remaining pectoral muscle boundary. This estimated boundary is then superimposed on original mammogram 4(d) and finally pectoral muscle is removed from breast region 4(e).

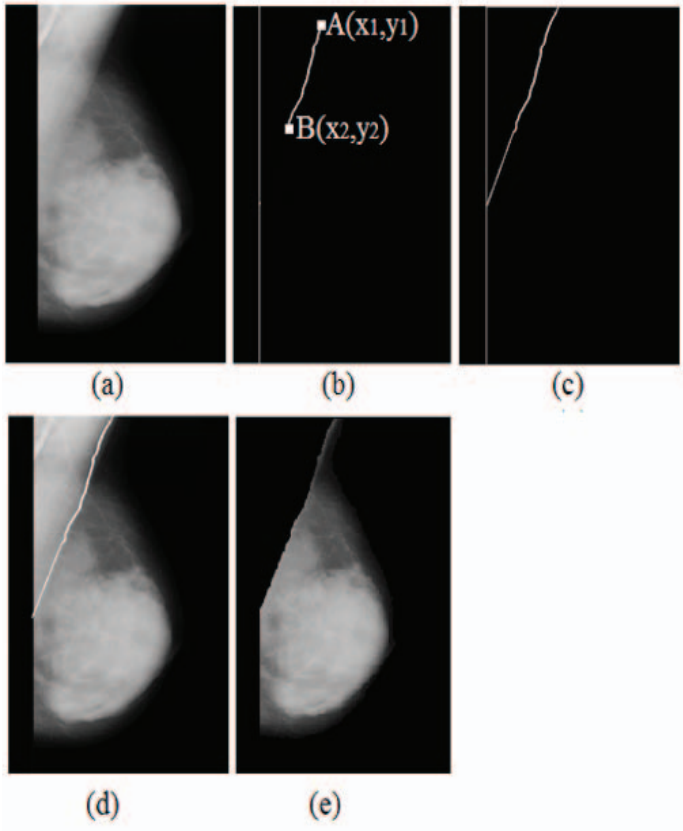

Fig. 4. (a) Original Image. (b) Detection of pectoral muscle using canny edge detection. (c) Straight line approximation. (d) Pectoral muscle boundary superimposed on original mammogram. (e) Removal of pectoral muscle.

D. Feature Extraction

Once the pectoral muscle is removed from breast muscle, some features were extracted from normal and abnormal breast tissues to express characteristics of cancerous tissues. Several feature extraction methods have been proposed as describe in literature review, but in our proposed methodology, features extraction was performed through GLCM. A GLCM describes occurrence of different combination of pixel intensities in an image. In our proposed methodology, a window (of 201*201 pixels size) was selected around cancerous tissue as shown in Fig. 5. Before creating GLCM, pixel values of image are first scale so that they become integer values between 0 and 3. Afterwards, 4 GLCM matrices were created at angle of -45, 45, -90 and 90 degrees. Four features i.e. contrast, correlation, energy and homogeneity were extracted from each GLCM matrices such that length of feature vector was 1*16. These features were extracted from normal as well as cancerous tissues.

E. SVM Based Classification

GLCM features obtained in previous steps are now classified using SVM. SVM uses linear discriminant function to classify features into two or more different classes with a linear separator in a feature space. Let feature space is m dimensional then the separator would be a hyper plane of m-1 dimensions. Linear discriminant function is given by,

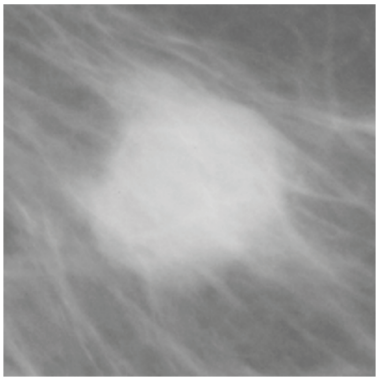

Fig. 5. Selection of rectangular window around cancer

$$f(x) = \mathbf{sgn}(\mathbf{w^T x} + b) \tag{3}$$

where,
$\mathbf{w^T x} = \sum_{j=1}^{m} w_j x_j$,
x_j is a feature vector of j^{th} example,
w is a weight vector and b is a biased value.
Here **sgn** function returns +1 labels if its argument is greater than 0, otherwise it returns -1 for argument less than 0. In SVM **w** and b are adjusted such that, following criteria is satisfied,

$$\mathbf{min} \quad \eta = \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

s.t,

$$\mathbf{w}^T\mathbf{x_i} + b \leq -1 \quad \forall \ i \ \ s.t \ \ \mathbf{y_i} = -1$$
$$\mathbf{w}^T\mathbf{x_i} + b \geq -1 \quad \forall \ i \ \ s.t \ \ \mathbf{y_i} = +1$$

Here, **y** represents output label. SVM classifier is robust and finds an optimum boundary between positive and negative class features. SVM Matlab tool box was used in current study, in which SVM was trained using linear kernel and misclassification rate was optimized on the basis of maximum AUC value obtained from receiver operating characteristic (ROC) curve of trained classifier.

IV. RESULTS AND DISCUSSION

The proposed methodology was validated on Mini-MIAS database, which is freely available. Mini-MIAS contains 322 breast images at resolution of 1024*1024 pixels. Each mammogram has spatial resolution of 50 micron in which each pixel is stored at 8 bits. All these images are in MLO view. Following section briefly explains the results obtained at each stage.

A. Results for Breast Region Segmentation

Segmentation of breast region by Otsus thresholding technique was performed on the entire database. In this segmentation step, our aims was to remove noise, markers and other artifacts from the mammograms. The results of segmentation was compared with ground truths. The results of segmentation are divided into three categories.

TABLE I. RESULTS FOR BREAST MUSCLE SEGMENTATION

Table 1		
Category	Number of Images	Percentage (%)
Good	312	96.89
Acceptable	8	2.48
Unacceptable	2	0.62

TABLE II. RESULTS OF REMOVAL OF PECTORAL MUSCLE

Table II		
Category	Number of Images	Percentage (%)
Successful	150	93
Unsuccessful	12	7

1) Good Category: This category includes the mammogram in which labels and other artifacts were removed along with successful extraction of breast region.

2) Acceptable Category: Acceptable category includes those mammograms in which labels and artifacts were removed but segmentation error was relatively large.

3) Unacceptable Category: Unacceptable category includes those mammograms in which labels or artifacts could not be removed successfully.
Table I shows the results of Breast Region Segmentation.

B. Results for Pectoral Muscle Removal

The proposed algorithm for removal of pectoral muscle was randomly applied on 161 mammograms. The results for removal of pectoral muscle are divided into two categories.

1) Successful: Mammograms from which pectoral muscle is successfully removed without over segmenting the breast tissue.

2) Unsuccessful: Mammograms from which either pectoral muscle is not removed or breast tissue is over segmented by pectoral muscle removal. Table II shows the results for pectoral muscle removal.

C. Classification of Normal and Abnormal Breast Tissues

For classification of normal and abnormal breast tissues, SVM was trained on 192 examples, in which 82 were positive and 110 were negative examples. The ROC curve of the SVM classifier is shown in the Fig. 6.
Once the SVM was trained, the performance of classifier was evaluated on 87 test examples (33 positive and 54 negative test examples). SVM classifier predicted correct labels for more than 96 % of all test examples. These classification results are shown in Table 3.

V. RESULT COMPARISON WITH OTHER RESEARCHERS

The results of proposed methodology were also compared with previously explored techniques and are enlisted below.

TABLE III. RESULTS OF REMOVAL OF PECTORAL MUSCLE

Table III				
Features	Testing-Training	Accuracy (%)	Sensitivity (%)	Specificity (%)
GLCM	87-191	96.55	96.97	96.29

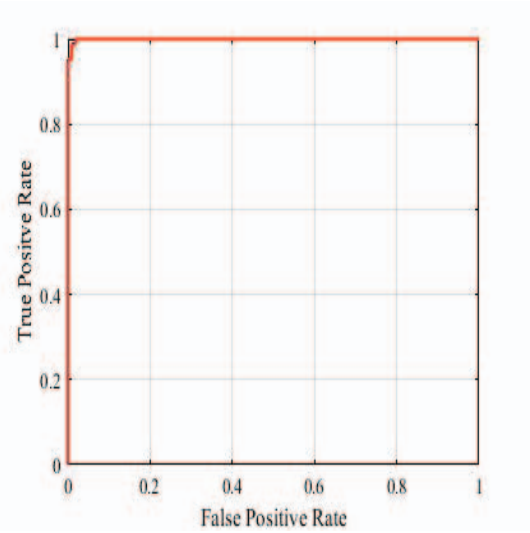

Fig. 6. ROC curve for SVM classifier

TABLE IV. RESULTS COMPARISON OF BREAST MUSCLE SEGMENTATION

Table IV		
Researchers	Acceptable (%)	Unacceptable (%)
Chandrasekhar et al. [16]	94	96.89
Raba et al. [17]	98	2
Proposed	99.38	0.62

A. Results Comparison of Breast Muscle Segmentation

Table 4 compares the results of breast muscle segmentation with previously proposed techniques, which shows that proposed methodology has significantly improved results.

B. Results Comparison of Pectoral Muscle Removal

Results of proposed algorithm are compared with previously proposed techniques in Table 5, which indicates that the results of proposed methodology are comparable with other researchers.

C. Results Comparison of Breast Cancer Detection

Results of breast tissue classification are compared with classification technique of AlQoud et al. [15] in Table 6. Although, the results of AlQoud et al. are slightly better than proposed classification technique. This might be because dataset used in proposed methodology was about 2.5 times larger than AlQoud dataset.

TABLE V. RESULTS COMPARISON OF PECTORAL MUSCLE REMOVAL

Table V		
Researchers	Acceptable (%)	Unacceptable (%)
Sreedevi et al. [5]	90.06	9.94
Ojo et al. [13]	96.27	3.73
Proposed	93	7

TABLE VI. RESULT COMPARISON OF BREAST CANCER DETECTION

Table VI		
Researchers	Training-Testing Examples (%)	Accuracy (%)
AlQoud et al. [15]	70-30	98.72
Proposed	190-87	96.55

VI. Conclusion

This paper proposed a simple methodology for segmentation of breast muscle to detect breast cancer. This methodology was validated on Mini-MIAS database. Experimental results shows that, the methodology presented in this paper can successfully be apply in CAD systems for detection of breast cancerous tissues in mammograms.

References

[1] S. Bano, M. Farhat, S, S. Arif, M. Mushtaq, M. Zafar, F. Khurshid, M. Abro, M, and A. Ahmad, "Awareness about cancer in pakistan by national acadmy of young scientists school of biological sciences university of punjab, lahore, pakistan," 2013.

[2] B. Levin and P. Boyel, "Cancer incidence and mortality worldwise," *International Agency for Research on Cancer*, 2008.

[3] A. N. Hisham and C.-H. Yip, "Overview of breast cancer in malaysian women: a problem with late diagnosis," *Asian Journal of Surgery*, vol. 27, no. 2, pp. 130–133, 2004.

[4] G. Nickel, "Diplomatic discourse, international conflict at the un. addresses and analysis," *IRAL, International Review of Applied Linguistics in Language Teaching*, vol. 37, no. 4, p. 337, 1999.

[5] S. Sreedevi and E. Sherly, "A novel approach for removal of pectoral muscles in digital mammogram," *Procedia Computer Science*, vol. 46, pp. 1724–1731, 2015.

[6] R. Takiar, D. Nadayil, and A. Nandakumar, "Projections of number of cancer cases in india (2010-2020) by cancer groups," *Asian Pac J Cancer Prev*, vol. 11, no. 4, pp. 1045–1049, 2010.

[7] M. Siddiqui, M. Anand, P. Mehrotra, R. Sarangi, and N. Mathur, "Biomonitoring of organochlorines in women with benign and malignant breast disease," *Environmental Research*, vol. 98, no. 2, pp. 250–257, 2005.

[8] R. D. Yapa and K. Harada, "Breast skin-line estimation and breast segmentation in mammograms using fast-marching method," *International Journal of Biological, Biomedical and Medical Sciences*, vol. 3, no. 1, pp. 54–62, 2008.

[9] S.-C. Tai, Z.-S. Chen, and W.-T. Tsai, "An automatic mass detection system in mammograms based on complex texture features," *IEEE journal of biomedical and health informatics*, vol. 18, no. 2, pp. 618–627, 2014.

[10] J. Nagi, S. A. Kareem, F. Nagi, and S. K. Ahmed, "Automated breast profile segmentation for roi detection using digital mammograms," in *Biomedical Engineering and Sciences (IECBES), 2010 IEEE EMBS Conference on*. IEEE, 2010, pp. 87–92.

[11] R. Nanayakkara, Y. Yapa, P. Hevawithana, and P. Wijekoon, "Automatic breast boundary segmentation of mammograms," *Int J. Soft Comput. Eng.(IJSCE)*, vol. 5, no. 1, 2015.

[12] M. Mustra and M. Grgic, "Robust automatic breast and pectoral muscle segmentation from scanned mammograms," *Signal processing*, vol. 93, no. 10, pp. 2817–2827, 2013.

[13] J. Ojo, T. Adepoju, E. Omdiora, O. Olabiyisi, and O. Bello, "Preprocessing method for extraction of pectoral muscle and removal of artefacts in mammogram," *IOSR Journal of Computer Engineering*, vol. 16, no. 3, pp. 06–09, 2014.

[14] R. Nithya and B. Santhi, "Comparative study on feature extraction method for breast cancer classification," *Journal of Theoretical and Applied Information Technology*, vol. 33, no. 2, pp. 1992–86, 2011.

[15] A. AlQoud and M. A. Jaffar, "Hybrid gabor based local binary patterns texture features for classification of breast mammograms," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 16, no. 4, p. 16, 2016.

[16] R. Chandrasekhar and Y. Attikiouzel, "Automatic breast border segmentation by background modeling and subtraction," in *5th International Workshop on Digital Mammography (IWDM),(Yaffe M. ed.), Medical Physics Publishing, Madison, USA*, 2000, pp. 560–565.

[17] D. Raba, A. Oliver, J. Martí, M. Peracaula, and J. Espunya, "Breast segmentation with pectoral muscle suppression on digital mammograms," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2005, pp. 471–478.