A Novel Approach of Classification of Mammograms using Longest Line Detection Algorithm and Decision Tree Classifier

A. M. Solanke, Manjunath, D. V. Jadhav

Abstract: Cancer is a second leading cause of mortality in a world. Approximately 9.6 million deaths are occurred due to cancer till 2018. As per the World Health Organisation, breast cancer is the major reason of mortality among women. Many lives can be saved by reliable detection and diagnosis of breast cancer in primitive stage. It is difficult to classify normal and abnormal mammograms accurately because of noise, dense breast tissues, unwanted parts such as labels and pectoral muscle. The methods based on Computer Aided Detection (CAD) can address these problems. These methods provide diagnosis assistance to radiologists and doctors for detection of cancer. Human errors can be reduced with the help of Computer Aided Detection algorithms. In this work mammogram images are preprocessed using longest line detection algorithm to remove pectoral muscle. Then texture and statistical features are extracted from preprocessed mammograms. Finally decision tree classifier is used to classify mammograms as normal and abnormal categories. The proposed methodology is applied to 322 mammograms. The performance analysis resulted into improved accuracy of 98.14%, sensitivity of 99.1% and specificity of 97.63%.

I. INTRODUCTION
Breast cancer is main reason of death in world. According to World Health Organisation out of 9.6 million cases of cancer, 2.09 cases are of breast cancer till 2018. Gender, family-history, gene mutations in BRCA1 and BRCA2, age, overweight (specifically after menopause), and consumption of alcohol are the few causes of breast cancer [1]. There are many imaging techniques for breast cancer screening such as mammography, ultrasound, Magnetic Resonance Imaging (MRI), thermography and elastography. Mammography is famous and commonly used screening technique for detection and clinical evaluation of breast cancer [2]. Masses, architectural distortion, calcifications and bilateral asymmetry are the indications of breast cancer. Masses are irregular in shape with rough boundary. Masses appear as dense tissues on mammogram, therefore distinction between masses and tissues with our eyes is a challenging task.

Specially women with dense breasts may suffer from false diagnosis. Normal breast structure gets distorted and seen as speculated lines from one source point in architectural distortion. Calcifications may appear as single spots or in clusters. Calcifications may appear as single spots or in clusters. Asymmetry of fibro-glandular tissue density between right and left breast is bilateral asymmetry [3][4]. Different CAD techniques are proposed for spotting abnormality on mammogram. Major steps of CAD system are preprocessing of mammograms, extraction of significant features and classification. These steps are implemented to focus on region of interest by removal of unnecessary information and extracting features [5]. Preprocessing is implemented by using series of morphological operations, median filter, thresholding, histogram equalization techniques, different edge detection techniques [6][7]. Region of interest is enhanced by contrast image enhancement techniques such as histogram equalization, convolutional mask, fixed neighbourhood, statistical enhancement, region based enhancement [8]-[10]. Separation of mammogram is carried out by features extraction of enhanced images. Specifically for mammogram, features extracted are related to texture properties, statistical properties, wavelet transform, Gray Level Co-occurrence Matrix (GLCM) and intensity properties. Further classifiers like Support Vector-machine (SVM), Bayesian, Artificial Neural Network (ANN), K Nearest Neighbour (KNN), Convolutional Neural Network (CNN) and Decision Tree (DT) are preferred for classification of mammograms [11]. CAD techniques are proposed to detect breast cancer accurately. Mammograms have fatty or dense background. Elshinawy proposed a technique to separate dense and fatty mammograms using Local Binary Pattern and SVM classifier [12]. Naveed proposed bagging ensemble method for mammogram classification by combining ANN, KNN, SVM and Bayesian classifiers and improved the accuracy as compared to individual classifiers using DWT features [13]. Nithya developed algorithm for classification of non cancerous and cancerous mammograms of DDSM database. Overall accuracy of 96% is obtained by using GLCM feature extraction and Neural network classifier [14].Ohmshankar formed two sets of mammogram images. One set includes normal images and another set contains images with mass and calcification as abnormal images. These sets separated using texture features and KNN classifier as normal and abnormal mammograms with 92% accuracy [15].
A Novel Approach for Classification of Mammograms using Longest Line Detection Algorithm and Decision Tree Classifier

Ranjit developed a system by extracting ROI of each abnormal image and calculated GLCM texture features of ROI. Separation of mammograms is carried out by SVM, ANN and KNN classifiers. KNN classifier achieved 95% accuracy that proved to be better than other two classifiers [16]. Anwar offered new algorithm for feature extraction and feature selection from mammograms. Discrete Cosine Transform (DCT) features are extracted and feature selection is carried out by new Weighted Sparse PCA (WSPCA) method. Classifiers such as Bayesian, DT and SVM are tried for classification of cancerous and non-cancerous mammograms of MIAS database. DT classifier performed with 97.8% accuracy better than the other two [17]. Proposed method classify mammograms with improved accuracy than current methods. Accurate diagnosis will detect breast cancer in early stage and reduce deaths. Our proposed system achieved 98.14% accuracy, sensitivity of 99.1% and specificity of 97.63% by using new longest line detection algorithm and decision tree classifier.

II. DATABASE

The Mammographic Image Analysis Society (MIAS) database is made available for researchers by an organisation of United Kingdom research groups. Total 322 mammogram images are available with left and right breasts of total one sixty one women. Even numbered file name mammogram is of left breast and odd numbered file name is of right breast of the same person. All mammogram are of size 1024 x 1024 pixels. Database consists of normal mammograms and abnormal mammograms with masses, calcifications, architectural distortion and bilateral asymmetry.

III. PROPOSED WORK

The proposed CAD system is a novel approach to analyse and separate mammograms. Total 16 features are extracted and classified which is resulted into high accuracy, sensitivity and specificity. Fig. 1 shows flow of implemented method.

![Fig. 1 Flow of proposed CAD method for classification of mammograms](image)

A. Pre-processing Technique:

Mammograms are pre-processed to get rid of labels, markers, noise and unwanted part such as pectoral muscle. Series of morphological operations are applied to filter out the unnecessary data. Pectoral muscle is removed by various techniques such as sobel filter with hough transform [18], geometrical based segmentation technique [19], operations such as thresholding, gamma correction followed by morphological operations [20]. In our proposed work pectoral muscle is masked out by longest line detection algorithm as shown in Fig.1(d). Long line detection is carried out by canny edge detector and hough transform. There are many edge detection filters like prewitt, sobel, roberts and laplacian. These filters are sensitive to noise as compared to canny edge detector. Canny edge detector uses maximal suppression method and hysteresis thresholding [21]. Algorithm for preprocessing is explained in detail as follows.

1. Mammogram images are in RGB colour format. These images are converted from RGB to grayscale as shown in equation 1.

\[
A = 0.2989 \times Rc + 0.5870 \times Gc + 0.1140 \times Bc
\]  

(1)

2. Gray scale mammograms are converted to binary image by using level of 0.04. Pixels in gray scale image with luminance greater than 0.04 are replaced by white and all other pixels with black.

3. Morphological operation is performed on image A with 5x5 disc structuring element B, as given by equation (2).

\[
g(x, y) = \begin{cases} 1 & \text{if } B \text{ fits } A \\ 0 & \text{otherwise} \end{cases}
\]

\[A \otimes B = \{z | (B)_z \cap A \neq \Phi\}
\]

4. Eight connected components having pixels less than 20000 are removed.

5. Dilation is performed on new binary image A and structuring element B, which is given by equation (3).

\[
g(x, y) = \begin{cases} 1 & \text{if } B \text{ hits } A \\ 0 & \text{otherwise} \end{cases}
\]

\[A \oplus B = \{z | (B)_z \cap A \neq \Phi\}
\]

6. Height wise image division: To concentrate on desired part, image is divided vertically in two parts.

\[A = A / 2
\]

7. Identification of left or right breast:

Mean is calculated for divided parts and decision of right or left breast mammogram is based on mean.

if AvgLeft > AvgRight
side = 'left';
Else
side = 'right';

8. Morphological opening with line structuring element with positive angle for mammogram of left breast and negative angle for right breast mammogram.

\[A \ominus B = (A \Theta B) \oplus B
\]

(5)

9. Edge detection with Canny edge (C.E.):
C.E. is optimal type of edge detector. This detector gives magnitude as well direction. Perfect detection, localization and only single response for each edge point are the criterion for C.E. edge detection. Steps for C.E. detection are as follows:

Step 1: Smoothing of image $A$ by Gaussian filter is given by equation (6)

$$S = A \ast g(x, y)$$  \hspace{1cm} (6)

$g(x, y)$ is given by equation (7)

$$g(x, y) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{x^2 + y^2}{2\sigma^2}}$$  \hspace{1cm} (7)

Standard deviation value $\sigma$ is 1.41.

Step 2: Derivative of filtered image is presented in equation (8).

$$\nabla S = \nabla(g \ast A) = (\nabla g) \ast A$$

$$\nabla S = \begin{bmatrix} g_x \\ g_y \end{bmatrix} \ast A = \begin{bmatrix} g_x \ast A \\ g_y \ast A \end{bmatrix}$$  \hspace{1cm} (8)

Step 3: Computation of magnitude $M$ and direction $D$ gradient

$$M = \sqrt{(S_x^2 + S_y^2)}$$

$$D = \tan^{-1} \frac{S_y}{S_x}$$  \hspace{1cm} (9)

Where $(S_x, S_y)$ are gradient vectors.

Step 4: Non Maximal Suppression:

Examination of each edge point leads to intensive computations. To overcome these computations non Maximal Suppression is applied. Direction of gradient is perpendicular to edge. Compare point $M(x, y)$ on edge to neighbouring pixels $M(x_1, y_1)$ and $M(x_2, y_2)$ that fall on the same gradient direction.

$M(x, y)$ is retained if it is greater than two neighbouring points otherwise suppressed as described in equation (10).

$$M = \left| \nabla S \right|(x, y) \quad \text{if} \quad \left| \nabla S \right|(x, y) > \left| \nabla S \right|(x_1, y_1) \& \left| \nabla S \right|(x, y) > \left| \nabla S \right|(x_2, y_2)$$\hspace{1cm} (10)

0 \quad \text{Otherwise}

Step 5: Hysteresis thresholding

Two threshold values are considered as low and high in hysteresis thresholding. If gradient magnitude is greater than high threshold then it is edge pixel and if lower than low threshold then non edge pixel. Values between high and low threshold are checked with eight point connected neighbours. Fig. 2 (b) shows detected edges of input mammogram.

Step 6: Pectoral muscle detection by identifying longest line among detected edges by Hough transform (H.T.): H.T. is used for edge linking to detect pectoral muscle. Hough transform maps x-y plane into c-m plane. Points in x-y plane become lines in c-m plane. Line is represented by equation (11).

$$Y = mX + C$$  \hspace{1cm} (11)

Where $m$ is a slope and $c$ is $y$ intercept. Rearranging the equation (11) in c-m plane we get equation (12).

$$c = (-x)m + y$$  \hspace{1cm} (12)

Edge points in x-y will have common intersection point in c-m plane. Line is drawn by finding local maxima in c-m plane. Line end points are determined as shown in Fig. 2 (c) and longest line is detected. Part above pectoral muscle is masked out as shown in Fig. 2 (d).

Fig. 2 Preprocessing of mammogram image

IV. EXTRACTION OF FEATURES AND CLASSIFICATION

Feature extraction is the crucial task due to availability of vast feature space. Features are divided in three categories:

1. Intensity, 2. Shape and 3. Texture features. More number of feature extraction affects accuracy as well as computation time [22]. As number of features are increasing accuracy increases up to certain features. After that accuracy starts decreasing [23]. In Proposed method total 16 features are extracted such as Mean, Entropy, Discrete Wavelet Transform (DWT) features, Principal Component Analysis (PCA) components, Kurtosis and GLCM from preprocessed images [24]. GLCM texture features such as contrast, correlation, homogeneity and energy are selected [25]. Mathematical explanation of these features along with Mean and entropy is given in Table 1. DWT separates data into various frequency components. Single level two dimensional DWT with Daubechies one filter is applied on processed image. Four different matrix coefficients are computed as Ca, Ch, Cv and Cd. Ca is approximation matrix coefficient whereas Ch, Cv and Cd are horizontal, vertical and diagonal matrix coefficients. Further principal components of Ca are extracted as features [26].
A Novel Approach for Classification of Mammograms using Longest Line Detection Algorithm and Decision Tree Classifier

Table 1. Mathematical representation of extracted features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mathematical Equation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\mu = \frac{1}{n} \left( \sum_{i=1}^{n} P_i \right)$</td>
<td>Mean is calculated for non zero pixels of preprocessed and segmented image.</td>
</tr>
<tr>
<td>Entropy</td>
<td>$H = -\sum_{i} P_i \log P_i$</td>
<td>Entropy is a one of the mathematical measure for texture information. It is a statistical measure of randomness.</td>
</tr>
<tr>
<td>Contrast</td>
<td>$\sum_{i,j=0}^{N-1} P_i j (i-j)^2$</td>
<td>It gives the contrast between two adjacent pixels $i$ and $j$. If image consists of uniform gray levels.result is zero contrast.</td>
</tr>
<tr>
<td>Homogeneity</td>
<td>$\sum_{i,j=0}^{N-1} P_i j (i-j)$</td>
<td>Homogeneity is nothing but similarity between adjacent pixels $i$ and $j$. Large values indicate the low contrast of image. Homogeneity is also known as Inverse Difference Moment (IDM).</td>
</tr>
<tr>
<td>Correlation</td>
<td>$\sum_{i,j=0}^{X-1} P_i j (i-\mu_i)(j-\mu_j) \sqrt{\sigma_i^2 \sigma_j^2}$</td>
<td>Correlation gives the value of a pixel $i$ correlated to neighbouring pixel $j$ over entire image. It ranges between 1 and -1 and gives NaN for entirely uniform gray level image.</td>
</tr>
<tr>
<td>Energy</td>
<td>$f_1 = \sum_{i} \sum_{j} \left( p(i,j) \right)^2$</td>
<td>Energy is a measure of orderliness. Order of two images will be different even though horizontal and vertical contrast is same. Energy is a square root of angular second moment. Value of energy is 1 for all identical pixel values in image.</td>
</tr>
</tbody>
</table>

DT classifier classifies mammogram images. Data is divided into training and testing. 97 false positive cases occurred for 60% training. For 70% training data false positive (FP) cases are 16. When training data is 80% resulted in 11 FP cases. Finally for 90% training data, false positive cases are reduced up to 6 only and outcome is improved accuracy [27]. Table 2 gives results of methodologies proposed by various researchers for classification of mammograms. Proposed method gives improved classification accuracy of 98.14%, specificity of 97.63% and sensitivity of 99.1% than other methods. Experimental results are presented in Table 3. Graphs of sensitivity, accuracy and specificity are shown in Fig.4.

Table 2. Comparison of methodologies and proposed work

<table>
<thead>
<tr>
<th>Author</th>
<th>Database</th>
<th>Law</th>
<th>Methodology</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Method</td>
<td>MIAS database</td>
<td>GLCM, DWT, statistical features,DT classifier</td>
<td>Accuracy: 98.14% Specificity: 97.63% Sensitivity: 99.1%</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Experimental results

<table>
<thead>
<tr>
<th>Training Percentage</th>
<th>F.P.</th>
<th>Sensitivity</th>
<th>Accuracy</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>97</td>
<td>95.62</td>
<td>69.88</td>
<td>79.89</td>
</tr>
<tr>
<td>70</td>
<td>16</td>
<td>93.04</td>
<td>95.03</td>
<td>96.14</td>
</tr>
<tr>
<td>80</td>
<td>11</td>
<td>95.61</td>
<td>96.08</td>
<td>97.12</td>
</tr>
<tr>
<td>90</td>
<td>06</td>
<td>99.1</td>
<td>98.14</td>
<td>97.63</td>
</tr>
</tbody>
</table>

V. CONCLUSION

This work presents a novel CAD method to classify mammograms into normal and abnormal categories. In this method, mammograms are preprocessed by removing noise and unnecessary part. Noise is removed by applying series of morphological operations. Further, area above pectoral muscle is removed by using longest line detection algorithm. This algorithm is a combination of canny edge detector and hough transform. The canny edge detector outperformed all other edge detectors. The features such as GLCM, DWT and statistical features are extracted from preprocessed mammograms. Finally decision tree classifier is used to classify mammograms as normal and abnormal categories. This method resulted into improved classification accuracy, specificity and sensitivity of 98.14%, 97.63% and 99.1% respectively.

ACKNOWLEDGMENT

Authors are grateful to MIAS organisation for providing database.
REFERENCES

AUTHORS PROFILE
A.M.Solankе received the M.E. degree from Government College of Engineering, Pune, India. She is presently a research scholar in Electrical & Electronics engineering, Jain (Deemed-to-be-University), Bangalore, India. She is assistant professor in MMCOE, Pune, India. The areas of her research interest are biomedical image processing, deep learning and communication engineering.

Dr. Manjunath R. has a blend of academic and industry experience over two decades. He has filed about 60 patents, authored 180 research papers in international conferences and journals and a book. He was a member of embedded systems special interest group at NASSCOM and advisor for Consumer lifestyle and Healthcare subcommittees. He represented the industry in international standardization bodies such as Wi-Fi Alliance, served as the editor for the regional profiles standard in Digital Living Network Alliance (DLNA) and as the industry liaison officer for the CE-Linux Forum. He figures in Marquis Who’s Who 2008 & ‘2000 outstanding intellectuals of the 21st century’. International Biographical Center, UK.

Dr. D.V. Jadhav received the Ph.D. degree from SGGS Institute of Engineering and Technology, Nanded, India. He authored many research papers in international conferences and journals and a book. He is Joint Director, Technical Education Amravati, India. His areas of interests include computer vision, biometrics and image processing.