Feature Extraction Values for Digital Mammograms

Arpana M.A, Prathiba Kiran

Abstract*— Currently digital mammography is the most efficient and widely used technology for early breast cancer detection. The major diagnosing elements such as masses, lesions in the digital mammograms are noisy and of low contrast. The aim of the proposal is to enhance the mammogram images by reducing the noise using median filter, image sharpening and image smoothing. The data clustering algorithm i,e Fuzzy C means clustering is used to segment the region of interest from which various statistical, gradient and geometrical features are extracted. The features extracted from the few images of the data base are used to train the neural networks for classification. The evaluated algorithm is tested on the digital mammograms from the Mammogram Image Analysis Society (MIAS) database. The experimental results show that the breast region extracted by the presented algorithm approximately follows that extracted by an expert radiologist. The detected mass is classified as normal or abnormal. Further abnormal can be classified into benign or malignant.

Index Terms*— Bio Medical Image processing, Mammograms, Breast Cancer, High Pass Spatial Filter, Fuzzy C means Clustering, Median Filtering, Gradient features

I. INTRODUCTION

Breast cancer is one among the life threatening diseases which are diagnosed in women in the world. Early detection contributes in lessening the mortality rate and improves the breast cancer prognosis. According to Globocan (WHO), for the year 2012, India recorded 70218 deaths due to breast cancer, more than any other country in the world [1-2]. Though the incidence rate in India is much lesser compared to the western countries, high mortality rate is due to the lack of instruments and techniques required for the early detection of breast cancer. About 10% of all women develop breast cancer and about 25% of all cancers diagnosed in women are breast cancers.

Mammography is currently the best method for detecting a breast cancer early, before the malignant tissue is substantial enough to feel or cause symptoms. However, the interpretation of a mammogram is often difficult and depends on the expertise and experience of the radiologist. The radiological interpretation of mammogram images is a difficult task since the appearance of even normal tissue is highly variable and complex, and signs of early disease are often small or indistinct. Suspicious findings are commonly clarified by follow-up images, ultrasound, or MRI. It has been estimated that 10−30% of cancers which could have been detected are missed.

Manuscript received on May 2014

Mrs. Arpana M.A, is currently pursuing her M.Tech Degree in Digital Electronics and Communications at AMC College of Engineering, Bangalore which is affiliated to VTU, Belgaum.

Mrs.Prathibha Kiran, was awarded with M.Tech in Biomedical Signal Processing and *Instrumentation* during 2011 from Dayanand Sagar College, which is affiliated to VTU, Belgaum.

Thus, improving both the specificity and the sensitivity of mammographic diagnoses is an important goal in improving prognoses while also reducing the number of unnecessary procedures or surgical operations.

Most of the limitations of conventional mammography can be overcome by using digital image processing. Thus, in order to improve the correct diagnosis rate of cancer, image enhancement techniques are often used to enhance the mammogram and assist radiologists in detecting it. Some of the efficient enhancement algorithm of digital mammograms based on wavelet analysis and modified mathematical morphology. Adopt wavelet-based level dependent thresholding algorithm and modified mathematical morphology algorithm[8] to increase the contrast in mammograms to ease extraction of suspicious regions known as regions of interest (ROIs) are used.Several segmentation techniques are used like the gradient vector flow snake (GVF Snake) with gradient map adjustment to obtain the accurate breast boundary from the rough breast boundary [7] and an improved multi-scale morphological gradient watershed segmentation method for automatic detection of clustered microcalcification in digitized mammograms [20]**.** The classification of the extracted features is performed using data clustering techniques namely K-means clustering, Fuzzy C-Means clustering and Subtractive clustering [6].In some methods the segmentation strategy is based on the assessment of density using multiscale wavelet transform. The density data obtained by processing with wavelet are used to train multilayer perceptron network (MLP) with one hidden layer with error back-propagation algorithm[9].The images can be classified as normal or abnormal using different classification techniques like Support vector Machine(SVM), Neural Networks.

II. METHODOLOGY

In the proposed algorithm the input mammogram images from the MINI MIAS data set are considered for evaluation in classification. The images are preprocessed using noise reduction and image enhancement techniques and the region of interest is segmented using Fuzzy c means clustering. From the ROI, various statistical, texture and gradient features of are extracted and clinical features are obtained directly from the dataset. The feature set is given as input to the neural network feed forward classifier for segregating the tumor region as benign, malignant or normal. The figure.1 shows the flow chart of the proposed system.

A. Data Set

Images from MINI MIAS (Mammographic Image Analysis Society) data base are used for the evaluation of the system. In the mini MIAS data base the original MIAS Database (digitized at 50 micron pixel edge) has been reduced to 200 micron pixel edge and clipped/padded so that every image is 1024 × 1024 pixels. The list is arranged in pairs of films, where each pair represents the left (even filename numbers)

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

and right mammograms (odd filename numbers) of a single patient. The size of *all* the images is 1024 pixels x 1024 pixels.. The database contains left and right breast images for 161 patients, and is available on a DAT-DDS tape. Its quantity consists of 322 images, which belong to three types such as Normal, benign and malignant. There are 208 normal, 63 benign and 51 malignant (abnormal) images.

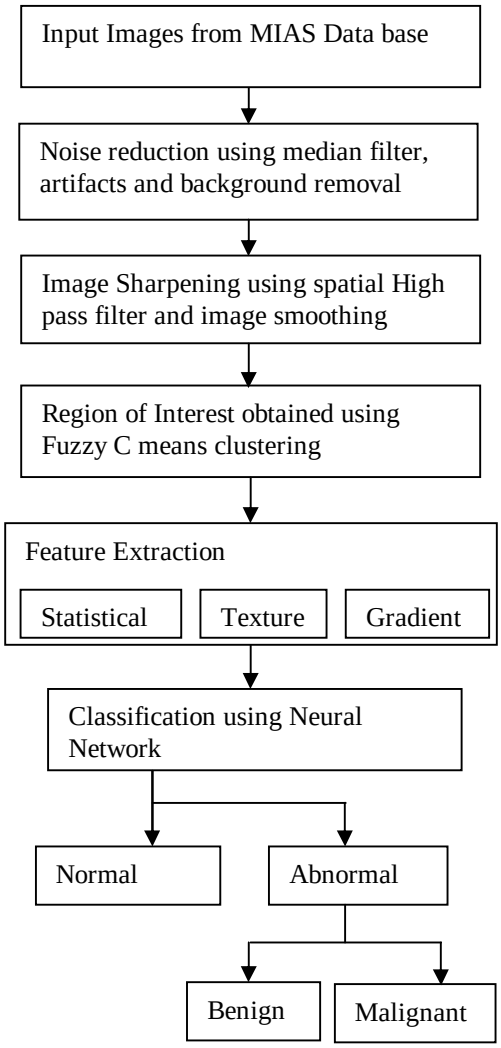


Fig.1 Flow Chart of the Proposed System

B. Preprocessing

Mammograms in their raw form contain noise, artifacts and are inconsistent to interpret. In order to produce a reliable representation of the breast anatomy we need to pre-process the mammograms. In preprocessing of the mammogram images Noise reduction and Image enhancement by sharpening and smoothing are the different steps involved.

1. Noise Reduction

The mammogram images have existing artifacts like written labels that need to be eliminated and this can done by cropping the images. The pruning of images removes nearly all background noise which is done by adding salt and pepper noise and then removing the noise using "Median Filtering".

2. Image Sharpening

Enhancing the high-frequency components of an image leads to an improvement in the visual quality. Image sharpening refers to any enhancement technique that highlights edges and fine details in an image. Image sharpening is widely used for increasing the local contrast and sharpening the images. In principle, image sharpening consists of adding to the original image a signal that is proportional to a high-pass filtered version of the original image, often referred to an unsharp masking on a one-dimensional signal. The original image is first filtered by a high-pass filter that extracts the high-frequency components, and then a scaled version of the high-pass filter output is added to the original image, thus producing a sharpened image of the original.

3. Image Smoothing

Smoothing is often used to reduce noise within an image or to produce a less pixelated image. Image smoothing is a key technology of image enhancement, which can remove noise in images. Excellent smoothing algorithm can both remove various noises and preserve details. In the proposed system Discrete Wavelet Transform Technique is used for image smoothing. These algorithms have the ability of preserving details.

C. Segmentation

Image segmentation is the process of partitioning a digital image into multiple segments. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. In this algorithm the fuzzy C means Clustering technique is used for image segmentation.

D. Feature Extraction

Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. Morphological operations like dilation and erosion are performed to extract the features from the image. Dilation adds pixels to the boundaries of objects in an image, while erosion removes pixels on object boundaries. The number of pixels added or removed from the objects in an image depends on the size and shape of the structuring element used to process the image.

E. Classification

The obtained features are applied to Feed Forward Neural Network classifier and are compared and tested with the trained ones. The output obtained from testing is classified normal or cancerous and then the cancerous images are further classified as benign or malignant and corresponding accuracy is calculated.

III. FUZZY C MEANS CLUSTERING

The fuzzy c-means (FCM) algorithm is a clustering algorithm developed by Dunn, and later on improved by Bezdek, useful when the required numbers of clusters are pre-determined; thus, the algorithm tries to put each of the data points to one of the clusters. FCM algorithm does not decide the absolute membership of a data point to a given cluster; instead, it calculates the likelihood (i.e., the degree of

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

membership) that a data point will belong to that cluster. Hence, depending on the accuracy of the clustering that is required in practice, appropriate tolerance measures can be put in place. Since the absolute membership is not calculated, FCM can be extremely fast because the number of iterations required to achieve a specific clustering exercise corresponds to the required accuracy.

Iterations

In each iteration of the FCM algorithm, the following objective function J is minimized:

$$J = \sum_{i=1}^{N}\sum_{j=1}^{C} \delta_{ij} \|x_i - c_j\|^2 \qquad (1)$$

where, N is the number of data points, C is the number of clusters required, c_j is the centre vector for cluster j , and δ_{ij} is the degree of membership for the i^{th} data point *cluster* c_j . The norm, $\|x_i - c_j\|$ measures the similarity (or closeness) of the data point x_i to the centre vector c_j of cluster j. Note that, in each iteration, the algorithm maintains a centre vector for each of the clusters. These data-points are calculated as the weighted average of the data-points, where the weights are given by the degrees of membership.

Degree of membership

For a given data point x_i, the degree of its membership to cluster j is calculated as follows:

$$\delta_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{2/(m-1)}} \qquad (2)$$

where, m is the fuzziness coefficient and the centre vector c_j is calculated as follows:

$$c_j = \frac{\sum_{i=1}^{N} \delta_{ij}^{m} * x_i}{\sum_{i=1}^{N} \delta_{ij}^{m}} \qquad (3)$$

In equation (3) above, δij is the value of the degree of membership calculated in the previous iteration. Note that at the start of the algorithm, the degree of membership for data point i to cluster j is initialized with a random value θij, $0 \le \theta ij \le 1$, such that

$$\sum_{j}^{C} \delta_{ij} = 1$$

Fuzziness coefficient

In equations (2) and (3) the fuzziness coefficient m , where $1 \le m < \infty$, measures the tolerance of the required clustering. This value determines how much the clusters can overlap with one another. The higher the value of m , the larger the overlap between clusters. In other words, the higher the fuzziness coefficient the algorithm uses, a larger number of data points will fall inside a *fuzzy* band where the degree of membership neither 0 nor 1, but somewhere in between.

Termination condition

The required accuracy of the degree of membership determines the number of iterations completed by the FCM algorithm. This measure of accuracy is calculated using the degree of membership from one iteration to the next, taking the largest of these values across all data points considering all of the clusters. If we represent the measure of accuracy between iteration k and $k+1$ with □ , we calculate its value as follows:

$$\epsilon = \Delta_i^N \Delta_j^C \left|\delta_{ij}^{k+1} - \delta_{ij}^{k}\right| \qquad (4)$$

where, δ_{ij}^{k} and δ_{ij}^{k+1} are respectively the degree of membership at iteration k and $k+1$, and the operator Δ , when supplied a vector of values, returns the largest value in that vector.

IV. STATISTICAL, TEXTURE AND GRADIENT FEATURES

Texture analysis refers to a class of mathematical procedures and models that characterize the spatial variations within image as a means of extracting information. Texture is a real construct that defines local spatial organization of spatially varying spectral values that is repeated in a region of larger spatial scale.

Statistical methods analyze the spatial distribution of gray values, by computing local features at each point in the image and deriving a set of statistics from the distributions of the local features. The reason behind this is the fact that the spatial distribution of gray values is one of the defining qualities of texture. Depending on the number of pixels defining the local feature, statistical methods can be further classified into first order (one pixel), second-order (two pixels) and higher-order (three or more pixels) statistics. The basic difference is that first-order statistics estimate properties (e.g. average and variance) of individual pixel values, ignoring the spatial interaction between image pixels, whereas second- and higher order statistics estimate properties of two or more pixel values occurring at specific locations relative to each other. Common features include moments such as mean, variance, dispersion, mean square value or average energy, entropy, skewness and kurtosis.

Mean

Mean is a measure of the average intensity of the neighbouring pixels of an image.

$$\text{Mean} = \sum_{i=0}^{L-1} Z_i * P(Z_i)$$

Variance

Variance is the average of squared deviation of all pixels from mean. The variance gives you an idea how the pixel values are spread. E.g. if mean pixel value is 50% gray, and most of the other pixels also 50% gray (small variance) or 50 black pixels and 50 white pixels (large variance)? So it gives an idea how well the mean summarizes the image (i.e. with zero variance, most of the information is captured by the mean).

$$\text{Variance} = \sum_{i=0}^{L-1} (Z_i - m)^2 * P(Z_i)$$

Skewness

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. The

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. .

$$\text{Skewness} = \sum_{i=0}^{L-1} (Z_i - m)^3 * P(Z_i)$$

Entropy

Entropy measures the purity of the clusters with respect to the given class labels. To compute the entropy of a set of clusters, the class distribution of the objects pij in each cluster j is calculated. Given this class distribution, the entropy of cluster j is calculated as:

$$\text{Entropy} = -\sum_{i=1}^{L-1} P(Z_i) * \log P(Z_i)$$

The total entropy for a set of clusters is computed as the weighted sum of entropies of all clusters.

Kurtosis

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean.

$$\text{Kurtosis} = \sum_{i=0}^{L-1} (Z_i - m)^4 * p(Z_i)$$

Contrast

Contrast is the difference in luminance and/or color that makes an object (or its representation in an image or display) distinguishable. In visual perception of the real world, contrast is determined by the difference in the color and brightness of the object and other objects within the same field of view.

$$\text{Contrast} = \sum_{i=0}^{L-1} \sqrt{(Z_i - m)^2 * P(Z_i)}$$

The geometrical features like centroid, area and gradient features like orientation are calculated.

V. RESULTS

The proposed method is tested by using the mini MIAS database of mammograms. All images are digitized at the resolution of 1024 × 1024 pixels and 8 bit accuracy (gray level). The testing images include 209 normal images, 23 images of CIRC (Circumscribed masses), 19 images of SPIC (Speculated masses),19 original images of MISC(ill-defined masses), 23 images of CALC(Calcification). The proposed algorithm was implemented in a MATLAB environment. The original image is shown as Fig. (a). and the preprocessing is done by median filtering ,image denoising and sharpening are shown in Fig. (b).Fig (c) and Fig (d) respectively. Fig (e) shows the segmented part of the ROI region.

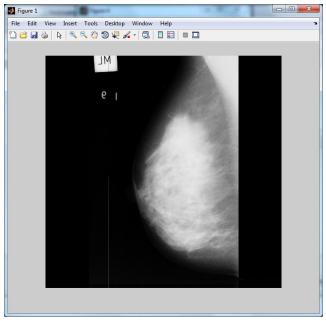


Fig (a).Original Image

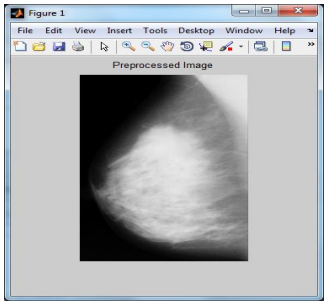
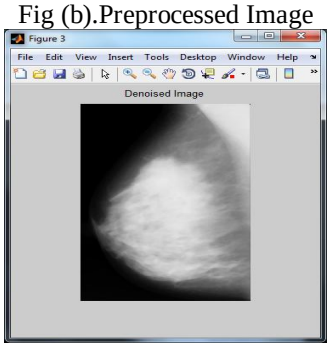
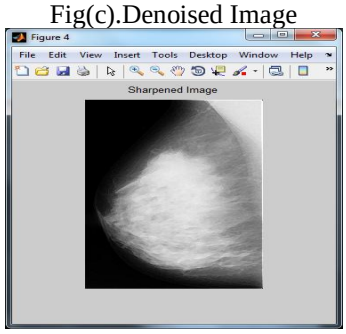


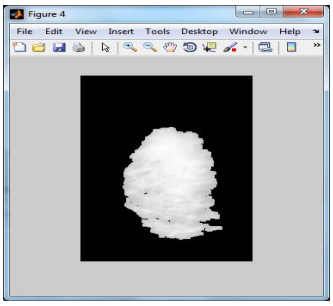
Fig (b).Preprocessed Image



Fig(c).Denoised Image



Fig(d). Sharpened Image



Fig(e).Segmented Image

REFERENCES

[1] Globocan 2012: Estimated Cancer Incidence, Mortality and Prevalence World Wide in 2012, www.globocan.ioac.fr

[2] Breast Cancer Statistics , www.breastcancer.org

[3] Michael Mavroforakis , Harris Georgiou , Nikos Dimitropoulos ,Dionisis Cavouras , Sergios Theodoridis “Significance analysis of

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

qualitative mammographic features, using linear classifiers, neural networks and support vector machines", European Journal of Radiology 54 (2005) 80–89

[4] D. Miglioretti, R. Smith-Bindman, L. Abraham, J. Brenner, E. Aiello, M. Buist, P. Carney,and J. Elmore, "Radiologist characteristics associated with interpretive performance of diagnostic mammography", Journal of the National Cancer Institute, vol. 99, no. 24, pp. 1854-63, 2007.

[5] Arianna Mencattini, Marcello Salmeri, Roberto Lojacono, Manuela Frigerio, and Federica Caselli, "Mammographic Images Enhancement and Denoising for Breast Cancer Detection Using Dyadic Wavelet Processing", IEEE Transactions on instrumentation and measurement, VOL. 57, NO. 7, JULY 2008.

[6] Mohammad Sameti, Rabab Kreidieh Ward, Jacqueline Morgan-Parkes, and Branko Palcic,"Image Feature Extraction in the Last Screening Mammograms Prior to Detection of Breast Cancer", IEEE Journal of selected topics in signal processing, VOL. 3,NO.1, FEBRUARY 2009

[7] Shyr-Shen Yu, Chung-Yen Tsai, Chen-Chung Liu, "A breast region extraction scheme for digital mammograms using gradient vector flow snake", New Trends in Information Science and Service Science (NISS), 2010 4th International Conference , IEEE2010

[8] Vijaya Kumar Gunturu, Ambalika Sharma, "Contrast enhancement of mammographic images using wavelet transform", 978-1-4244-5540-9/10©2010 IEEE

[9] Tiago T. Wirtti, Evandro O. T. Salles, "Segmentation of masses in digital mammograms", October 31, 2010. CISNE– UFES

[10] Jawad Nagi, Sameem Abdul Kareem, Farrukh Nagi, Syed Khaleel Ahmed "Automated Breast Profile Segmentation for ROI Detection Using Digital Mammograms", 978-1-4244-7600-8/10/ ©2010 IEEE

[11] Erin K. Hamilton, Seonghye Jeon, Pepa Ram´ırez Cobo, Kichun Sky Lee and Brani Vidakovic, "Diagnostic Classification of Digital Mammograms by Wavelet-Based Spectral Tools: A Comparative Study", 978-0-7695-4574-5/11© 2011 IEEE

[12] Aarthi.R, Divya.K, Komala.N, Kavitha.S, " Application of Feature Extraction and Clustering in Mammogram Classification using Support Vector Machine", IEEE 978 1 4673 0671, 2011

[13] Vishnukumar K. Patel, Prof. Syed Uvaid, Prof. A. C. Suthar,"Mammogram of Breast Cancer Detection Based using Image Enhancement Algorithm", IJETAE ISSN 2250-2459, Volume 2, Issue 8, August 2012

[14] Sheeba Jenifer Sujit, S.Parasuraman, Amudha Kadirvelu," Fuzzy Clustering In Digital Mammograms Using Gray Level Co-occurrence Matrices", IEEE 2012

[15] Sharanya Padmanabhan and Raji Sundararajan "Enhanced Accuracy of Breast Cancer Detection in Digital Mammograms using Wavelet Analysis", IEEE 2012

[16] Pradeep N, Girisha H, Sreepati B and Karibasappa K,"Feature Extraction of mammograms",International Journal of Bioinformatics Research ,ISSN: 0975–3087 & E-ISSN:

[17] G.Bharatha Sreeja, Dr. P. Rathika, Dr. D. Devaraj ,"Detection of Tumors in Digital Mammograms Using Wavelet Based Adaptive Windowing Method", I.J. Modern Education and Computer Science, 2012, 3, 57-65.

[18] Zaheeruddin , Z. A. Jaffery and Laxman Singh," Detection and Shape Feature Extraction of Breast Tumor in Mammograms", Proceedings of the World Congress on Engineering 2012 Vol II, WCE 2012, July 4 - 6, 2012, London, U.K.

[19] Karthikeyan Ganesan, U. Rajendra Acharya, Chua Kuang Chua, Lim Choo Min, K. Thomas Abraham, and Kwan-Hoong Ng, "Computer-Aided Breast Cancer Detection Using Mammograms: A Review", IEEE Reviews in biomedical engineering, VOL. 6, 2013

[20] Shrinivas D Desai ,Megha G, Avinash B, Sudhanva K, Rasiya S, Linganagouda K"Detection of Microcalcification in Digital Mammograms by Improved-MMGW Segmentation Algorithm", 2013 International Conference on Cloud & Ubiquitous Computing & Emerging Technologies

Mrs. Arpana M.A is currently pursuing her M.Tech Degree in Digital Electronics and Communications at AMC College of Engineering, Bangalore which is affiliated to VTU, Belgaum. Her area of interests includes Bio Medical Image Processing, Control Systems, and Digital Electronics.

Mrs.Prathibha Kiran was awarded with M.Tech in Biomedical Signal Processing and *Instrumentation* during 2011 from Dayanand Sagar College, which is affiliated to VTU, Belgaum. Presently working as Assistant Professor in AMC Engineering College, Bangalore in the department of Electronics & Communication Engineering. She secured 2'nd rank in M.Tech in the stream of Biomedical Signal Processing and Instrumentation from Visvesvaraya Technological University. It is to her credit that she has already contributed papers at National conference, International conference and International Journals: Her Area of research interest includes Signal processing, Wavelets, Advanced digital image processing, Neural Network and Fuzzy logic, Bio-medical Instrumentation and Wireless Communication.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.