Detection and Classification of Tumors in a Digital Mammogram

Rajashekar K.R
Mtech
Manipal Institute of technology
Manipal India

ABSTRACT
Mammography is an effective way that has demonstrated the ability to detect breast cancer at early stages with high sensitivity and specificity. Due to textural variation in image intensity , diagnosis performance varies from 60% to 80% in manual reading of mammogram .This paper demonstrates a novel approach for classifying mammograms by computer aided design using image processing and data mining techniques. This experiment consists of four stages namely preprocessing, segmentation, extraction of features and classification. In preprocessing the breast image is standardized. Then suspicious regions of cancer are acquired from mammogram by K-means clustering technique. Features are extracted from these region and are given as input to the pretrained decision tree based classifier, which in turn classifies the mammogram into normal, benign and malignant. The system has very high accuracy and has been verified with the ground truth given in the database (mini-MIAS database & DDSM). The false negative rate was as very low compared to the other existing methods.

Keywords
Breast cancer, classifier, Digital mammography, Feature extraction, Segmentation

1. INTRODUCTION
Breast cancer ranks as one of the leading cancer types in the number of new cases diagnosed and is second only to lung cancer as the most prevalent cause of cancer death in women. In 2010 the American Cancer Society reported approximately 209,060 new cases of breast cancer and 40,230 deaths due to breast cancer[12]. It is important to note that men also develop breast cancer. Approximately 390 of the reported deaths due to breast cancer in 2010 was men[12]**.** The high incidence of breast cancer in women, especially from developed countries, has increased significantly in recent years. The etiologies of this disease are not clear and neither are the reasons for the increased number of cases. Currently there are no methods to prevent breast cancer that is why early detection represents a very important factor in cancer treatment and allows reaching a high survival rate.
Mammography is the process of using low-dose amplitude-X-rays (usually around 0.7 mSv) to examine the human breast and is used as a diagnostic as well as a screening tool. In a mammogram, breast regions can be divided into background, fat tissue, breast parenchyma and tumors with increasing intensity levels. Tumors are radiolucent and appear bright in the image**.** Mammograms are considered the most reliable method in early detection of cancer[8]. Normally mammogram readings are performed by radiologists. Large number of mammograms generated by screening of population must be diagnosed by relatively few radiologists. It is difficult to provide accurate diagnosis due to variety of factors such as poor quality of image, benign appearance of lesions, eye fatigue factor, and difficulty due to bright zone of the objects on mammogram. So that the performance of the radiologists varies from 65% to 85% [12]. Due to the above mentioned regions a variety of computer assisted detection techniques have been proposed. In order to improve the accuracy of interpretation CAD involves two major process computer aided detection (CADe) and Computer Aided Diagnosis (CADi). Developing CAD algorithm using extracted textures from breast profile region would reduce number of unnecessary biopsies in patients with benign disease and thus avoid physical and mental suffering of patients. Thus CAD acts as a second reader and assists radiologist for accurate and efficient detection of cancer cells in the earlier stages. Thus the combination of CAD scheme and expert's knowledge will greatly improve the detection accuracy.
In this experimental study, digital mammogram images that were provided from online mammogram database (MIAS database]. Each image is segmented and significant features are computed from them.

2. PROBLEM DEFNITION
Digital mammograms are among the most difficult medical images to be read due to their low contrast and differences in the types of tissues. Important visual clues of breast cancer include preliminary signs of masses and calcification clusters. Unfortunately, at the early stages of breast cancer, these signs are very subtle and varied in appearance, making diagnosis difficult, challenging even for specialists for classifying a benign and malignant tumors. This is the main reason for the development of classification systems using image processing and data mining techniques to assist specialists in medical institutions. This paper attempts to classifies the digital mammograms into three categories: normal, benign and malignant. The normal ones those characterizing a healthy patient, the benign ones represent mammograms showing a tumor, but that tumor is not formed by cancerous cells, and the malignant ones are those mammograms taken from patients with cancerous tumors.

2. SYSTEM DEVELOPMENT METHODOLOG
This include mainly six stages as shown in figure 1[4].

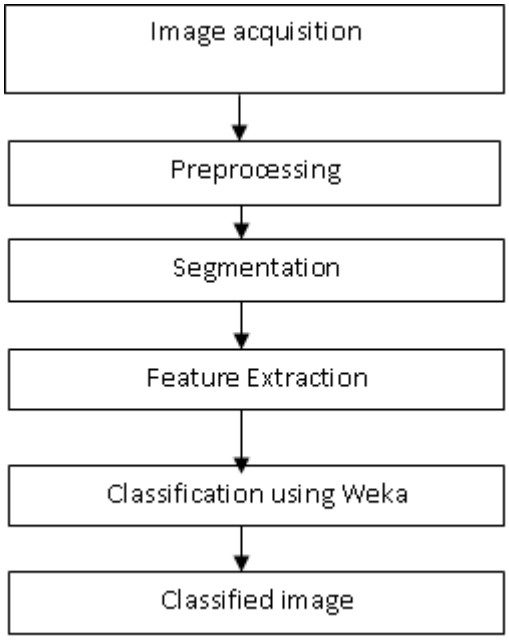

Fig. 1: System development methodology

2.1 Image Acquisition

Obtaining the mammogram in the digital format is Image Acquisition. The images required for training the classifier and testing the classifier are stored in the data base. We use the digital database for screening mammography (DDSM) which is a publicly available database of digitized screen filmed mammograms. It contains 2620cases. In image acquisition no actual scanning is involved, but the images are accessed from pre-defined data base [10]. The images used for training and testing are in **".pgm**" format. .

2.2 Preprocessing

Almost 50% of the mammogram image contain background .Two steps involved in preprocessing is noise removal and contrast enhancement[2].Existing artifacts like written labels are removed from the image by image cropping operations. The pruning of images removes nearly all the background noise using median filters[1].Image enhancement helps in qualitative improvement of the image with respect to a specific application. Enhancement can be done either in the spatial domain or in the frequency domain[2]. Here we work with the spatial domain and directly deal with the image plane itself. In order to diminish the effect of over-brightness or over-darkness in images, and at the same time accentuate the image features, we applied the Max-Min Equalization method, which is a widely used technique. In this type of image equalization technique the maximum pixel value of the image is made 255. The minimum pixel value is made zero. The intermediate pixel values are varied according to equation given below,

$$D_{pixel} = [(D_{in} - D_{min}) / (D_{max} - D_{min})] * 255$$

Enhanced image is shown in figure 2.

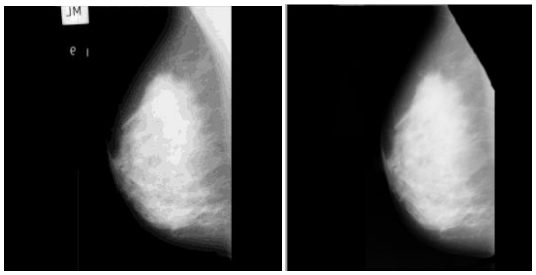

Fig. 2: Image before and after enhancement

2.3 Segmentation

The aim of the segmentation process is to extract the abnormal regions or regions with high probability of abnormality that is the Region of Interest (ROI). The result of image segmentation is a set of segments that collectively cover the entire image, or set of contours extracted from the image[2]. Each of the pixels in the region are similar with respect to some characteristic or computed properly, such as color, intensity, or texture are significantly different with respect to the same characteristic. Clustering algorithms can be applied to solve the segmentation problem. They consist in choosing an initial pixel or region that belongs to one object of interest, followed by an iterative process of neighborhood's analysis, deciding if whether each neighboring pixel belongs or not to the same object. This paper proposes a K-means segmentation method for detection of masses in digital mammograms[10].

Consider a histogram of a digital image with gray levels in the range [0, L-1].It is a discrete function $h(r_k) = n_k$, where r_k is the k^{th} gray level and n_k is the number of pixels in the image having gray level r_k. A histogram is a function m_i that counts the number of observations that fall into each of the disjoint categories (known as bins). Let n be the total number of observations and k be the total number of bins, the histogram m_i meets the following conditions:

$$n = \sum_{i=1}^{k} m_i .$$

Given a set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets $(k \leq n)$ $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Where $\boldsymbol{\mu}_i$ is the mean of points in S_i. Given an initial set of k means $m_1 \ldots m_k$, the algorithm proceeds by alternating between two steps: Assign each observation to the cluster with the closest mean

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\| \leq \|x_p - m_j^{(t)}\|$$

Calculate the new means to be the centroid of the observations in the cluster.

$$\mathbf{m}_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{\mathbf{x}_j \in S_i^{(t)}} \mathbf{x}_j$$

The Random Partition method first randomly assigns a cluster to each observation and then proceeds to compute the initial means which is the centroid of the cluster's randomly assigned points. The results of the segmentation of Mammogram image is shown below:

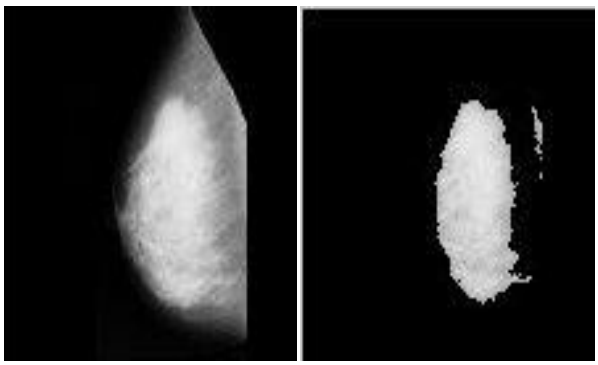

Fig 3: Segmented region

2.4 Feature Extraction

Transforming the input data into the set of features is called features extraction. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. The traditional goal of feature extraction is to characterize an object to be recognized by measurements whose values very similar for objects in same category and very different for object in different category[3]. Following feature values are extracted from the segmented images[7].

1.Mean: The mean, m of the pixel values in the defined window, estimates the value in the image in which central clustering occurs. The mean can be calculated using the formula:

$$\mu = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j)$$

Where p (i, j) is the pixel value at point (i, j) of an image of size M x N.

2. Standard Deviation: The Standard Deviation, s is the estimate of the mean square deviation of grey pixel value p(i, j) from its mean value m. Standard deviation describes the dispersion within a local region. It is determined using the formula:

$$\sigma = \sqrt{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (p(i,j) - \mu)^2}$$

3. Energy: Energy returns the sum of squared elements in Image. Energy is also known as uniformity. The formula for finding energy is given in below equation:

$$E = \sum_{i,j} P(i,j)^2$$

4. Entropy: Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. Entropy, h can also be used to describe the distribution variation in a region. Overall Entropy of the image can be calculated as:

$$h = -\sum_{k=0}^{L-1} Pr_k(log_2 Pr_k)$$

Where, Pr is the probability of the k-th grey level.

5. Skewness: Skewness, S characterizes the degree of asymmetry of a pixel distribution in the specified window around its mean. Skewness is a pure number that characterizes only the shape of the distribution. It is given by

$$S = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \left(\frac{p(i,j) - \mu}{\sigma}\right)^3$$

Where, m and σ are the mean and standard deviation respectively.

6. Kurtosis: Kurtosis, K measures the peakness or flatness of a distribution relative to a normal distribution. The conventional definition of kurtosis is:

$$K = \left\{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \left[\frac{p(i,j) - \mu}{\sigma}\right]^4\right\} - 3$$

Where the -3 term makes the value zero for a normal distribution.

7.Autocorrelation: Autocorrelation is the cross-correlation of a signal with itself. Informally, it is the similarity between observations as a function of the time separation between them. It is a mathematical tool for finding repeating patterns[11].

$$\text{Autocorrelation} = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1}(p_x - \mu_x)(p_y - \mu_y)/\sigma_x\sigma_y$$

2.5 Classification

For classification of samples, we have employed the freely available Machine Learning package, WEKA (Waikato Environment for Knowledge Analysis) to train our data set using J48 decision tree method. A special type of classifier is the decision tree which is trained by an iterative selection of individual features that are most salient at each node of the tree[5] .The main advantage of the tree classifier, besides its speed, is the possibility to interpret the decision rule in terms of individual features.

3. EXPERIMENTAL RESULTS

Ninety mammogram images (30 images from each class) are given as test data to the classifier. It was found that out of 90 images, 83 images were rightly classified and remaining 7 were misclassified to other categories.

Confusion matrix is shown in Table 1,

TABLE 1 CONFUSION MATRIX

Normal	Benign	Malignant	
29	**0**	**0**	**Normal**
1	**26**	**3**	**Benign**
0	**4**	**27**	**Malignant**

The matrix clearly shows that out of 30 normal images 29 were correctly classified as normal and 1 was misclassified as benign. Among the 30 benign images fed to the system 26 were rightly classified as benign and 4 was wrongly classified to malignant. Among 30 malignant images 27 were correctly classified and 3 was misclassified to benign.

4. ACKNOWLEDGMENTS

Author expresses his sincere gratitude Dr. Ashok Kumar T, Principal, VCET, Puthur for his valuable suggestions during the course of the work.The author also thank to Mr.Kirthiraj, Mr. Akshay, Mr.Harsha and Mr. Donal from SJEC. Mangalore for their help while conducting this experiment

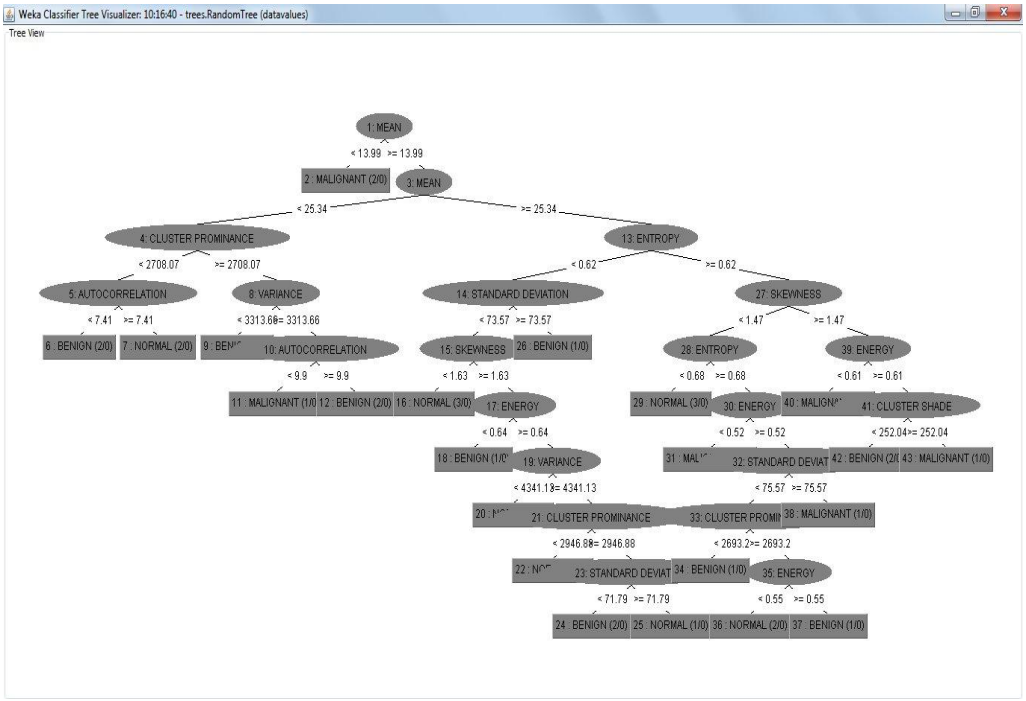

Fig 4: Structure of Decision tree

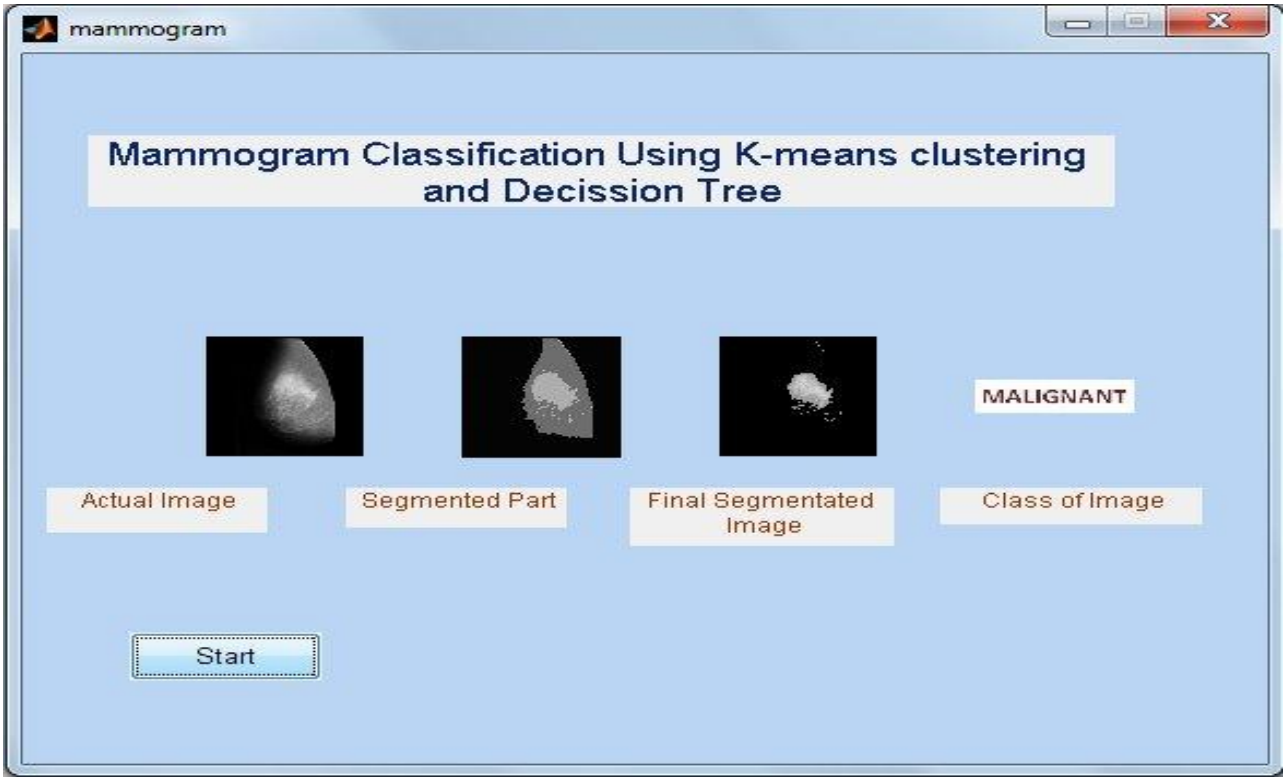

Fig.5: Output Snapshot

6. REFERENCES

[1] Gonzalez R.C, Woods R.E and Eddins S.,2002 Digital Image Processing Using MATLAB® -, 2e, Prentice Hall.

[2] Gonzalez R.C and Woods R.E 2002Digital Image Processing, 2e,Prentice Hall.

[3] Witten H. and Frank E, Data Mining Practical Machine Learning Tools and Techniques, Second Edition

[4] Duda R.,Hart P., Stork D, 2001 Pattern Classification , 2e, Wiley.

[5] Hand D , Mannila H, Smyth P., Principles Of Data Mining, Eastern Economy Edition , Prentice Hall

[6] Osmar R. Za¨ıane,Luiza M. Antonie, Alexandru C "Mammography Classification by an Association Rulebased Classifier".

[7] Vijayakumar C., Damayanti G. Chanda S., Sreedhar C and Bhargava B, "Wavelet and Co-occurance Matrix Based Artificial Neural Network Tools for the Segmentation of Multiple sclerosis Lesions on MR Images", IEEE Trans. Signal and Image Processing, Vol. 1, 2006, pp. 333-337

[8] Pisano E.D,Yaffe M.J, Kuzmiak C.M, "Digital Mammography",Lippincott Williams & Wilkins ,2004

.[9] Vasantha M .Bharathi S., "Classifications of Mammogram Images using HybridFeatures", Euro Journals Publishing, 2011.

[10] Nalini S.,Mohapatra A.G, Gurukalyan K., "Breast Cancer MassDetection in Mammograms using K-means and Fuzzy C-means Clustering", International Journal of computer applications, 2011.

[11] Vibha L, Harshavardhan G M, Pranaw K, P DeepaShenoy, Venugopal K R, L M Patnaik,"Classification of Mammograms Using Decision Trees", 10th International database Engineering and Applications symposiums (IDEAS ' 06), 2006

[12] T.N.C.I.W. site. Available: www.cancer.gov