International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 4, April - 2014

A Review: Breast Mass Classification in Mammogram Images

Santhoshi Sukumaran[1]
Department of Computer Science and Engineering
College of Engineering, Karunagappally,
Cochin University of Science and Technology,
Kerala, India.

Sunitha S[2]
Department of Computer Science and Engineering
College of Engineering, Karunagappally,
Cochin University of Science and Technology,
Kerala, India.

***Abstract* - Breast cancer is one of the most common forms of cancer among women and the second foremost cause of cancer deaths worldwide. Early detection of cancer is a key factor to a successful treatment. This study reviews different methods used by a Computer Aided Diagnosis (CAD) system for the classification of breast masses related to breast cancer. A mass is one of the main signs of breast cancer. The dataset consisting of digital mammograms is used for the study. The breast mass is classified as either benign or malignant. Accurate determination of the nature of the mass is very important in reducing the false positives and unnecessary biopsies.**

***Index Terms* – Breast mass classification, mammogram, fuzzy subtractive clustering.**

I. INTRODUCTION

The high prevalence of breast cancer has increased drastically in the recent years. It is a principal cause of fatality in women, with approximately 1 in 12 women affected by the disease during their lifetime [12]. In the past two decades, breast cancer has been the second leading cause of cancer deaths among women in the United States of America, next only to lung cancer. The increasing number of deaths is mainly due to the fact that the cancer is not identified at an early stage. Chances of survival in case of breast cancer are found to be stage-dependent and the best survival is attained when disease is diagnosed at an early stage. False-positive results are more common for younger women, women who have had previous breast biopsies, women with a family history of breast cancer etc.Mammography is an effective tool for early detection because in many cases it can detect abnormalities such as masses and calcifications up to two years before they are palpable. Moreover it reduces the false positives thereby reducing the anxiety and other forms of psychological distress in affected women [15].

With the advancement in the field of computer technology, radiologists have a chance to improve their interpretation of the image using computer capabilities that can improve the image quality of mammograms. Over the past two decades, many attempts have been made to aid the radiologists in the detection and diagnosis of masses by developing computer-aided tools for mammogram interpretation. Image processing and intelligent systems are two main areas of computer technologies that have been persistently explored in the development of computer-aided mammography systems.

A. *Abnormalities of the Breast*

This happens as a result of mutations, or abnormal changes, in the genes that are responsible for regulating the cell growth in the breast tissues and keeping them healthy. These abnormal changes give cells the ability to keep dividing without control thereby producing more cells similar to it and forming a tumor. There are two main abnormalities that lead to cancer in the breast region. These abnormalities include a mass in the breast and calcification. Calcifications are small mineral (calcium) deposits within the breast and these appear as localized high-intensity regions (spots) in the mammogram. The location, size, shape, density, and margins of the mass or lumps are used by the radiologist in evaluating the likelihood of cancer. A breast mass may be benign or malignant in nature. The benign mass is non-cancerous in nature and do not spread to other parts of the body. On the other hand a malignant mass is cancerous and can invade and damage nearby tissues.

B. *Imaging modalities*

A wide range of imaging modalities is used by researchers in different works. Some of these include ultrasound, ductogram (galactogram), thermography (thermal imaging), Magnetic Resonance Imaging (MRI), and mammography [15]. Ultrasound, also known as sonography, uses sound waves to view inside a part of the body. A gel is put on the skin of the breast and an instrument called transducer is rubbed with gel and pressed against the skin. It emits sound waves and picks up the echoes as they bounce from body tissues. The echoes are converted by a computer into a black and white image on a computer screen. This test is painless and does not involve an exposure to radiation.

A ductogram, also known as a galactogram, is sometimes used to help find the cause of nipple discharge. Here test, a very thin plastic tube is put into the opening of a duct in the nipple that the discharge coming from. A small amount of contrast material is put in. It outlines the shape of the duct on x-ray and can show whether there is a mass inside the duct.

Thermography is a method that measures and maps the heat on the surface of the breast using a special heat-sensing camera. It is based on the idea that the temperature rises occurs in areas where there is increase in blood flow and metabolism, which could be a sign of a tumor. Thermography has been around for many years, but studies have shown that it is not an effective screening tool for finding breast cancer early.

Mammography is an x-ray image of the breast formed by a diverging x-ray beam. Thus, the breast volume attenuation is represented by light and dark shadows captured in a film screen combination process; the resulting image is planar projection of the three dimensional breast. Mammogram can be applied to check for breast cancer in women who have no symptoms of the disease. This type of mammogram is referred to as a screening mammogram. Screening mammograms usually involves two x-ray pictures, or images, of each breast. The x-ray images make it possible to detect tumours that cannot be felt.

II. LITERATURE REVIEW

A. *Image Pre-processing*

In image pre-processing techniques are necessary in order to find the orientation of the mammogram to remove the noise and to enhance the quality of the image. Usually the procedure used for denoising, is dependent on the features of the image, aim of processing and also post-processing algorithms. Denoising by the use of low-pass filters not only reduces the noise but also blurs the edges. Spatial and frequency domain filters are commonly used as tools for image enhancement. Low pass filters smoothens the image by blocking detail information.

The most commonly used filters in the preprocessing of mammograms are Adaptive Median Filter, Adaptive Mean filter, Mean filter etc. Adaptive Median filtering has been found to smooth the non repulsive noise from two-dimensional signals without blurring edges and preserve image details [13]. The mean filter replaces each pixel by the average value of the intensities in its neighborhood. It can locally reduce the variance and is easy to implement. In order to alleviate the blurring effect, the adaptive mean filters have been proposed to achieve a balance between straightforward averaging (in homogeneous regions) and all-pass filtering (where edges exist). They adapt to the properties of the image locally and selectively remove speckles from different parts of the image. They use local image statistics such as mean, variance and spatial correlation to effectively detect and preserve edges and features. The adaptive mean filters outperform mean filters, and generally reduce speckles while preserving the edges.

Another technique used in preprocessing is Histogram Equalization. This technique corresponds to redistribution of gray levels in order to obtain uniform histogram. In this case every pixel is replaced by integral of the histogram of the image in that pixel. Through this adjustment, the intensities can be better distributed on the histogram. This allows for areas of lower local contrast to get better contrast. Histogram equalization accomplishes this by efficiently spreading out the most frequent intensity values. The method is useful in images with backgrounds and foregrounds that are both bright or both dark.

B. *Extraction of Region of Interest*

Extraction of Region of Interest (ROI) is an important step. It is actually image segmentation and partitions the image into groups of pixels which are homogeneous with respect to some criterion. These may include values of intensity, texture, color, range, surface normal and surface curvatures. The ROI is extracted to separate the suspicious regions that may contain masses from the background and locate the suspicious mass candidates from ROIs. The suspicious area is an area that appears brighter than its surroundings, has almost uniform density, has a regular shape with varying size, and has fuzzy boundaries. Different techniques can be used to obtain the ROI.

Thresholding is one of the common procedures for image segmentation. It is a simple and effective method for images with different intensities. These thresholding techniques are very much useful for image binarization which is very essential task for any type of segmentation. It assumes that images are composed of regions with different gray levels. Thresholding can be local or global in nature based upon how the threshold or intensity value is chosen. A local thresholding process obtains the threshold locally which separates the desired classes. Global thresholding is based on global information, such as the histogram of the mammograms.

Based on classification of pixels, two different types of segmentation are defined: region growing and region clustering [3]. Region growing is one of the popular techniques for segmenting masses in digitized mammograms. The basic idea of the algorithm is to find a set of seed pixels in the image first, and then to grow iteratively and aggregate with the pixels that have similar properties. If the region is not growing any more, then the grown region and surrounding region are obtained. Region clustering searches the region directly without any prior information. The clustering process separates one or more disjoint objects within the ROIs, which were filled, grown in a local neighborhood, and eroded and dilated by morphological operators.

Multiscale techniques can also be applied to segment the suspicious areas, and this may improve the detection rate. Tumors with the radii between 2 and 30mm can be detected at different scales. Discrete wavelet transform (DWT) is a powerful mathematical tool for image analysis, and DWT is one of the multiscale techniques [3].

C. *Feature Extraction*

Feature refers to a piece of information that has relevance in solving the computational task related to a certain application. Feature extraction can be defined as a quantitative measurement or analysis of the medical images. To deal with the abnormalities of the mammograms, many types of features can be extracted. The different types of features calculated from the extracted ROI broadly come under the category statistical, geometrical or structural.

The statistical features are the simplest ones and they include mean, standard deviation, variance etc .Mean of a defined window denotes the value in the image where central clustering takes place. Standard deviation refers to estimate of the mean square deviation of a pixel from its mean. It describes the dispersion within a local region. Variance is the square root of standard deviation. Skewness is another statistical feature that describes the degree of asymmetry of a pixel distribution in the specified window around its mean. It characterizes the shape of the distribution [2].

Texture provides information about the spatial distribution of intensity levels in a neighborhood. So, it cannot be defined for a point. Smoothness is a texture feature that measures the grey level contrast. Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. Entropy describes the distribution variation in a region. Energy provides the sum of squared elements in the Grey Level Co-Occurrence Matrix (GLCM). Energy is also known as uniformity. Contrast returns a measure of the intensity contrast between a pixel and its neighbor over the whole image [5]. Correlation returns a measure of how correlated a pixel is to its neighbor over the whole image. Homogeneity returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal [7].

The main structural or geometrical features are related to the geometry i.e. size, geometrical shape and boundary of the concerned region. In a region Area is the actual scalar count of pixels, Centroid is the center of the region. Bounding box is the smallest rectangle that contains the region. Filled Area is the number of pixels in the filled image. Equiv Diameter is the diameter of a circle with the same area as the region.

The features related with geometrical shape are as follows: Euler Number denotes the number of objects in the region minus the number of holes in those objects. Extrema denotes the extremal points in the region. Convex Hull is the smallest convex polygon that can contain the region. Solidity is the proportion of the pixels in the convex hull that are also in the region.

The features related with the boundary are as follows: Major Axis Length denotes the length (in pixels) of the major axis of the ellipse that has the same second-moments as the region. Minor Axis Length denotes the length (in pixels) of the minor axis of the ellipse that has the same second moments as the region. Eccentricity denotes the eccentricity of the ellipse that has the same second-moments as the region and it is shown as the ratio of the distance between the foci of the ellipse and its major axis length. Orientation indicates the angle (in degrees) between the x-axis and the major axis of the ellipse that has the same second-moments as the region. Extent represents the proportion of pixels in the bounding box that are also in the region [1].

D. Clustering Techniques

Clustering is a process of partitioning or grouping a given set of unlabeled patterns into a number of clusters in which similar patterns are assigned to one cluster. Each pattern can be represented by a vector having many parameters or attributes. The computation of a measure of similarity or distance between the respective patterns is fundamental to the use of any clustering technique. Clustering identifies a suitable group of a dataset, such that patterns in the data gives a concise representation of the behaviour of the similar data in a group that has similar characteristics, and different clusters have characteristics that are not identical. In most cases a cluster is represented by a cluster centre or a centroid, which will then used to form the membership function. Over the years, many methods have been developed for clustering patterns. Each method can have its own technique (i.e. partitioning or hierarchical), mode (on-line or offline), approach (fuzzy or crisp clustering), or special purpose (i.e. for sequential data set, very large database, etc.)

Hierarchical clustering is a graphical representation of data. Partitional clustering is considered the second general category of clustering. It concerns with building partitions (clusters) of data sets according to the relative proximity of the points in the data sets to each other. The K-means algorithm (or Hard C-means clustering), is a crisp clustering algorithm and it partitions a collection of n vector into c groups and finds a cluster center in each group such that a cost function (or an objection function) of dissimilarity (or distance) measure is minimized. Fuzzy C-Means clustering (FCM), is a supervised clustering method in which each data point belongs to a cluster to a degree specified by a membership grade [10].

The mountain clustering method is a relatively simple and effective approach to approximate estimation of cluster centers on the basis of a density measure called the mountain function. This method determines the cluster centers using three steps. The first step involves forming a grid the data space, where the intersections of the grid lines constitute the candidates for cluster centers. The second step involves constructing a mountain function representing a data density measure. The third step involves selecting the cluster centers by sequentially destructing the mountain function. Its computation grows exponentially with the dimension of the patterns because the method must evaluate the mountain function over all grid points [8].

Subtractive clustering is an alternative approach to the mountain clustering method. It was proposed by Chiu and is a method that can automatically extract fuzzy rules from data. In this method, data points (not grid points) are considered as the candidates for cluster centers. This can automatically generate multiple rules .The subtractive algorithm is based on a measure of the density of data points in which a data point with many neighboring points has the potential to be the cluster centre. By using this method, the computation is simply proportional to the number of data points and independent of the dimension problem [10].

E. Classification Techniques

Once the features related to masses are extracted and selected, the features are input into a classifier to classify the detected suspicious areas into normal tissues, benign masses, or malignant masses. Classifiers such as artificial neural network (ANN) have performed well in mass classification. ANNs have been applied in business, medicine, robotics, manufacturing, industrial, as solutions to a variety of problems such as forecasting, decision making, speech recognition, classification of text, signal processing and controllers, image processing, pattern recognition, neurological and cognitive modeling [6].

In the medical field, ANNs have been used since the late 1980s, initially to identify accuracy of survival prediction, classification of cancerous tumors etc. The advantage of ANNs is their capability of self-learning, and often suitable to solve the problems that are too complex to use the conventional techniques, or hard to find algorithmic solutions. Generally, a known database of mammograms, including the selected features and the desire results, is selected to train the ANN. After the weights are determined, the ANN is ready to classify the masses. The classification process is divided into the training phase and the testing phase. During training, the features are extracted from the images in which the diagnosis is known. After training is over, the trained networks are stored to be used in the algorithm. Whenever an image is taken as input in the algorithm, it is simulated with the trained net-works and goes for testing the data.

The k-Nearest Neighbours classifier (kNN) consists of the assignment of an unclassified vector using the closest K vectors found in the training set. Usually, the Euclidean distance is used. Due to the fact that kNN is based on distances between sample points in the feature space, features need to be normalized to avoid that some features are weighted more strongly than others [3]. Hence, all features have been normalized to unit variance and zero mean.

A binary decision tree recursively uses a threshold to separate mammogram data into two classes every time by choosing a threshold to split input data into two classes each time. An ordered list of binary threshold operations on the features is organized as a tree [3]. Each node has a threshold associating with one or more features to divide the data into its two descendents. The process stops when it only contains patterns of one class. It uses intensity features, shape features and texture features.

LDA (Linear Discriminant Analysis) is a conventional method for classification. The key idea of this method is to construct the decision boundaries directly by optimizing the error criterion to separate the classes of objects.

The SVM (Support Vector Machine) is based on the statistical learning theory which describes the properties of learning machines that allow them to give reliable predictions.SVM performs an implicit mapping of data into a higher dimensional feature space, where linear algebra and geometry can be used to separate data [14] .SVM algorithm constructs a separating hyper surface in the input space by mapping the input space into a high dimensional features space through some non linear mapping chosen a priori (kernel) or constructing in this features space the Maximal Margin Hyper plane.

Fuzzy logic has been used subsequently in many medical fields for over a decade in data classification, decision analysis, diagnosis and prognosis [5].A combination of ANN and fuzzy logic is implemented in ANFIS (Adaptive Neuro Fuzzy Inference System).

III. CONCLUSION

Here different steps in a Computer Aided Diagnosis (CAD) system for breast mass classification have been discussed for mammogram images. The main steps for an efficient classification of mass involves enhancement, extraction of ROI, feature extraction, clustering etc. Each step can be implemented using a number of methods. Classification of breast mass is one of the most challenging and active research topics in the field of image processing for the last decade. The researchers use a combination of various methods to improve the accuracy and performance of the system.

REFERENCES

[1] Pelin Gorgel, Ahmet Sertbas and Osman N.Ucan, "A Fuzzy Inference System Combined with Wavelet Transform for Breast Mass Classification."

[2] Pradeep N, Girisha H, Sreepathi B, Karibasappa K, "Feature Extraction Of Mammograms", *International Journal of Bioinformatics Research ,ISSN: 0975–3087 & E-ISSN: 0975–9115 , Volume 4, Issue 1, 2012, pp.-241-244*

[3] H.D. Cheng, X.J. Shi, R. Min, L.M. Hu, X.P. Cai, H.N. Du, "Approaches for automated detection and classification of masses in mammograms". *Pattern Recognition 39 (2006) 646 – 668.*

[4] R.Ramani, Dr.S.Suthanthiravanitha, S.Valarmathy, "A Survey of Current Image Segmentation Techniques for Detection of Breast Cancer", *International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 5, September- October 2012, pp.1124-1129.*

[5] K. Yuvaraj1, U.S. Ragupathy, "Computer Aided Segmentation and Classification of Mass in Mammographic Images using ANFIS", *EJBI 2013; 9(2):37–41, published: August 30, 2013.*

[6] Mohammed J. Islam, Majid Ahmadi and Maher A. Sid-Ahmed, "An Efficient Automatic Method in Digitized Mammograms Using Artificial Neural Network", *International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.3, July 2010.*

[7] Sulochana Wadhwani, A.K.Wadhwani, Monika Saraswat, "Classification of Breast Cancer Detection", Current Research in Engineering, Science and Technology (CREST) Journals, Vol 01, Issue 03, May 2013.

[8] Khaled Hammouda, Prof. Fakhreddine Karray, "A Comparative Study of Data Clustering Techniques".

[9] Pelin Gorgel, Ahmet Sertbas and Osman N.Ucan, "Feature Extraction Based Wavelet Transform in Breast Cancer Diagnosis Using Fuzzy and Non-fuzzy Classification", *International Journal of Electronics, Mechanical and Mechatronics Engineering, Vol.2 Num.4 pp. (327-333).*

[10] Veronica S.Moertini, "Introduction to Five Data Clustering Algorithms", *Integral Vol.7, No.2, October 2002.*

[11] Monireh Sheikh Hosseini, Maryam Zekri, "Review of Medical Image Classification Using the Adaptive Neuro Fuzzy Inference System", *Journal of Medial Signals and Sensors, Vol.2, Issue 1, March 2012.*

[12] American Cancer Society. Breast Cancer Facts and Figures, *http://www.cancer.org/*

[13] D.Narain Ponraj, M.Evangelin Jenifer, P. Poongodi, J.Samuel Manoharan, "A Survey on the Preprocessing Techniques of Mammogram for the Detection of Breast Cancer", *Journal of Emerging Trends in Computing and Information Sciences,Vol .2,No.12,December 2011.*

[14] Nasseer M.Basheer, Mustafa H.Mohammed, "Classification of Breast Masses in Digital Mammograms Using Support Vector Machines", *International Journal of Advanced Research in Computer Science and Software Engineering ,Vol.3,Issue 10,October 2013.*

[15] National Cancer Institute. Facts Sheet, *http://www.cancer.gov/cancertopics/factsheet/detection/mammograms*