IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010
ISSN (Online): 1694-0814
www.IJCSI.org

A Theoretical Methodology and Prototype Implementation for Detection Segmentation Classification of Digital Mammogram Tumor by Machine Learning and Problem Solving Approach

***VALLIAPPAN Raman 1, PUTRA Sumari 2 and MANDAVA Rajeswari 3**

**[1] School of Computer Science, University Sains Malaysia,
George town, Penang 11800, Malaysia**

**[2] School of Computer Science, University Sains Malaysia,
George town, Penang 11800, Malaysia**

**[3] School of Computer Science, University Sains Malaysia,
George town, Penang 11800, Malaysia**

Abstract

Breast cancer continues to be a significant public health problem in the world. Early detection is the key for improving breast cancer prognosis. The CAD systems can provide such help and they are important and necessary for breast cancer control. Microcalcifications and masses are the two most important indicators of malignancy, and their automated detection is very valuable for early breast cancer diagnosis. The main objective of this paper is to detect, segment and classify the tumor from mammogram images that helps to provide support for the clinical decision to perform biopsy of the breast. In this paper, a classification system for the analysis of mammographic tumor using machine learning techniques is presented. CBR uses a similar philosophy to that which humans sometimes use: it tries to solve new cases of a problem by using old previously solved cases. The paper focus on segmentation and classification by machine learning and problem solving approach, theoretical review have been undergone with more explanations. The paper also describes the theoretical methods of weighting the feature relevance in case base reasoning system.

Key words: *Digital Mammogram, Segmentation, Feature Extraction and Classification.*

1. Introduction

Breast cancer is the most common cancer of western women and is the leading cause of cancer-related death among women aged 15-54 [14]. Survival from breast cancer is directly related to the stage at diagnosis. Earlier the detection, higher chances of successful treatments. In an attempt to improve early detection, a study has been undertaken to process the screening mammograms of breast cancer patients in order to analyze the mass/microcalcifications features that help to differentiate benign from malignant cases.

In this paper we propose the selected shape-based features in order to classify clustered masses between benign and malignant [15]. The computerized analysis of mammographic masses performed in this work can be divided into four stages: 1) digitization of mammograms and enhancement of images; 2) detection of suspicious areas; 3) extraction of features for every segmented tumors in the digitized mammogram; and 4) analysis of the features using Case Based Reasoning techniques. A Case-Based Reasoning algorithm is used for classifying these cases into benign or malignant cases. We have to be aware that Case-based Reasoning means using previous experience in form of cases to understand and solve new problems [13].

The main objective of this paper is to focus on the segmentation and theoretical review for classification of tumor by case base reasoning approach and how to apply weights to the features, and improve the accuracy rate. The paper is organized as follows: In Section 2, it clearly explains the existing works of mammography detection, then machine learning technique is explained in section 3, Next in section 4 problem solving capabilities are explained, next experimental results are shown in Section 5. Section 6 discusses the shown experimental results. Finally conclusion and future works are specified in

section 6.

2. Existing Research Works

The problem of image processing has been divided into several research areas and medical research has been quite receptive of image processing in applications like x ray, computer aided tomography, ultrasound and magnetic resonance [8]. There are several existing approaches were made to detect the abnormal tissues in breast images and to detect the cancer earlier.

Zhang et al. [3] noted that the presence of spiculated lesions led to changes in the local mammographic texture. They proposed that such a change could be detected in the Hough domain, which is computed using the Hough transform. They partitioned an image into overlapping ROIs and computed the Hough transform for each ROI. The Hough domain of each ROI was thresholded to detect local changes in mammographic texture and to determine the presence or absence of a spiculated mass.

Brzakovic et al. [4] use a two stage multi-resolution approach for detection of masses. First they identified suspicious ROIs using Gaussian pyramids and a pyramid linking technique, based on the intensity of edge links. Edges were linked across various levels of resolution. This was followed by a classification stage, where the ROI were classified as malignant, benign or normal based on features like shape descriptors, edge descriptors and area

Petrick et al. [5] developed a two-stage algorithm for the enhancement of suspicious objects. In the first stage they proposed an adaptive density weighted contrast enhancement filter (DWCE) to enhance objects and suppress background structures. The central idea of this filtering technique was that it used the density value of each pixel to weight its local contrast. In the first stage the DWCE filter and a simple edge detector (Laplacian of Gaussian) was used to extract ROIs containing potential masses. In the second stage the DWCE was re-applied to the ROI. Finally, to reduce the number of false positives, they used a set of texture features for classifying detected objects as masses or normal. They further improved the detection algorithm by adding an object-based region-growing algorithm to it.

Lai [6] made an approach based on a multiresolution Markov random field model detect mass lesions. Its initial window size for segmentation influences the sensitivity of detection. Li [7] proposed a method on iris filter was developed to detect mass lesions of rounded convex regions with low contrast. The iris filter enhances most round malignant masses. However, some malignant masses are shaped irregularly.

The above methods show less than five false positives per image with a true positive detection rate of approximately 90%.It is difficult to compare the performance of these methods because their databases are different.

3. Machine Learning Approach

Compare to all the existing works; we developed mammographic tumor segmentation by region growing approach and classification using case base reasoning approach [15]. In this paper there are two stages; first stage includes machine learning approach such as digitizing the images, preprocessing and segmentation. Tumor segmentation classified as two types: Mass and Microclacification.It is more difficult to detect masses than microcalcifications because their features can be obscured or similar to normal breast parenchyma. Masses are quite subtle, and often occurred in the dense areas of the breast tissue, have smoother boundaries than microcalcifications, and have many shapes such as circumscribed, speculated lobulated or ill-defined. Second stage is the problem solving approach using Case Base Reasoning method; new cases are solved by previous solved old cases, which is the main focus of the paper. Figure 1 illustrates the overall block diagram of tumor Classification method.

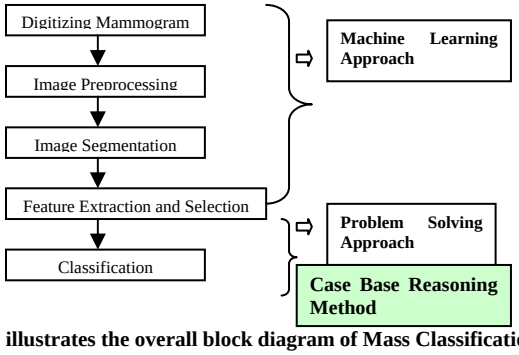

Fig 1 illustrates the overall block diagram of Mass Classification method

3.1 Digitization

First, the X-ray mammograms are digitized with an image resolution of 100 × 100 μm^2 and 12 bits per pixel by a laser film digitizer. To detect microcalcifications on the mammogram, the X-ray film is digitized with a high resolution. Because small masses are usually larger than 3mm in diameter, the digitized mammograms are decimated with a resolution of 400 × 400 mm 2 by

averaging 4 × 4 pixels into one pixel in order to save the computation time.

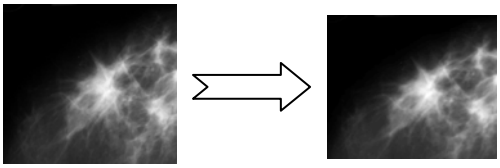

Fig 2 illustrates original image [16] by reducing image dimension in averaging of 8×8 Matrix

3.2 Preprocessing

Preprocessing is an important issue in low-level image processing. The underlying principle of preprocessing is to enlarge the intensity difference between objects and background and to produce reliable representations of breast tissue structures. An effective method for mammogram enhancement must aim to enhance the texture and features of tumors. The reasons are: (1) low-contrast of mammographic images; (2) hard to read masses in mammogram because it is highly connected to surrounding tissues; the enhancement methods are grouped as global histogram modification approach and local processing approach. Current work is carried out in global histogram modification approach.

Preprocessing Approach	Description	Advantage
Global Histogram Modification Approach	Re-assign the intensity values of pixels to make the new distribution of the intensities uniform to the utmost extent	Effective in enhancing the entire image with low contrast
Local Approach	Feature-based or using nonlinear mapping locally	Effective in local texture enhancement

Table 1 illustrates the preprocessing approach

3.3 Segmentation

After preprocessing, next stage is to separate the suspicious regions that may contain masses from the background parenchyma, that is to partition the mammogram into several non-overlapping regions, then extract regions of interests (ROIs), and locate the suspicious mass candidates from ROIs. The suspicious area is an area that is brighter than its surroundings, has almost uniform density, has a regular shape with varying size, and has fuzzy boundaries. The Segmentation methods do not need to be excruciating in finding mass locations but the result for segmentation is supposed to include the regions containing all masses even with some false positives (FP). FPs will be removed at a later stage. We chose region growing process for segmentation of a mammographic tumor.

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 growing may be applied globally or locally. If the grown region of a seed has an average intensity greater than that of the surrounding, the region is classified as the parenchyma, or fat, tissue.The accuracy reaches 70% for classifying the tissue patterns. The key issue of region growing is to find a criterion that checks whether the gray level values of its neighbors are within a specified deviation from the seed. The performance of the algorithm depends on the enhancement method; therefore the algorithm will get a better result if a better enhancement method is applied. Global histogram Modification Enhancement method was applied to enhance the images before region growing. Second issue of region growing is to find the suitable seeds. An automatic seed selection was applied. There are three parts in mammograms: a fat region, a fatty and glandular region, and a dense region. According to the intensity values and local contrast between a seed pixel and its neighbors in the three partitions, three sets of seed pixels are selected from the partitioned regions.

The region growing process starts from seed pixels. The gray level mapping shows local valleys at the boundary of two neighboring regions. The local peak just after the local valley in the gray level mapping gives a sign of the switch between the absorption of pixels in the boundary of the current region and the absorption of pixels in the neighboring region. When the grown region size is equal to or greater than a minimum region size with the stopping condition such as speckle noise, touching previous region, new adjacent region, contrast limitation. Once the stopping condition is achieved, region growing is applied and the masses are segmented. Below algorithm summarizes the region growing procedures for segmenting the masses.

Algorithm

1. Pull the top item from the growth list.
2. Mark this pixel in the output image - it is part of the region.
3. Examine each neighboring pixel. For each pixel, if it has not already been visited and it fits the growth criteria, mark it as visited and add it to the growth list.
4. Go back to step 1 and repeat until there are no more items in the growth list and extract the part of tumor.

4. Feature Extraction

After segmenting the tumors in mammogram, The ROI hunter provides the "regions of interest" without giving further information [12]. To this purpose suitable features should be selected so that a decision making system can correctly classify possible pathological regions from healthy ones. Feature extraction plays a fundamental role in many pattern recognition tasks. In this paper twelve features (global and local features) are extracted from the segmented tumors. Below table illustrates the features.

Feature of Selection	Description
Skewness	$\frac{1}{N}\frac{\sum_{i,j=0}^{N-1}\left[g(i,j)-\overline{g(i,j)}\right]^3}{\sqrt{\sum_{i,j=0}^{N-1}\left[g(i,j)-\overline{g(i,j)}\right]^3}}$
Kurtosis	$\frac{1}{N}\frac{\sum_{i,j=0}^{N-1}\left[g(i,j)-\overline{g(i,j)}\right]^4}{\sqrt{\sum_{i,j=0}^{N-1}\left[g(i,j)-\overline{g(i,j)}\right]^4}}$
Circularity	$\frac{A_1}{A}$
Compactness	$\frac{P^2}{A}$
Contrast	$\frac{P_1-P_2}{P_1}$
Standard deviation	σ^2
Intensity	$\overline{g(i,j)} = 1/N\sum_{i,j=0}^{N-1} g(i,j)$
Area	*tumor area*
Length	*True Length of Mass*
Breadth	*True Breadth of Mass*
Convex Perimeter	*Perimeter of the convex hull of the mass*
Roughness	*Perimeter/Convex Perimeter*

Table 2 illustrates the Local and Global Features

5. Case Base Reasoning Approach

Case-Based Reasoning (CBR) integrates in one system two different characteristics: machine learning capabilities and problem solving capabilities. CBR uses a similar philosophy to that which humans sometimes use: it tries to solve new cases (examples) of a problem by using old previously solved cases. The process of solving new cases contributes with new information and new knowledge to the system. This new information can be used for solving other future cases. The basic method can be easily described in terms of its four phases .The first phase retrieves old solved cases similar to the new one. In the second phase, the system tries to reuse the solutions of the previously retrieved cases for solving the new case. The third phase revises the proposed solution. Finally, the fourth phase retains the useful information obtained when solving the new case. In a Case-Based Classifier System, it is possible to simplify the reuse phase. Classifying the new case with the same class as the most similar retrieved case can do reuse [13].

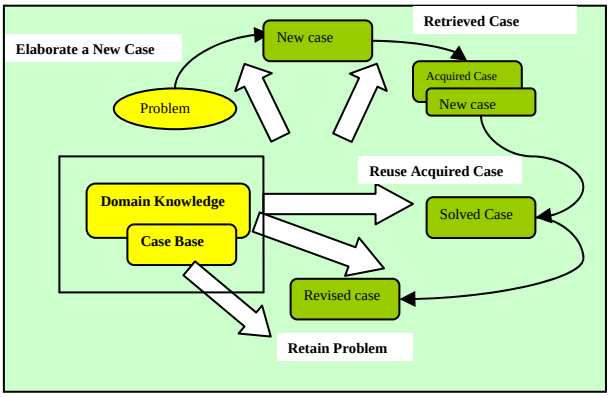

Fig 3 illustrates the Case Base Reasoning Approach

The kernel in a Case-Based Reasoning system is the retrieval phase (phase 1). Phase 1 retrieves the most similar case or cases to the new case. Obviously, the meaning of most similar will be a key concept in the whole system. Similarity between two cases is computed using different similarity functions. For our purpose in this paper, we use the similarity functions based on the distance concept. The most used similarity function is the Nearest Neighbor algorithm, which computes the similarity between two cases using a global similarity measure. The future practical implementation (used in our system) of this function is based on the Minkowski's metric.

Minkowski's metric is defined as:

$$Similarity(Case\_x, Case\_y) = \sqrt[r]{\sum_{i-1}^{F} W_i \times |x_i - y_i|^r}$$

(1)

Where *Case_x*, *Case_y* are two cases, whose similarity is computed; F is the number of features that describes the case; $x_i y_i$ represent the value of the ith feature of case *Case_x* and *Case_y* respectively; and w_i is the weight of the ith feature. In this study we test the Minkowsky's metric for three different values of r: Hamming distance (r = 1), Euclidean distance (r = 2), and Cubic distance (r = 3). This similarity function needs to compute the feature relevance (w_i) for each problem to be solved. Assuming an accurate weight setting, a case-based reasoning system can increase their prediction accuracy rate. We use also the Clark's and the Cosine distance, both are based on distance concept and also use weighting features. Sometimes human experts can not adjust the feature relevance, automatic method can solve this limitation.

5.1 Feature Selection Based on Rough Set theory

This paper presents a review on weighting method based on the Rough Sets theory introduced by Pawlak [10]. It is a single weighting method (RSWeight) that computes the feature weights from the initial set of train cases in the CBR system. We also introduce a weighting method that computes the Sample Correlation among the features and the classes that the cases may belong to. The idea of the rough set consists of the approximation of a set by a pair of sets, called the lower and the upper approximation of this set. In fact, these approximations are inner and closure operations in a certain topology generated by the available data about elements of the set. The main research trends in Rough Sets theory which try to extends the capabilities of reasoning systems are: (1) the treatment of incomplete knowledge; (2) the management of inconsistent pieces of information; (3) the manipulation of various levels of representation, moving from refined universes of discourse to coarser ones and conversely .

We compute from our universe (finite set of objects that describe our problem, the case memory) the concepts (objects or cases) that form partitions of that Universe. The union of all the concepts made the entire Universe. Using all the concepts we can describe all the equivalence relations (R) over the universe. Let an equivalence relation be a set of features that describe a specific concept. The universe and the relations form the knowledge base, defined as KB = (U; R). Every relation over the universe is an elementary concept in the knowledge base [10].

All the concepts are formed by a set of equivalence relations that describe them. So we search for the minimum set of equivalence relations that define the same concept as the initial set. The set of minimum equivalence relations is called reduct. A reduct is the essential part, which suffices to define the basic concepts occurring in the knowledge. The core is the set of all indispensable equivalence relations over the universe, in a certain sense the most important part of the knowledge. The core is defined as the intersection of all the reducts. Reducts contain the dependencies from the knowledge. We can use this information to weigh the relevance of each feature in the system [10]. An attribute that does not appear in the reduct has a feature weight value of 0.0, whereas an attribute that appears in the core has a feature weight value of 1.0. The rest has a feature weight value depending on the proportional appearance in the reducts. This is the weight feature information that we use in the case-based classifier system.

5.2 Sample Correlation

Sample Correlation computes the weights w_i computing the sample correlation which exists between each feature x_i and the class z [10].

The Sample Correlation is defined as:

$$Sample\_Correlation(x_i, z) = \frac{1}{N-1}\sum_{j-1}^{N}\left(\frac{x_{ij-\bar{x}_i}}{S_{x_i}}\right)\left(\frac{z_{j-\bar{z}}}{S_z}\right) \quad (2)$$

Where N is the number of cases; x_{ij} is the value of ith feature for the case j; z_j is the class which belong to the case j. $\in_i$ is the mean of the ith feature;| is the mean if the classes; S_{xi} is the standard deviation of the feature x_i ; and S_z is the standard deviation of class z.

Therefore weighting feature method needs a huge amount of cases to develop a good weighting feature selection during the retrieval phase.

If the system accuracy rate increases, then there is enough information in the system to develop a good weighting policy

6. Experimental Results

Currently the project is in the initial stage (prototype) and first phase of implementations are done in matlab. Therefore there are forty six X-ray mammograms taken for testing the method. The mammograms were taken from the patient files in the Free Mammogram Database (MIAS). In addition, 10 mammograms were used for training of the classifier. The 46 mammograms include 15

malignant and 10 benign masses that are in dense regions with glandular tissues, various breast areas involving ducts, breast boundaries, blood vessels, and/or glandular tissues. After segmentation, feature extraction and classification need to performed and tested. The below results show the various stages of mammogram segmentation. Feature extraction and Classification need to be refined and implemented in future works.

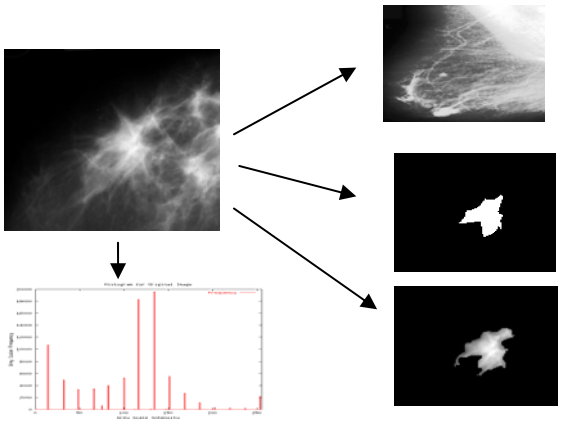

Fig 4 illustrates the original image from MIAS database [16].Preprocessing the Image, Hunting for ROI, Histogram of Original Image and segmenting the tumor.

Experiment	Segmentation Result	Accuracy
Experiment 1	Benign	63.3%
Experiment 2	Malignant	73.7%
Experiment 3	Benign	68.6%

Table 2 illustrates the results of Benign and Malignant tumors

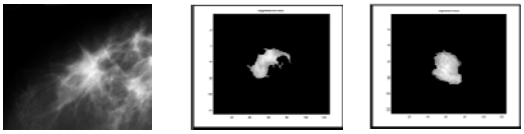

Fig 5 illustrates the original mammogram image [16], segmentation tumor for malignant cases and benign cases using region growing method.

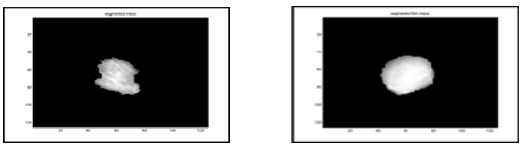

Fig 6 illustrates the results of segmented malignant tumors

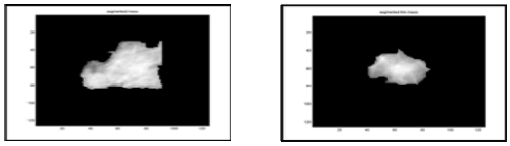

Fig 7 illustrates the results of segmented benign tumors

The efficiency and complexity of this system was improving than other systems presented in literature. The performance of the algorithm on the training set was a TP rate of 62% with 0.4 FP per image. The algorithm was tested using a set of 46 images, consisting of 15 malignant cases and 10 benign cases. Performance of the algorithm on the testing set was a TP of 60% and average of 0.5 false clusters per image. Table 2 shows that average accuracy of detecting benign and malignant ranges in 60% -70%.

Discussion

The tests were divided into 2 stages. The first part concerns subjective tests of the initial detection of masses and another one focused on the classification using the selected feature sets. The subjective tests were performed by image processing and experts in radiology. Radiologist's suggestions were used regarding mass shapes and their occurrence terms. The initial detection of mass was optimized with regard to influence of image pre-processing, size and shape of structuring element in global histogram modification approach, whereas to re-assign the intensity values of pixels to make the new distribution of the intensities uniform to the atmost extent. After that region growing segmentation is applied to detect and segment the tumor. On the segmented tumor part, features are selected for classification. Therefore an m1×12 real valued matrix is obtained for each mammography, which contains as many rows (m) as the number of masses are analyzed in the image, while the number of columns (12) is related to the computed shape features for every mass. In order to feed this information to the system of case base classification, the matrix is flattened into a vector. This process is achieved computing the mean value of each feature of the mass present in the image. Therefore, an image can be reduced to a real-valued vector with 12 features. The human experts also decided which training and test sets must be used. The training set contained 10 samples, while the test set had 36 samples. The intial results were evaluated comparing the result of this classification with the diagnosis

given by the biopsies. Currently the classification of benign and malignant was identified with accuracy range of 60-70%.

Conclusion

The paper provides the methodology with partial results of segmentation and explains theoretically how mammogram tumor classification is performed through case base reasoning method. First stage of mammogram mass segmentation result is shown in this paper, second stage is under implementation, so the conceptual framework of classification method is described on the paper. Info structure presented in this paper when successfully implemented would have an immense impact in the area of computer-aided diagnosis system. In future the methodology can be applied in a variety of medical image applications

Acknowledgments

I would like to thank School of Computer Science and Institute of Post graduate Studies, University Sains Malaysia for supporting to progress my research activities.

References

[1] Kohonen T, “Self Organization and Associative Memory”, Spring-Verlag, Hidelbarg, (1998)
[2] Hall EL, “Computer Image Processing and Recognition “, Academic Press, New York, (1978).
[3] Woods R.E, “Digital Image Processing”, Adisson Welsely, Reading, (1992).
[4] Kapur T, “Model based Three Dimensional Medical Image Segmentation”, MIT, (1992).
[5] Sheshadri HS, Kandaswamy A.,” Detection of breast cancer by mammogram image segmentation“. JCRT journal, Page no 232-234(2005).
[6] S.M. Lai, X. Li, and W.F. Bischof, On techniques for detecting circum- scribed masses in mammograms, IEEE Trans Med Imaging 8, 377– 386 (1989).
[7] H.D. Li, M. Kallergi, L.P. Clarke, V.K. Jain, and R.A. Clark, Markov random field for tumor detection in digital mammography, IEEE Trans Med Imaging 14, 565– 576, (1995).
[8] H. Kobatake, M. Murakami, H. Takeo, and S. Nawano, Computerized detection of malignant tumors on digital mammograms, IEEE Trans Med Imaging 18, 369–378, (1999).
[9] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans System Man Cybernet SMC-9, 62–66 (1979).
[10] Z.Pawalak, Rough Sets: Thoertical Aspects of Reasoning Data, Kluwer Academic Publication (1991).
[11] Jiang, Y., R.M. Nishikawa and J. Papaioannou, "Requirement of Microcalcification Detection for Computerized Classification of Malignant and Benign Clustered Microcalcifications" In Proceedings of the SPIE Conference on Image Processing, vol. 3338,pp. 313-317. San Diego (USA), (1998).
[12] Dhawan, A.P., and Y. Chitre,"Analysis of Mammographic Microcalcifications using Gray-level Image Structure Features" IEEE Transactions of Medical Imaging, (15) pp.246-259,(1996).
[13] Aamodt, A., and E. Plaza, “Case-based reasoning: Foundations issues, methodological variations, and system approaches”. AI Communications, 7: 39-59, (1994).
[14] Fajardo, L.J., and M.B. Williams ,” The Clinical Potential of Digital Mammography”, In Proceedings of the 3rd International Workshop on Digital Mammography, pp. 43-52. Chicago (USA), (1996).
[15] Raman Valliappan and Putra Sumari “Digital Mammogram Segmentation: An Initial Stage “in 4th IASTED International conference on Advanced Computing Science and Technology, Langawi, Malaysia, (2008).
[16] Mammogram Image Analysis Database, UK.

Valliappan Raman is currently doing his PhD in mammogram tumor segmentation and classification at University Sains Malaysia. He has completed his Masters Degree in 2005 at University Sains Malaysia and completed his Bachelor of Engineering in Computer Science in 2002 at Madurai Kamaraj University. He has been working as lecturer for past four years in well established university. He have undergone various research projects under major government grants and published papers, articles and journals. His research interest is in medical imaging, watermarking and health informatics.

Putra Sumari, is currently an Associate professor in School of Computer Science, University Sains Malaysia. He has undergone various research projects under government and university grants. He has supervised many postgraduate and undergraduate students. His research areas are in applied informatics, multimedia and image processing. He has published many papers in highly repuatated journal and conferences.

Mandava Rajeswari is currently an Associate Professor in School of Computer Science, University Sains Malaysia. His research areas are in computer vision and medical imaging.