Classification of Breast Masses in Mammograms using Support Vector Machine

G.Vaira Suganthi
Assistant Professor
Sethu Institute of Technology, Virudhunagar District, India.

J.Sutha
Head of the Department of Computer Science and Engineering
Sethu Institute of Technology, Virudhunagar District, India.

ABSTRACT
A Multi view CADx System for the mammography images is implemented. Two types of systems are widely used in mammography. They are Computer-aided detection (CADe) and Computer-aided diagnosis (CADx). The different views of mammography images MLO (Mediolateral Oblique) and CC (Crani-caudal) are assessed. Segmentation will be implemented for the images obtained from the two views for extracting the mass contour. A set of features related to the geometry of the boundary and the structure inside it will be computed for both of the images. An optimal subset of similar features will be extracted. Using the ranked features extracted the classification will be implemented using SVM. A Monte – Carlo method owing to the iterative and complex structure of the algorithms is used. The validation of the results is based on confidence intervals for given coverage probabilities and performance metrics.

General Terms
Image Processing

Keywords
Monte-Carlo method, Computer-aided detection, Support Vector Machine.

1. INTRODUCTION
The objective of this study is to investigate the use of pattern classification methods for distinguishing different types of breast tumors. For years, cancer has been one of the biggest threats to human life; it is expected to become the leading cause of death over the next few decades. Based on statistics from the World Health Organization (WHO), cancer accounted for 13% of all deaths in the world in 2004; deaths caused by cancer are expected to increase in the future, with an estimated 12 million people dying from cancer in 2030. Of all the known cancers, breast cancer is a major concern among women. It is the second-most common and leading cause of cancer deaths among women. According to published statistics, breast cancer has become a major health problem in both developed and developing countries over the past 50 years, and its incidence has increased in recent years.

At present, there are no effective ways to prevent breast cancer, because its cause remains unknown. However, efficient diagnosis of breast cancer in its early stages can give a woman a better chance of full recovery. Therefore, early detection of breast cancer can play an important role in reducing the associated morbidity and mortality rates. Computer-aided detection or diagnosis (CAD) systems, which use computer technologies to detect abnormalities in mammograms such as calcifications, masses, and architectural distortion, and the use of these results by radiologists for diagnosis, can play a key role in the early detection of breast cancer and help to reduce the death rate among women with breast cancer. Thus, in the past several years, CAD systems and related techniques have attracted the attention of both research scientists and radiologists. For research scientists, there are several interesting research topics in cancer detection and diagnosis systems, such as high-efficiency, high-accuracy lesion detection algorithms, including the detection of masses, detection of architectural distortion, and the detection of bilateral asymmetry. Radiologists, on the other hand, are attracted by the effectiveness of clinical applications of CAD systems.

One of the difficulties with mammography is that mammograms generally have low contrast. This makes it difficult for radiologists to interpret the results. Mammography is susceptible to a high rate of false positives as well as false negatives, causing a high proportion of women without cancer to undergo further clinical evaluation or breast biopsy, or miss the best time interval for the treatment of cancer. Several solutions have been proposed to increase the accuracy, specificity, and sensitivity of mammography and reduce unnecessary biopsies.

Double reading of mammograms has been advocated to reduce the proportion of missed cancers. The basic idea of double reading is to have two radiologists read the same mammograms. However, the workload and cost associated with double reading are high. Instead of double reading, CAD, which is referred to as the "second pair of eyes of the radiologists," With a CAD system, only one radiologist is needed to read each mammogram rather than two. The adoption of a CAD system could reduce the experts' workload. There are two types of examinations performed using mammography: screening mammography and diagnostic mammography. Screening mammography is performed to detect breast cancer in an asymptomatic population. Screening mammography generally consists of four views, with two views of each breast: the craniocaudal (CC) view and the mediolateral oblique (MLO) view.

This paper focuses on the development of a CAD system for the detection of masses that utilizes correspondence between MLO and CC views. Radiologists compare the two ipsilateral mammography views to decide whether or not a suspicious lesion is present. If a suspicious region in one view has certain features in common with a suspicious region in the other view, there is a higher probability that the region is a true lesion. The word "benign" means harmless. Benign tumors are not cancerous. Malignant tumors are also called breast

cancer. They can have irregular borders and they are made up of abnormally shaped cells. The CAD system uses computer based procedures to detect tumor blocks or lesions and classify the type of tumor.

2. RELATED WORK

In [1], the assessment of a CAD for the tumoral masses classification in mammograms by the uncertainty propagation through the system was performed. Based on the metrological characterization of the developed CAD, the features extraction, features selection, and classification steps were validated. In particular, suitable metrics such as the Receiving Operating Curve (ROC) and the Area under ROC (AUC) were widely used in order to provide a quantitative evaluation of the performance. Finally, a Monte Carlo simulation was implemented in order to provide the confidence interval for some coverage probabilities for all involved parameters.

In [2], an algorithm [12] already proposed by the authors was improved and assessed. The procedure succeeded in the case of very low contrast because it depended only on the orientation of the gradient vectors in the image but not on their amplitude. The mass detection procedure was carried out by performing two distinct stages. The procedure consisted of two separate steps: the detection of the suspicious regions and the final classification of those regions as masses or normal tissue. Starting from the original idea of the iris filter in a new method for the automatic identification of the masses in the mammographic images was proposed.

In [3], the assessment of a tumoral mass segmentation and characterization algorithm was performed by implementing the uncertainty propagation through the blocks. A Monte Carlo method owing to the iterative and very complex structure of the algorithms was used. The validation of the results was based on confidence intervals for given coverage probabilities and ad hoc performance metrics. A preliminary metrological validation of the mass segmentation algorithm was provided to extract shape and margins of tumoral masses with the aim of classify them as benign or malignant. The assessment was performed by uncertainty propagation through the whole system concerning both the segmentation step and the features extraction procedure.

In [4], uncertainty handling and propagation was considered by means of random fuzzy variables (RFVs) through a computer-aided-diagnosis (CADx) system for the early diagnosis of breast cancer. In particular, the denoising and the contrast enhancement of microcalcifications was specifically addressed, providing a novel methodology for separating the foreground and the background in the image to selectively process them. The whole system was then assessed by metrological aspects. It was assumed that the uncertainty associated to each pixel of the image has both a random and a non-negligible systematic contribution.

In [5], a data set of 10 digital mammograms containing benign tumors was presented to four radiologists for diagnosis in order to prove the variability between them. Then, several statistical features and their combinations were investigated in order to determine the best combination for diagnosis. It was found that a combination of the mean and median in a MATLAB algorithm was the best combination for mammographic benign tumor detection.

In [6], the ranges of feature extraction values for breast cancer mammography images were determined. After calculating the features for each mammogram image signs as cancer tumor, a decision about the frequency range value of breast cancer feature extraction was made. Geometric, Texture and Gradient features were analyzed.

In [7], a complete method for fast detection of circumscribed mass in mammograms employing an RBFNN (radial-basis-function neural networks) was presented. This method could distinguish between timorous and healthy tissue among various parenchymal tissue patterns. A decision was made to check whether a mammogram was normal or not and then detecting the masses position by performing sub-image windowing analysis.

3. ALGORITHM FOR BREAST MASSES CLASSIFICATION

Fig.1 Shows the various phases involved in the classification of Breast Masses in mammograms. The phases are briefly explained below.

3.1. Extraction of a Region of Interest (ROI)

The Region of Interest (ROI) is manually extracted from the original mammography image.

3.2. Segmentation

The segmentation step is needed to extract the mass contour. A region-growing algorithm which is a luminance-region based approach for the image segmentation is implemented to extract the mass contour. In particular, the segmentation step is divided into three main phases which are briefly described in the following list.

Artifact removing. An histogram-based approach is applied to the ROI to remove artifacts appearing as bright regular spots that can alter the segmented area.

Contrast enhancement by a nonlinear mapping of the pixel intensity is applied to improve local contrast on the ROI.

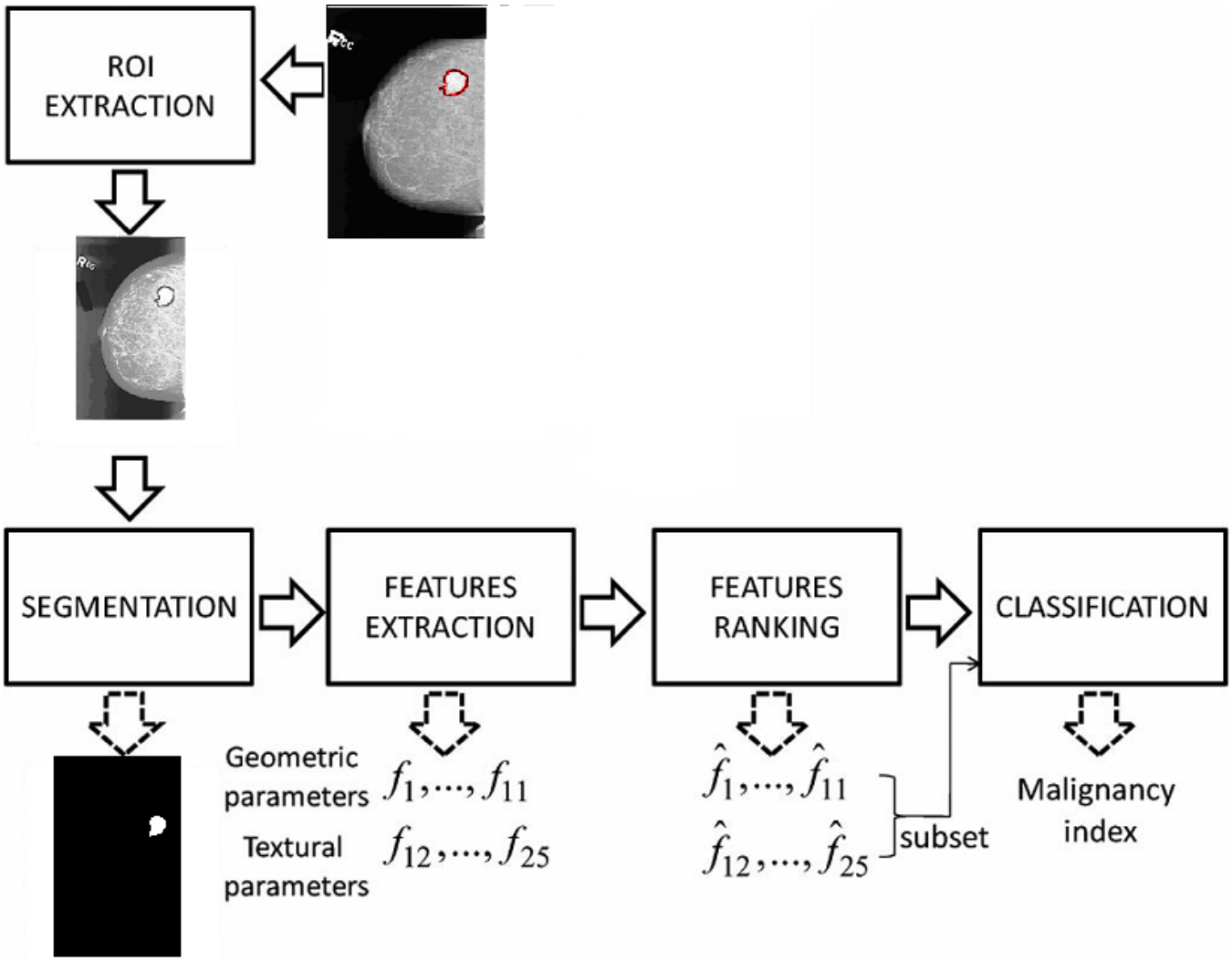

Figure 1. Phases in the classification of Breast masses

Region-growing. The algorithm starts from a seed pixel manually pointed inside the suspicious region, and then, expands the area around the seed to include nearby pixels falling within a threshold range. The seed is set to the average of the 15 × 15 neighboring pixels so that the algorithm is robust to noise. The region-growing algorithm is preliminarily applied to a decimated image to extract a coarse contour. The ROI is grid in regions of 3 × 3 pixels evaluating the average value of each square of the ROI. This value is assigned to the corresponding pixel in the decimated image. At this point, the region growing algorithm on this decimated image is applied to obtain a coarse contour. The contour in the original image is remapped, and reapplied the algorithm using the coarse contour as seeds to obtain a refined contour.

3.3. Features Extraction

Once the mass boundary is identified, a set of features related to the geometry of the boundary and the structure inside it is computed. The geometric parameters extracted are, area of the segmented mass; perimeter of the boundary of the segmented mass; statistical parameters of the radius of the segmented mass with respect to its centroid; circularity of the segmented mass boundary; eccentricity of the segmented mass boundary; rectangularity of the segmented mass boundary, boundary roughness of the segmented mass boundary related to the gradient of the radius; zero crossing of the segmented mass boundary. Table 1 illustrates the geometrical features extracted from 5 images.

Textural features considered are: entropy of the segmented mass;the sum of squares; the correlation; the contrast; kurtosis, dispersion, variance, skewness, mean, energy, entropy, correlation, homogenity. Table 2 illustrates the textural features extracted from 5 images. Textural features deal with the internal structure of the mass and are mainly based on the notion of co-occurrence matrix. Textural features are fundamental to correctly classify tumoral masses since the geometry of the boundary is often an ambiguous characteristic.

Table 1. Geometrical features extracted

Area	Perimeter	Eccentricity	Circularity	Rectangularity	Image
578	121.53	0.6657	2.034	0.0069	can1
712	112.67	0.62780	1.4195	0.00561	can2
15	20.48	0.77401	2.2274	0.26666	can3
766	182.16	0.76683	3.4492	0.00522	b1-cc
744	192.07	0.37759	3.9476	0.005376	b1-mlo

Table 2. Textural features extracted

Kurtosis	Dispersion	Variance	Skewness	Mean	Energy	Entropy	correlation	contrast	Homogenity	image
120.6449	0.016175354	0.008088	10.93823	0.008154	0.981922	5.77E-02	0.88835789	0.08898	0.998411077	can1
104.3072	0.01863807	0.009319	10.16402	0.009408	0.979534	6.33E-02	0.90783594	0.08464	0.998488573	can2
773.9236	0.002574256	0.001287	27.8015	0.001289	0.99613	1.62E-02	0.50421183	0.062887	0.998877019	can3
119.5042	0.016325973	0.008163	10.88596	0.008231	0.980213	6.46E-02	0.79401168	0.165703	0.997041017	b1-cc
79.55938	0.024224989	0.012113	8.863373	0.012263	0.971165	0.088242	0.81540322	0.22033	0.99606554	b1-mlo

3.4. Features Selection

The features selection is needed to extract an optimal subset of features for the classification. In this paper, the features selection is manually performed after ranking the features. In order to choose the selection criterion, it is verified that the normality of parameters is rejected in most of the cases. For this reason, we use a method with no assumptions about the data normality. To rank the features, the AUC for each feature is evaluated. First, to evaluate the ROC for each feature, we consider a threshold ranging from the minimum value of the feature to the maximum value of it, with a certain number of steps. Then, the feature is computed for every segmented mass. Obviously, according to the selected feature, the comparison between the feature value and the threshold can have a different result. For example, consider the circularity of a mass (f7). It is well known that high values of f7 are often related to benign mass while low values are often related to malignant cases. Thus, at every step, all the values of the circularity with the threshold are compared. Four cases can be encountered:

1) if the circularity is greater than the threshold for a benign mass then this case represents a True Negative (TN);

2) if the circularity is smaller than the threshold for a malignant mass then this case represents a True Positive (TP);

3) if the circularity is greater than the threshold for a malignant mass then this case represents a False Negative (FN);

4) if the circularity is smaller than the threshold for a benign mass then this case represents a False Positive (FP).

Obviously, since the threshold changes, also the assignment of TN, TP, FN, and FP to the cases changes at a different step. At every iteration, the sensitivitySE (also true positive rate), which is the probability of having a positive test among positive diagnosed patients is computed:

SE = TP ;TP + FN and the specificity SP (also true negative rate), the probability of having a negative test among negative diagnosed patients: SP = TN; FP + TN

The ROC curve has the sensitivity plotted along the vertical axis and the reversed scale of the specificity plotted on the horizontal axis. Then, the AUC is the area under the ROC curve. The greater the AUC value is, the higher is the position in the ranking of the feature under test. The same value for the circularity leads to a TP or to a FN according to the threshold value.

3.5. Breast Masses Classification

Using the ranked features extracted at the previous step, a SVM classifier is implemented. This method is useful in this study to provide preliminary results for the mass classification. The classification procedure allows us to assign the probability of being malignant (or of being benign) to each detected mass, starting from the values of the ranked features. The leave one out approach which selects one of the observations to hold out for the evaluation set while uses the remaining observations for the training set, was used as a cross-validation method. This procedure is then repeated for every observation of the data set. The validation of this step is performed by computing the AUC related to the classification result.

4. RESULTS AND DISCUSSION

4.1. Database

All experiments reported in this section were performed on a Pentium-Core2 2.66 GHz machine with 2GB of main memory, running Windows operating system. All algorithms were implemented in MATLAB version 7.9.0. The Figure.2 shows the result of applying the segmentation algorithm and the geometric features centroid & bounding box. In order to validate the proposed system, mammography images are taken from the DDSM database, with a pixel depth of 12bpp and a spatial resolution in the range [43–50] μm. The data set is represented by the Digital Database for Screening Mammography (DDSM) public database [13], a resource for use by the mammographic image analysis research community.

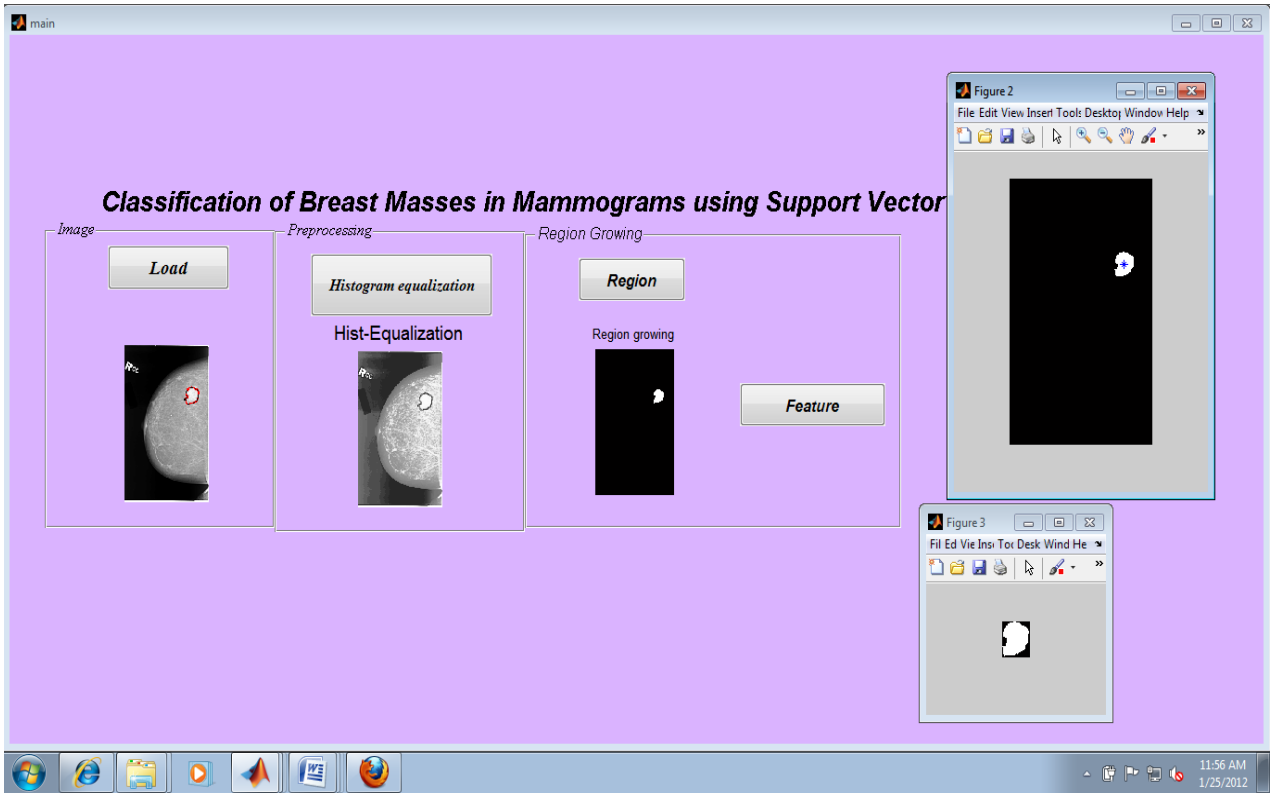

Figure 2. Region Growing Algorithm & Extraction of geometric features

4.2. Parameter Selection and Training

Once the training samples are obtained, the next step is to determine the optimal parametric settings of SVM. In this process, the following variables: the type of kernel function, its associated parameter, and the regularization parameter C must be decided. Various parameters for the SVM like regularization parameter C, degree of polynomial, sigma of RBF etc. are varied as: C from 1 to 1.05, degree of polynomial (p) from 2 to 9, and Gamma from 0.13 to 2.5 to choose the best parameters for SVM. It is found that polynomial kernel with degree (p) to be 3 and regularization parameter (C) having value 100 gives highest area of 0.958 under ROC curve (Az). For Gaussian RBF kernel it has been observed that gamma = 1.50 and C=10 gave highest area of 0.962 under ROC curve (Az). The SVM classifier is trained with the *training-set* using the optimal parameters of polynomial and gaussian RBF kernel respectively.

4.3. Performance Evaluation

As ROC analysis is commonly used approach for classification performance evaluation. Table I shows area under ROC curves (Az) of the proposed SVM approach with training and test data sets. In this study, cases actually benign and malignant are considered as true negative (TN) and true positive (TP) respectively. Table 3 shows that Az for Gaussian RBF kernel is more than that of polynomial kernel.

Table 3. Az values of ROC curves with different kernel of SVM

Kernel	TN	FN	FP	TP	Az	
					Training	Test
Polynomial	16	3	2	18	1.00	0.97
RBF	16	1	2	18	1.00	0.98

TN=True Negative, FN= False Negative, FP= False Positive, and TP= True Positive

5. ACKNOWLEDGMENT

Our sincere thanks to our colleagues and family members for the successful completion of the project work.

6. REFERENCES

[1] Arianna Mencattini, Marcello Salmeri., “Metrological Characterization of a CADx System for the Classification of Breast Masses in Mammograms”, IEEE

Transactions on Instrumentation and Measurement, Vol.59, No.11.Nov. 2010.

[2] Arianna Mencattini, Marcello Salmeri, "Assessment of a breast mass identification procedure using an Iris detector", IEEE Transactions on Instrumentation and Measurement, Vol.59, No.10, Oct.2010.

[3] Arianna Mencattini, Marcello Salmeri, et. al. "Uncertainty propagation for the assessment of tumoral masses segmentation", AMUEM 2009–International Workshop, July 2009.

[4] Arianna Mencattini, Marcello Salmeri, et. al. "Uncertainty modeling and propagation through RFVs for the assessment of CADx systems in Digital mammography", IEEE Transactions on Instrumentation and Measurement, Vol.59, No.1, Jan.2010.

[5] El-Sanosi M.D., Habbani.A.K., et. al. "Computer Aided Detection of Benign Tumors of the Female Breast", Proceedings of the 2008 IEEE, CIBEC'08.

[6] Hala Al-Shamlan, Ali El-Zaart, "Feature Extraction values for Breast cancer Mammography Images", 2010 International Conference on Bioinformatics and Biomedical Technology.

[7] Ioanna Christoyianni, Evalgelos Dermatas, et. al, "Fast detection of masses in Computer-Aided Mammography", IEEE Signal Processing Magazie, Jan.2000.

[8] Jinshan Tang, Rangaraj M.Rangayyan, et. al. "Computer Aided Detection and Diagnosis of Breast Cancer With Mammography:Recent Advances", IEEE Transactions on Biomedicine, Vol.13, No.2, March 2009.

[9] Maurice Samulski, Nico Karssemeijer., "Optimizing Case-Based Detection Performance in a Multiview CAD System for Mammography" IEEE Transactions on Medical Imaging, Vol.30, No.4. Apr. 2011.

[10] Nalini Singh, Ambarish G.Mohapatra, "Breast Cancer Mass Detection in Mammograms using K-means and Fuzzy C-means Clustering", International Journal of Computer Applications, Vol.22, No.2, May 2011.

[11] Sukhwinder Singh, Vinod Kumar, "SVM Based System for Classification of Microcalcifications in Digital Mammograms",EMBS Annual International Conference, Sep. 2006.

[12] H. Kobatake and M. Murakami, "Adaptive filter to detect rounded convex regions: Iris filter," in Proc. IEEE Int. Conf. Pattern Recog., 1996, vol. 2, pp. 340–344.

[13] University of South Florida Digital Mammography Home Page, Univ. South Florida, 2000. [Online] Available: http://marathon.csee.usf.edu/Mammography/Database.html