International Journal of Innovative Computing, Information and Control
Volume **9**, Number **6**, June **2013**

ICIC International ©2013 ISSN 1349-4198
pp. 2651–2664

COMPUTER AIDED DETECTION OF RADIO OPAQUE LESIONS IN DIGITIZED MAMMOGRAMS THROUGH DISTRIBUTION FREE KNOWLEDGE BASED CLASSIFICATION TECHNIQUES

W. Jai Singh[1], B. Nagarajan[2] and S. Arun Kumar[3]

[1]Department of Master of Computer Applications
Park College of Engineering and Technology
Coimbatore, Tamil Nadu, India
jaisingh_w@yahoo.com

[2]Department of Computer Applications
Bannari Amman Institute of Technology
Sathyamangalam, Tamil Nadu, India

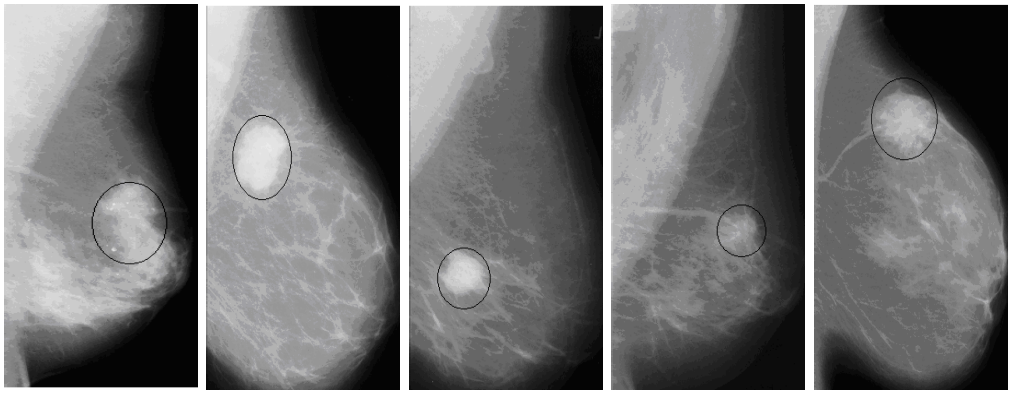
[3]Kovai Medical Center and Hospital
Coimbatore, Tamil Nadu, India

Received January 2012; revised June 2012

Abstract. *Segmentation of Lesions is a vital step in computerized mass detection scheme for digitized mammograms. In this paper, two robust algorithms have been developed for the segmentation of Radio Opaque Lesions known to be K-Means Bootstrap Subgroup (KMBS) and Expectation Maximization Bootstrap Subgroup (EMBS). The proposed algorithms are capable of segmenting the regions of varying intensity distribution in a mammogram. A number of image regions being 8 and 5, it yields True Positive (TP) rate of 94.5% and 92.3% for EMBS with false positive per Image of 0.26 and 0.33 respectively. When KMBS method is applied for the same data set, it results in TP rate of 93.4% and 91.3% with false positive per image of 0.33 and 0.37. The regions of Radio Opaque Lesions are segmented and the assessments of the segmentation results by radiologist are compared. The efficiency of algorithm is measured using Free Receiver Operating Characteristics (FROC) curve and the results are highlighted.*
Keywords: Bootstrap subgroup, Expectation-Maximization (EM), K-Means, Radio opaque lesions, Mammography, Segmentation

1. **Introduction.** Most cancer cells eventually form a lump or mass called tumor, which is named after the part of the body where the tumor originates. For instance, breast cancer begins in breast tissue, which is made up of glands for milk production, called lobules, and the ducts that connect lobules to the nipple. Breast cancer includes some different types, such as clustered microcalcifications, Speculated lesions, Circumscribed masses, Ill defined masses and Architectural distortions. It is to be noted that, 97% of deaths occurred due to breast cancer in women aged 40 and above [1]. The National Cancer Institute estimated that approximately 2.6 million US women with a history of breast cancer were alive in January 2008, and that more than half were diagnosed less than 10 years earlier. Most of these individuals were cancer-free, while others still had evidence of cancer and may have been undergoing treatment [2]. Mammography is a low-dose X-ray procedure that allows visualization of the internal structure of the breast. Mammography procedure detects about 80%-90% of the breast cancers in women without any symptoms. Figure 1 shows the various classes of breast cancer prevail among women.

The main objective of this paper is to demonstrate the potential of distribution free technique to segment suspicious regions in mammographic images and proposing a methodology that includes an innovative use of hybrid features and knowledge based classifier



(a) (b) (c) (d) (e)

FIGURE 1. Different classes of breast cancer which is marked by circles: (a) clustered microcalcifications, (b) speculated lesion, (c) circumscribed mass, (d) ill defined masses, (e) architectural distortion

for detection of lesions. The quality of the results obtained by this work can allow this methodology to be added to a computer tool for the medical area, providing support to specialists especially in cases in which visualization is difficult.

Breast parenchyma density pattern is an important factor in the risk and development of breast cancer [3]. N. Karssemeijer has developed a model for automated determining the parenchymal patterns in mammograms. This study exhibits the relation between breast cancer risk and changes in mammographic density [4]. S. R. Aylward et al. have developed a model on breast density estimation for breast cancer risk assessment. The model accuracy was quantified by comparing the estimates with semi-automated estimates generated by experts [5]. To some extent, different tissue regions in mammograms such as fat, parenchymal tissues, and masses can be distinguished based on the pixel gray levels [6]. However, conventional segmentation methods often fail when applied to digital mammograms, because of low contrast, complicated structured background, and variation in breast tissue contrast due to acquisition differences [7]. Therefore, to improve the segmentation result, distribution free segmentation algorithms have been developed to detect Radio Opaque Lesions in digitized mammograms.

In recent years, many methods for automated digital mammography processing have been published, and among them Expectation Maximization (EM) technique is the most popular one [8]. M. A. Kupinski and M. L. Giger suggested probabilistic mammogram model using a small-variance Gaussian function in the lesion region [9]. H. Li et al. have used a finite generalized gaussian mixture to model the histogram of a digitized, preprocessed mammogram; the parameters of the model were estimated by the EM algorithm [10]. T. Lei and W. Sewchand have developed a model X-ray Computed Tomography (CT) images as finite normal mixture in which parameters were estimated by the EM and Classification Maximization (CM) algorithms [11].

T. Kanungo et al. [12] have developed a simple and efficient implementation of Lioyd's K-means clustering algorithm for image analysis. A. K. C. Wong et al. [13] have proposed a method called discrete values clustering algorithm with applications to bio-molecular data. C. W. Chen et al. [14] have developed a robust segmentation algorithm for three dimensional image data based on a novel combination of adaptive K-means clustering and knowledge based morphological operations. H. P. Ng et al. [15] have developed a model for medical image segmentation using K-means clustering and improved watershed algorithm.

In mammogram, there are three main lesion features called texture, shape and gray level. In recent years, several lesion feature based schemes for mass detection and segmentation have been developed. A. Cao et al. [16] have proposed a method called robust information clustering incorporating texture features for detection of lesions in mammograms. H. Kobatake et al. [17] have used iris filter method to segment the suspicious regions and shape features are used to detect the malignant tumor. X. P. Zhang and M. D. Desai [18] have developed a systematic method for the detection and segmentation of bright targets by using gray level features. A. K. Santra et al. [19] have developed a new algorithm Pixcals to identify and detect microcalcifications. A. K. Santra et al. [20] have developed a new algorithm, Pixcals refined bandwidth algorithm to identify and detect microcalcifications. W. JaiSingh and B. Nagarajan [21] have developed a new novel approach for detection of lesions in digitized mammogram using hybrid features with learning classifier. R. Sammouda et al. [22] have developed a method for the detection of lung cancer using bit-plane slicing and modified version of hopfield neural network. M. S. Mohamad et al. [23] have developed a cyclic hybrid method to select a smaller subset of informative genes for cancer classification.

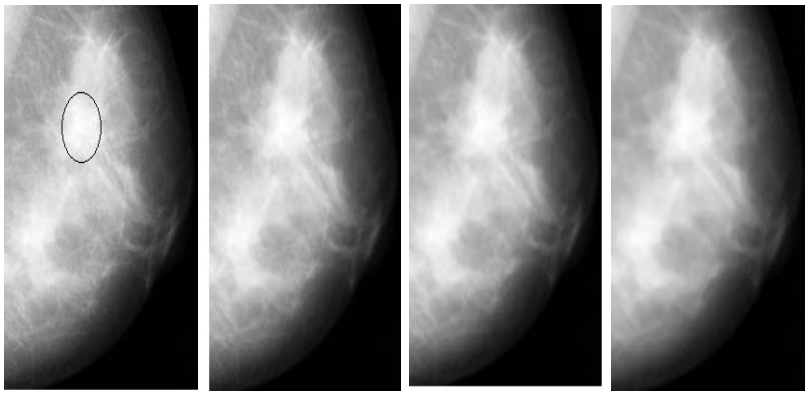
The present work aims at developing new segmentation algorithms for the detection of Radio Opaque lesions. In this paper, at the initial stage, median filter is applied to remove the noise, and unsharp masking techniques are used to enhance the quality of the mammograms. Second, the combination of Expectation Maximization Bootstrap Subgroup (EMBS) and K-Means Bootstrap Subgroup (KMBS) algorithms is employed to detect suspicious pixels. In Bootstrap subgroup technique, the pixel values are considered as a universal population. Therefore, each pixel is taken into consideration to detect suspicious pixels. K-means clustering is a method of cluster analysis, which aims at partition of 'n' observations into k clusters. Each observation belongs to the cluster with the nearest mean. EM algorithm and K-means clustering method both aim at finding the centers of natural clusters in the data through iterative refinement approach. Therefore, we have developed EMBS and KMBS algorithms to detect suspicious pixels. Third, binary morphological operators and eight-connected component labeling methods are employed to reconstruct the shape, to remove isolated pixels and to segment the suspicious regions. Fourth, hybrid features are extracted from the segmented regions and finally, a support vector machine (SVM) classifier is used to pinpoint the lesions. The experimental outcome shows better results when compared with classical EM and K-Means.

This paper is organized in a sequential form. In Section 2, the Image preprocessing and enhancement techniques are briefly explained. In Section 3, details of proposed techniques called K-Means Bootstrap Subgroup (KMBS) and Expectation Maximization Bootstrap Subgroup (EMBS) are portrayed. The experimental results and performance of the proposed approach are given in Section 4. Finally, summary, comparison and conclusion are given in Section 5.

2. **Preprocessing and Enhancement.**

2.1. **Median filtering.**

Median filtering is found to be very powerful in removing noise from two-dimensional signals without blurring edges. This makes it particularly suitable for enhancing mammogram images [24]. A median filter is a nonlinear spatial filter that replaces the value of a given pixel with the median pixel value within a region of interest. A median filter with properly chosen support can smooth the noise in the original image [25]. On the other hand, the median filter may also virtually eliminate the Lesions from the original-image. Thus, there will be a tradeoff between noise removal and the preservation of signals from Lesions. Figure 2 illustrates different support regions applied



(a) (b) (c) (d)

FIGURE 2. Different support region applied on the mammogram: (a) original image from MIAS database (mdb102), (b) resulting image – 5×5 median filter, (c) resulting image – 7×7 median filter, (d) resulting image – 11×11 median filter

in the mammogram *mdb102*. In Figures 2(b)-2(d), portray the feature images produced on original image when a median filter with a support region of size 5×5, 7×7 and 11×11 respectively. Clearly, by increasing the size of the support region both in noise signals and signals from Lesions are being suppressed. For the 7×7 and 11×11 cases, it eliminates more of the distinct Lesions than the 5×5. Thus, the 5×5 support region is best for noise removal and preservation of signals from Lesions.

2.2. **Image enhancement.** Enhancement is aimed at realizing improvement in the quality of a given image [26]. It can be accomplished by enhancing contrast and edges. Applying Unsharp Masking Filters, contrast enhancement is improved and the readability of areas with subtle changes in contrast is achieved. Many image enhancement techniques are based on spatial operations performed on local neighbourhoods of input pixels [27,28]. Often, the image is convolved with a finite impulse response filter called spatial mask.

The unsharp masking filter is a simple sharpening operator which derives its name from the fact that it enhances edges (and other high frequency components in an image) via a procedure that subtracts an unsharp, or smoothed, version of an image from the original image.

Unsharp masking produces an edge image $g(x, y)$ from an input image $f(x, y)$ via

$$g(x, y) = f(x, y) - f_{smooth}(x, y) \tag{1}$$

where f_{smooth} is a smoothed version of $f(x, y)$.

This edge can be used for sharpening if one adds it back into the original signal. The enhanced image $f_{sharp}(x, y)$ is obtained from the input image $f(x, y)$ as

$$f_{sharp}(x, y) = f(x, y) + \lambda g(x, y) \tag{2}$$

where λ controls the shape of the laplacian and must be in the range 0.0 to 1.0 and $g(x, y)$ is suitably defined gradient at (x, y). A commonly used gradient function is the discrete Laplacian.

$$\begin{aligned} g(x, y) \triangleq f(x, y) - \frac{1}{4}[f(x-1, y) + f(x, y-1) \\ + f(x+1, y) + f(x, y-1)] \end{aligned} \tag{3}$$

3. **Detection of Radio Opaque Lesions Using Distribution Free Knowledge Based Classification Techniques.** Once an image has been pre-processed, the enhanced images contain the Lesions, which are the brightest and they may exist within regions of high average gray levels. Hence, it is difficult to segment the required region in a reliable manner. Therefore, each pixel is taken into account to detect Lesions. Many authors have assumed various types of distributions to assess the pixels to segment suspicious region. For example, Z. Liang et al. [8] have assumed the mixture of three Gaussian distribution to image intensities with each type of tissues. M. A. Kupinski and M. L. Giger [9] have assumed the mixture of two normal distribution for automated seeded lesion segmentation on digitized mammograms. H. Li et al. [10] have used the finite generalized Gaussian mixture to segment the lesions. T. Lei and W. Sewchand [11] utilizes the finite Gaussian random field as the stochastic image model for segmentation purpose. However, the assumption of distributions is not applicable in general. It may lead to wrong diagnosis if the assumption fails. Hence, a distribution free knowledge based classification algorithms have been developed to detect Radio Opaque Lesions.

The Bootstrap, originally proposed and named by B. Efron (1979), is a computational technique that can be used effectively to estimate the sampling distribution of Statistics [29]. In particular, one can use the non-parametric Bootstrap to estimate the sampling distribution of a statistics, while assuming only that the sample is representative of the population from which it is drawn. The observations are independent and identically distributed. In its simplest form, the nonparametric Bootstrap does not rely on any distributional assumptions about the underlying population. However, it gives equal chances to all pixels in the image to be selected. Let $x = \{X_1, X_2, \ldots, X_n\}$ be a sample, i.e., a collection of n number of pixels drawn at random from a completely unspecified distribution, F.

$$F_n(x) = \frac{1}{n} \cdot (\text{Number of } X_i \leq x) \tag{4}$$

T. Seppala et al. (1995) have proposed a technique called the subgroup bootstrap chart [30] that is used to monitor the process or system. The bootstrap subgroup model is given below.

$$X_{ij} = \mu_i + \varepsilon_{ij} \tag{5}$$

where $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n$. Here, μ_i is the true mean of the i^{th} subgroup, and ε_{ij} is a random error term. Distribution free knowledge based classification algorithms KMBS and EMBS are given below.

3.1. **K-Means Bootstrap Subgroup (KMBS) algorithm.** Image segmentation remains as one of the major challenges in image analysis. Clustering algorithms are useful in segmentation problem. In this paper, K-means clustering algorithm [14] is used with a primary segmentation of the image which is an unsupervised method. The clustering algorithm is used in selecting an initial pixel or region that belongs to an object of interest, followed by an interactive process of neighborhood analysis. It decides whether each neighboring pixel belongs or does not belong to the same object. K-means clustering is most suitable for biomedical image segmentation since the numbers of clusters (K) which are usually known as images of particular regions form subgroups. The K-means algorithm process the data, which form features of vector space and finds natural clustering. In this paper, the model with grayscale image forming regions of varying clusters is being developed. The number of subgroup clusters (or tissue types) is k. The K-Means Bootstrap Subgroup can be used to monitor the mean or standard deviation of k clusters after sub grouping the data. In the present context, the implementation of the K-Means Bootstrap Subgroup (KMBS) algorithm is as follows.

Step 1: Read the mammogram image and store it in a two dimensional matrix.
Step 2: Apply the preprocessing technique to remove noise, to enhance contrast and to enhance edges.
Step 3: Implement the K means algorithm to segment the mammogram into K subgroups.

Given a set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional real vector, then k-means clustering aims to partition the n observations into k sets $(k < n)$ $S = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the Within-Cluster Sum of Squares (WCSS):

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - m_i||^2 \tag{6}$$

Given an initial set of k means $m_1^{(1)}, \ldots, m_k^{(1)}$, which may be specified randomly or by some heuristic, the algorithm proceeds alternatively between two steps:

Assignment step: Assign each observation to the cluster with the closest mean.

$$S_i^{(t)} = \left\{ x_j : ||x_j - m_i^{(t)}|| \leq ||x_j - m_{i^*}^{(t)}|| \text{ for all } i^* = 1, \ldots, k \right\} \tag{7}$$

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

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \tag{8}$$

The algorithm is deemed to have converged when the assignments no longer change.
Step 4: Observe k subgroups (clusters) of size n for total of n time's k observations.
Step 5: Compute $e_{ij} = x_{ij} - \bar{x}_i$, for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n$, where $\bar{x}_i$ is the sample mean of the i^{th} subgroup.
Step 6: Compute $x_{ij}^* = \bar{x}_i + c.e_{ij}$ for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n$. Here, $c = \sqrt{n/(n-1)}$ is a correction factor used to adjust the variance of the subgroups.
Step 7: Compute the sample mean, $\bar{x}^*$, from $x_1^*, x_2^*, \ldots, x_k^*$. This sample is a bootstrap sample.
Step 8: Sort the k estimated bootstrap sample, $\bar{x}_1^*, \bar{x}_2^*, \ldots, \bar{x}_k^*$.
Step 9: Find the smallest ordered x^* such that $(1 - (\alpha/2)).K$ values below it. This is the Bootstrap Upper Bandwidth Limit (BUBL).

Here, α is the desired false alarm rate. It must be in the range of $0 \leq \alpha \leq 1$.
Step 10: Segment the Region of Interest (ROI) based on the threshold value BUBL. The threshold image $R(x, y)$ is defined as

$$R(x, y) = \begin{cases} 1 \text{ if } f_{sharp}(x, y) > \text{BUBL} \\ 0 \text{ if } f_{sharp}(x, y) \leq \text{BUBL} \end{cases} \tag{9}$$

Step 11: The resulting image $R(x, y)$ which contains the white pixels is termed as suspicious pixels.
Step 12: Binary morphological operators are employed in the resulting image $R(x, y)$ to reconstruct the shapes of the suspicious region and to remove the isolated pixels.

3.2. **Expectation-Maximization Bootstrap Subgroup algorithm (EMBS).** The EM algorithm is widely used for the estimation of model parameters in medical Images. Statistical models are used to represent appropriate system in the image data. The selection of pixels is stochastic in nature. The EM Bootstrap Subgroup Algorithm (EMBS) is proposed to estimate the parameters of distribution from the samples of a known image. The method first classifies the tissue types or estimates the class parameters associated

with the tissue types. The density of a random variable $X^* = (x_1^*, x_2^*, \ldots, x_n^*)$ is defined as:

$$f(x^*) = \sum_{k=1}^{K} p_k f_k(x^*|\theta_k) \tag{10}$$

where,
K is the number of classes that need to be segmented from the image.
$\theta_k = (\mu_k, \sigma_k^2)^t$, $t \to t^{\text{th}}$ iteration.
$f_k(x^*|\theta_k)$ is a Bootstrap density with mean μ_k and variance σ_k^2.
$(p_1, p_2, \ldots, p_k)$ is a vector of mixture probabilities such that $k = 1, 2, \ldots, K$ and $\sum_{k=1}^{K} p_k = 1$.

Once the estimated values of the parameters are determined, then the class of corresponding pixel in the image can be easily classified. By computing the probability of the pixels, one can determine the belonging of a pixel to a particular class or subgroup. The Expectation Maximization Bootstrap Subgroup can be used to estimate the mean or standard deviation of each class or sub groups. In the present context, the implementation of the algorithm is as follows.
Step 1: Read the mammogram image and store it in a two dimensional matrix.
Step 2: Apply the preprocessing techniques to remove noise, to enhance contrast and to enhance edges.
Step 3: Employ EM algorithm to estimate parameters of the simulated Bootstrap samples.

(i) *Initialization*: The first step is the initialization of a priori probabilities, averages and variances of every class. With the help of the image histogram, initialize the parameters in the following way:

$$\hat{p}_k^0 = \frac{\hat{n}_k^0}{n}; \quad \hat{\theta}_k^0 = (\mu_k^0, \sigma_k^0) \tag{11}$$

where $\hat{n}_k^0$ is the total number of observations in the class k.

(ii) *Expectation*: In this step we estimate a posterior probability $\hat{p}_k^m(x_i^*)$ for the pixel x_i^* that belongs to the class k at the m^{th} iteration by:

$$\hat{p}_k^m(x_i^*) = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_k^m \tag{12}$$

(iii) *Maximization*: at the $(m+1)$th iteration we can estimate a priori probability $\hat{p}_k^{m+1}$, the mean $\hat{\mu}_k^{m+1}$ and the variance to each class by

$$\hat{p}_k^{m+1} = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_k^m(x_i^*) \tag{13}$$

$$\hat{\mu}_k^{m+1} = \frac{\sum_{i=1}^{n} x_i^* \hat{p}_k^m(x_i^*)}{\sum_{i=1}^{n} \hat{p}_k^m(x_i^*)} \tag{14}$$

$$(\hat{\sigma}_k^{m+1})^2 = \frac{\sum_{i=1}^{n} (x_i^* - \hat{\mu}_k^{m+1})^2 \hat{p}_k^m(x_i^*)}{\sum_{i=1}^{n} \hat{p}_k^m(x_i^*)} \tag{15}$$

(iv) Abort the algorithm when the following condition is satisfied.

$$\left| \hat{p}_k^{m+1} - \hat{p}_k^m \right| < \varepsilon$$

where ε is a very small number.

(v) Classification: The Bayesian Rule (BR). After the mixture identification, the BR is applied in order to classify the pixels according to their gray level x^*:

$$K(x^*) = \arg\left[\max_{1\leq k\leq K}\left\{p_k f(x^*|\theta_k)\right\}\right] \tag{16}$$

where $K(x^*)$ represents the label of the class of the pixel x^*.
Step 4: Observe k subgroups (clusters) of size n for total of n times k observations.
Step 5: Compute $e_{ij} = x_{ij} - \bar{x}_i$, for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n$, where $\bar{x}_i$ is the sample mean of the i^{th} subgroup.
Step 6: Compute $x^*_{ij} = \bar{x}_i + c.e_{ij}$ for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n$. Here, $c = \sqrt{n/(n-1)}$ is a correction factor used to adjust the variance of the subgroups.
Step 7: Compute the sample mean, $\bar{x}^*$, from $x^*_1, x^*_2, \ldots, x^*_k$. This sample is a bootstrap sample.
Step 8: Sort the k estimated bootstrap sample, $\bar{x}^*_1, \bar{x}^*_2, \ldots, \bar{x}^*_k$.
Step 9: Find the smallest ordered x^* such that $(1-(\alpha/2)).K$ values below it. This is the Bootstrap Upper Bandwidth Limit (BUBL).

Here, α is the desired false alarm rate. It must be in the range of $0 \leq \alpha \leq 1$.
Step 10: Segment the Region of Interest (ROI) based on the threshold value BUBL. The threshold image $R(x, y)$ is defined as

$$R(x,y) = \begin{cases} 1 \text{ if } f_{sharp}(x,y) > \text{BUBL} \\ 0 \text{ if } f_{sharp}(x,y) \leq \text{BUBL} \end{cases} \tag{17}$$

Step 11: The resulting image $R(x, y)$ which contains the white pixels is termed as suspicious pixels.
Step 12: Binary morphological operators are employed in the resulting image $R(x, y)$ to reconstruct the shapes of the suspicious region and to remove the isolated pixels.

3.3. **Segmentation of suspicious region.** A Lesion is an abnormality or alteration in the tissue's integrity. Breast lesions usually come in the form of lumps or swellings in or around the breast. For correct interpretation, image must be partitioned into regions that correspond to objects or parts of an object. After applying the mathematical morphological operators, the binary image will be segmented into different regions by using 8-connected component labeling method. A set of pixels in a binary image that form a connected group is known to be the suspicious region. The features of each region will be generated in the classification step for lesion detection.

3.4. **Hybrid feature extraction and classification.** To locate the regions that are suspicious of tumors, certain features that can be used in the CAD system. A set of 21 hybrid features is calculated for each suspicious region. These features fall into three categories related with the texture, shape and gray level properties of each region. The hybrid feature groups are presented in Table 1. Based on the feature extraction, Support Vector Machine (SVM) was used as classifier to further classify the suspicious region into lesion or normal. SVM is a machine learning method for creating a classification function from a set of labeled training data. Classification of breast abnormality has been performed as a two class problem where the two classes are lesion and normal.

4. Experimental Results and FROC Analysis.

4.1. **Results.** The new two methods KMBS and EMBS have been developed to segment the Radio Opaque Lesions in mammogram. The experiments were conducted on digitized mammograms with a spatial resolution of 200 μm from the Mini-MIAS database. They were clipped or padded, so that every image is of the size 1024 $\times$ 1024 pixels, comprising

TABLE 1. Hybrid features

Category	Features
Texture	Mean Gradient, Variance of gradient, Inertia, Correlation, Energy, Homogeneity, Entropy
Shape	Area of the lesion, Perimeter, Euler number, Orientation, Extent of Lesion, Convex Area of the lesion, Solidity, Lesion Eccentricity, Lesion's Equivalent Diameter
Gray Level	Mean Intensity, Variance Intensity, Standard Deviation of Intensity, Min Intensity, Max Intensity

tumor cases of 89 images and 75 non tumor cases. The test database comprised 19 Speculated Masses (SPIC), 22 Circumscribed Masses (CIRC), 14 Ill-defined Masses (MISC), 15 cases with Asymmetry (ASYM), 19 cases with Architectural Distortion (ARCH), and 75 normal cases. To evaluate the computer aided diagnosis results, the findings of the proposed method is considered as correct result, if its area is overlapped by at least 50% of a true lesion. The detection results are evaluated by terms of sensitivity and the number of False Positives per Image (FP/I). The CAD system was developed by using MATLAB 7.0.

The KMBS and EMBS algorithm was implemented by selecting the number of image regions $k = 8$. This was premised on the determination of the k value for digital mammograms by H. Li et al. using model selection [10]. Later, the experiments were repeated by selecting the number of major components of a mammogram with $k = 5$, based on the model proposed by S. R. Aylward et al. [5], which took into account the background and the breast tissue classes, uncompressed fat, fat, dense tissue, and muscle.

According to literature survey, for the CIRC lesions, their shapes are usually close to round or oval, and the morphological filter is an effective method to enhance the segmentation capability for these objects. In MISC shape mammograms, the centers of most of the lesions have higher gray-level values than background regions. The shapes of ASYM usually do not have a significant feature of texture or shape but gray level is a more representative feature. Since the two shapes of lesions have relatively large pixel values, the gray level features are obvious. Texture is the most used feature for SPIC and ARCH type lesions. Therefore, hybrid features are taken into account by the proposed methods to obtain high accuracy in computer aided diagnosis. We have used SVMs as a potential mechanism for the design of a classifier responsible for delineating between lesion and normal region in mammogram. There were a number of motivations for selecting SVMs as a classification mechanism. SVMs have been shown to perform well in medical diagnosis applications and have also been shown to perform well when dealing with relatively small training sets [31]. The hybrid features described in Section 3.4 were computed and SVM classifier is used to identify lesion and normal regions.

The results obtained from the proposed and existing methods were evaluated by a radiologist, an expert in mammography analysis. The original and resulting images were simultaneously presented on a computer monitor for subjective assessment. By visual comparison, the radiologist assigned one of four ranking options to the segmentation results. According to the radiologist's rating, 89 tumor cases which contain 92 lesions were analyzed. Out of which 87 and 85 were categorized as either excellent or good for EMBS, 86 and 84 for KMBS with $k = 8$ and $k = 5$ respectively as listed in Table 2. The results of the experiments conducted indicate that the proposed approaches significantly

TABLE 2. Radiologist evaluation for proposed methods

Radiologist Rating	Proposed Methods			
	EMBS		KMBS	
	$k = 5$	$k = 8$	$k = 5$	$k = 8$
Excellent	43	47	41	44
Good	42	40	43	42
Average	4	3	5	4
Poor	3	2	3	2

reduce the error in segmenting the cancer region, compared with the conventional EM and K means method, irrespective of the choice of k. According to the radiologist report, fixing of $k = 8$ helps the system to get the result accurately. From Table 2, the choice $k = 5$ is also meaningful, as each distribution corresponds to a specific tissue type in the breast.

Figure 3 is used to demonstrate the robustness of the proposed approach. The original MIAS mammogram Circumscribed mass (mdb028), Ill-defined mass (mdb134), Asymmetry (mdb072), Architectural distortion (mdb150), Speculated mass (mdb184) are given in Figure 3(a). In the EM and K-means algorithm, it eliminates one or more regions in the context mammogram. This difficulty is usually encountered in specific cases that have relatively small sized labeled region(s) after the initial segmentation is performed. By violating the initial assumption on the number of meaningful regions contained in the mammogram, there are possibilities for missing the region(s), which enclose the details of radiological importance. Figures 3(b) and 3(c) show the initial segmentation of the mammogram mdb028, mdb134, mdb072, mdb150 and mdb184 for $k = 8$. From Figures 3(b) and 3(c), the tumor candidate completely disappeared in one of the bright region. Interestingly, Figures 3(d) and 3(e) demonstrate how the KMBS and EMBS procedure preserved the brightest class representing the tumor (indicated in white), in which the size confirmed to the detail provided in the MIAS database and by the radiologist.

Comparisons of proposed methods with that of A. Cao et al. [16] and K. Hu et al. [32] are made. A. Cao et al. proposed a detection method, which has been verified with 60 mammograms in mini MIAS database and achieved a TPR of 90.7%. In K. Hu et al. method with 170 mammograms one can achieve a TPR of 91.3%. Our proposed methods have been verified with 164 mammograms, achieved a TPR of 93.4% and 94.5% for KMBS and EMBS respectively. Table 3 shows the comparison of detection rates between the ground truth database, previous works of A. Cao et al. and K. Hu et al. and the proposed methods.

TABLE 3. Comparison of detection rates

Authors and References	Type of Methods	Ground Truth database Amount of lesions/Images	Detection Result	Detection Rates (TPR)
A. Cao et al. [16]	Robust Information Clustering	54/54	49	90.7%
K. Hu et al. [32]	Adaptive Thresholding based on Multiresolution Analysis	92/89	84	91.3%
Proposed Methods	KMBS	92/89	86	93.4%
	EMBS	92/89	87	94.5%

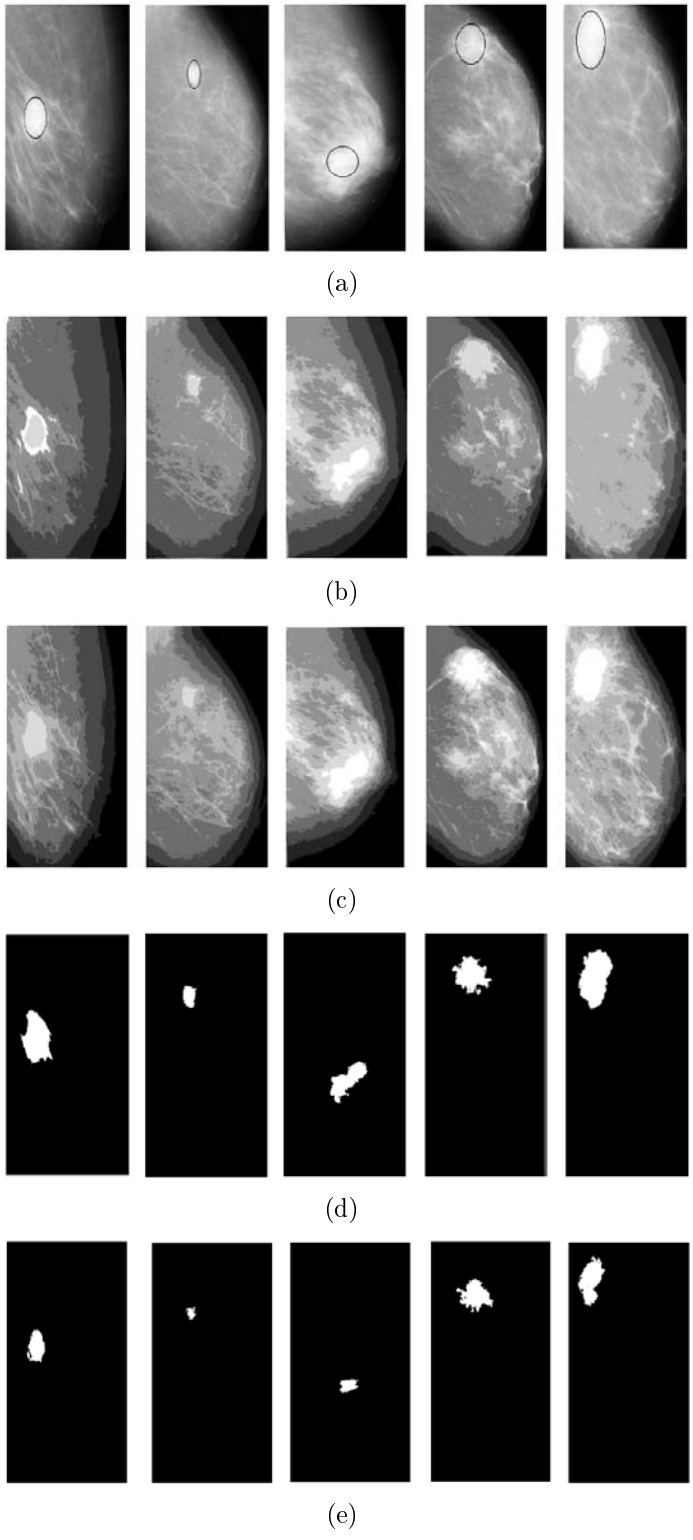


FIGURE 3. Results obtained by various approach with MIAS mammogram: (a) Original MIAS mammogram mdb028, mdb134, mdb072, mdb150 and mdb184. The circled region contains lesion; (b) and (c) Section of the segmented mammogram after applying the Expectation Maximization (EM) and K-Means algorithm with $k = 8$. The pixels corresponding to the relatively small sized brightest region constitute the tumor candidate; (d) and (e) segmentation of tumor after applying KMBS and EMBS. The method EMBS preserves the actual size of the tumor as specified in the MIAS database.

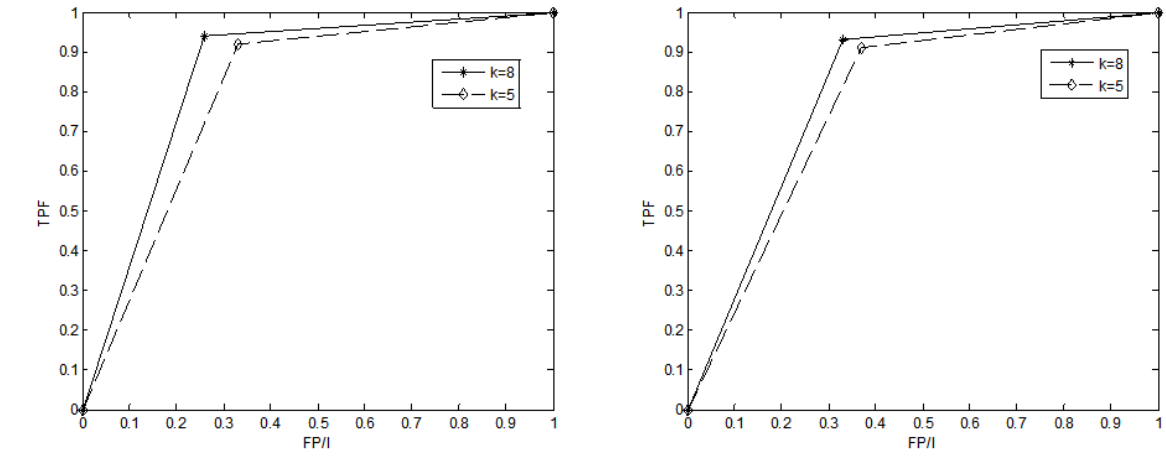


FIGURE 4. FROC curves for the proposed systems with $k = 8$ and $k = 5$: (a) EMBS, (b) KMBS

4.2. **Free Receiver Operating Characteristics (FROC) curve analysis.** Diagnostic tests have particular importance in medicine, where early and accurate diagnosis can decrease the morbidity and mortality of disease. For many years, diagnostic performance was reported by the accuracy of test. To evaluate, the performance of proposed system, Free Receiver Operating Characteristic (FROC) analysis is performed.

Figure 4 shows the FROC curve of the proposed detection method. In the FROC curve, x-axis represents the number of FP/I, and y-axis represents the true positive fraction (TPF), also known as sensitivity. The distinction between positive and negative in a segmentation result is sometimes artificial, because it is not always easy to confirm that a mammogram has a mass or not. In this paper, we used radiologist evaluation to count the number of detected lesions by the proposed algorithms. According to the radiologist rating, sensitivity and FP/I are calculated.

By analyzing different figures [Figures 4(a) and 4(b)], we find that the best FROC curve is given in Figure 4(a). This curve has been generated using the Expectation Maximization Bootstrap Subgroup (EMBS) method. In this method, it is observed that the TP rate of about 94.5% and 92.3%, with corresponding False positive per Image is 0.26 and 0.33 with $k = 8$ and $k = 5$ respectively. Figure 4(b) shows the FROC curve for K-means Bootstrap Subgroup (KMBS). In this method, we observe a TP rate of about 93.4% and 91.3%, with False positive per Image is 0.33 and 0.37 with $k = 8$ and $k = 5$. From this observation, fixing $k = 8$ helps the model to improve the accuracy of diagnostic test. The choice of $k = 5$ is also meaningful when compared to classical EM and K-Means approaches.

5. **Conclusion.** In this article, a practical CAD system is developed by designing new detection algorithms especially for the segmentation of Radio Opaque Lesions by using K-Means Bootstrap Subgroup (KMBS) and Expectation Maximization Bootstrap Subgroup (EMBS). The newly designed algorithms are capable of segmenting the regions of smoothly varying intensity distribution in mammogram. The experimental result shows, the number of image regions being 8 and 5, it yields True Positive (TP) rate of 94.5% and 92.3% for EMBS with false positive per Image of 0.26 and 0.33 respectively. When KMBS method is applied, it results in TP rate of 93.4% and 91.3% with false positive per image of 0.33 and

0.37. FROC curve is generated to evaluate the performance of the proposed approach. With these new segmentation results, the proposed systems are capable of detecting Radio Opaque Lesions of different types at low false positive rates.

REFERENCES

[1] *Cancer Facts and Figures 2011*, American Cancer Society, Inc., Atlanda, www.cancer.org, 2011.

[2] *Breast Cancer Facts & Figures 2011-2012*, American Cancer Society, Inc., Atlanta, 2012.

[3] J. W. Byng, N. F. Boyd, E. Fishell, R. A. Jong and M. J. Yaffe, Automated analysis of mammographic densities, *Phys. Med. Biol.*, vol.41, pp.909-923, 1996.

[4] N. Karssemeijer, Automated classification of parenchymal patterns in mammograms, *Phys. Med. Biol.*, vol.43, pp.365-378, 1998.

[5] S. R. Aylward, B. M. Hemminger and E. D. Pisano, Mixture modeling for digital mammogram display and analysis, in *Digital Mammography, (Computational Imaging and Vision Series)*, N. Karssemeijer et al. (eds.), The Netherlands, Kluwer, Dordrecht, 1998.

[6] L. Tabar and P. B. Dean, *Teaching Atlas of Mammography*, Georg Thieme Verlag, New York, 1985.

[7] M. M. Anguh and A. C. Silva, Multiscale segmentation and enhancement in mammograms, *Proc. of SIBGRAPI'97*, Campos do Jordao, pp.136-139, 1997.

[8] Z. Liang, J. R. MacFall and D. P. Harrington, Parameter estimation and tissue segmentation from multispectral MR images, *IEEE Trans. Med. Imag.*, vol.13, no.3, pp.441-449, 1994.

[9] M. A. Kupinski and M. L. Giger, Automated seeded lesion segmentation on digital mammograms, *IEEE Trans. Med. Imag.*, vol.17, no.4, pp.510-517, 1998.

[10] H. Li, Y. Wang, K. J. R. Liu, S.-C. B. Lo and M. T. Freedman, Computerized radiographic mass detection – Part I: Lesion site selection by morphological enhancement and contextual segmentation, *IEEE Trans. Med. Imag.*, vol.20, no.4, pp.289-301, 2001.

[11] T. Lei and W. Sewchand, Statistical approach to X-ray CT imaging and its applications in image analysis – Part II: A new stochastic model-based image segmentation technique for X-ray CT image, *IEEE Trans. Med. Imag.*, vol.11, no.1, pp.62-69, 1992.

[12] T. Kanungo, D. M. Mount, N. S. Netanyahu et al., An efficient k-means clustering algorithm: Analysis and implementation, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.24, no.7, pp.881-892, 2002.

[13] A. K. C. Wong, D. K. Y. Chiu and W. Huang, A discrete valued clustering algorithm with applications to biomolecular data, *Information Sciences*, vol.139, pp.97-112, 2001.

[14] C. W. Chen, J. Luo and K. J. Parker, Image segmentation via adaptive K-mean clustering and knowledge-based morphological operations with biomedical applications, *IEEE Trans. on Image Processing*, vol.7, no.12, pp.1673-1683, 1998.

[15] H. P. Ng, S. H. Ong, K. W. C. Foong, P. S. Goh and W. L. Nowinski, Medical image segmentation using K-means clustering and improved watershed algorithm, *IEEE Southwest Symposium on Image Analysis and Interpretation*, pp.61-65, 2006.

[16] A. Cao, Q. Song and X. Yang, Robust information clustering incorporating spatial information for breast mass detection in digitized mammograms, *Computer Vision and Image Understanding*, vol.109, pp.86-96, 2008.

[17] H. Kobatake, M. Murakami, H. Takeo and S. Nawano, Computerized detection of malignant tumors on digital mammograms, *IEEE Trans. Med. Imag.*, vol.18, no.5, pp.369-378, 1999.

[18] X. P. Zhang and M. D. Desai, Segmentation of bright targets using wavelets and adaptive thresholding, *IEEE Trans. on Image Processing*, vol.10, no.7, pp.1020-1030, 2001.

[19] A. K. Santra, W. JaiSingh and S. Devaarul, Pixcals statistical based algorithm to detect microcalcifications on mammograms, *International Journal of Computational Intelligence Research*, vol.6, no.2, pp.275-288, 2010.

[20] A. K. Santra, W. JaiSingh and S. Devaarul, Detection of microcalcifications using pixcals refined bandwidth algorithm in digitized mammograms, *International Journal of Tomography and Statistics, Winter 2011*, vol.16, no.W11, pp.81-90, 2011.

[21] W. JaiSingh and B. Nagarajan, Automatic diagnosis of mammographic abnormalities based on hybrid features with learning classifier, *Computer Methods in Biomechanics and Biomedical Engineering*, DOI:10.1080/10255842.2011.639015, 2012.

[22] R. Sammouda, J. Hassan and M. Sammouda, CT image analysis for early detection of lung cancer, *International Journal of Innovative Computing, Information and Control*, vol.4, no.11, pp.2847-2860, 2008.

[23] M. S. Mohamad, S. Omatu and M. Yoshioka, A cyclic hybrid method to select a smaller subset of informative genes for cancer classification, *International Journal of Innovative Computing, Information and Control*, vol.5, no.8, pp.2189-2202, 2009.

[24] W. Qian, L. P. Clarke, M. Kallergi and R. A. Clark, Tree-S tructured nonlinear filters in digital mammography, *IEEE Trans. on Medical Imaging*, vol.13, no.1, pp.25-36, 1994.

[25] T. O. Gulsrud and J. H. Husøy, Optimal filter-based detection of microcalcifications, *IEEE Trans. on Biomedical Engineering*, vol.48, no.11, pp.1272-1280, 2001.

[26] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd Edition, Prantice Hall of India Private Ltd., 2007.

[27] B. J. Leiner, V. Q. Lorena, T. M. Cesar and M. V. Lorenzo, Microcalcifications detection system through discrete wavelet analysis and contrast enhancement techniques, *IEEE International Conference on Electronics, Robotics and Automotive Mechanics*, pp.272-276, 2008.

[28] R. Aufrichtig and D. L. Wilson, X-ray fluoroscopy spatio-temporal filtering with object detection, *IEEE Trans. on Medical Imaging*, vol.14, no.4, pp.733-746, 1995.

[29] B. Efron, Bootstrap methods: Another look at the Jacknife, *Annals of Statistics*, vol.7, no.1, pp.1-26, 1979.

[30] T. Seppala, H. Moskowttz, R. Plante and J. Tang, Statistical process control via the subgroup bootstrap, *Journal of Quality Technology*, vol.27, pp.139-153, 1995.

[31] I. El-Naqa, Y. Yang, M. N. Wernick, N. P. Galatsanos and R. M. Nishikawa, A support vector machine approach for detection of microcalcifications, *IEEE Trans. on Medical Imaging*, vol.21, no.12, pp.1552-1563, 2002.

[32] K. Hu, X. Gao and F. Li, Detection of suspicious lesions by adaptive thresholding based on multiresolution analysis in mammograms, *IEEE Trans. on Instrumentation and Measurement*, vol.60, no.2, pp.462-472, 2011.