CHAPTER 3

A STUDY ON BREAST ABNORMALITY DETECTION

3.1 INTRODUCTION

A Computer-Aided Detection (CAD) system is used to aid radiologists in detecting mammographic lesions that may indicate the presence of breast cancer. This system acts only as a second reader, and the final decision is made by the radiologist. Recent studies have also shown that a CAD system, when used as an aid, has improved radiologists’ accuracy for detection of breast cancer (Giger et al 2001).

The steps in the general framework of a CAD system are:

1. Input: Region of Interest Selection (ROI, portion of a mammogram containing a breast masses and microcalcifications).
2. Preprocessing and enhancement.
3. Segmentation of the suspicious regions.
4. Hybrid feature extraction.
5. Classification of the breast mass and microcalcifications.
6. Output: Detection of the breast mass or microcalcifications.

The CAD system proposed in this thesis is also based on the above framework but proposes it is a variety mixture of segmentation, feature extraction and classification techniques in order to increase the reliability of the proposed system.
The rest of this chapter is structured as follows: Section 3.2 describes a proposed framework for detection of breast abnormalities in digitized mammograms. The sub section 3.2.1 shows the data set used in the experiment for detection of masses and cluster of microcalcification. In section 3.2.2, a brief description of Region of Interest selection is presented. Section 3.2.3 shows the preprocessing and enhancement techniques applied in the mammogram. A detailed description of segmentation of suspicious region is given in section 3.2.4. The reconstruction of suspicious region and separation of suspicious regions are brought out in section 3.2.5 and 3.2.6 respectively. Hybrid feature extraction and classification methods are given in section 3.2.7 and section 3.2.8 respectively. Experimental setup is elaborated in section 3.3. Finally, performance evaluation of the proposed algorithms and conclusion are given in section 3.4 and section 3.5 respectively.

3.2 PROPOSED FRAMEWORK FOR DETECTION OF BREAST ABNORMALITIES

The proposed framework aims to detect various cancerous parts in mammograms automatically. Figure 3.1 illustrates the underlying principle of the proposed CAD system. At the first stage, the Region of Interest (ROI) is selected from the mammogram as a sample image. In the second stage, the median filter is applied to remove the noise, and unsharp masking technique is used to enhance the quality of the sample image. During the third stage, a novel algorithm is employed to detect suspicious pixels. In the fourth stage, binary morphological operators and 8-connected component labeling methods are employed to reconstruct the shape, to remove isolated pixels and to segment the suspicious regions. Hybrid features are extracted from the segmented regions during the fifth stage. In the sixth stage, Support Vector Machine (SVM) classifier is used to pinpoint the lesions. Finally, the
efficiency of algorithm is measured using FROC curve, and the results are compared with the existing methods.

### 3.2.1 Data Set

At present, CAD schemes are frequently evaluated with a database generated by the investigators, which may contain different proportions of subtle cases and obvious cases. As a consequence, it is not possible to perform meaningful comparisons of different schemes. A common database is an important towards achieving consistency in performance comparison and the objective testing of algorithms. A digital mammography database, created by Mammographic Image Analysis Society (MIAS) (Suckling et al 1994), which is an organization of the United Kingdom research groups, is used in this research. It consists of 322 images, which belong to normal and abnormal categories. In addition, the abnormal cases are divided into six categories: circumscribed masses, microcalcifications, spiculated masses, ill-defined masses, architectural distortion and asymmetry. It is generally accepted that the essential characteristic of a high quality mammogram is, the ability to visualize these six features. All images are digitized at a resolution of 1024x1024 pixels and eight-bit accuracy (gray level). They additionally embody the locations of any abnormalities that may be present. The database is organized in pairs of films, where every pair represents the left (even filename numbers) and right (odd filename numbers) mammograms of a single patient, in Medio-Lateral Oblique (MLO) views. There are a number of variables that characterize each mammogram in MIAS database, as shown in Table 3.1.
Figure 3.1 Schematic representation of a proposed CAD System
Table 3.1 Characteristics of the MIAS database

<table>
<thead>
<tr>
<th>Column No.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MIAS database reference number.</td>
</tr>
<tr>
<td>2</td>
<td>Nature of background tissue: Fatty, Fatty-glandular, Dense-glandular</td>
</tr>
<tr>
<td>3</td>
<td>Class of abnormality present: Circumscribed masses 25 cases, Spiculated masses 19 cases, Ill-defined masses 15 cases, Architectural distortion 19 cases, Asymmetry 19 cases, Normal 207 cases.</td>
</tr>
<tr>
<td>4</td>
<td>Severity of abnormality: Benign 54/115 cases, Malignant 61/115 cases</td>
</tr>
<tr>
<td>5</td>
<td>Abnormality location: x,y image coordinates of center of abnormality</td>
</tr>
<tr>
<td>6</td>
<td>Approximate radius (in pixels) of a circle enclosing the abnormality</td>
</tr>
</tbody>
</table>

The existing data consist of the location of the abnormality (like the middle of a circle close the tumor), its radius, breast position (left or right), kind of breast tissues (fatty, fatty-glandular and dense) and tumor type if exists (benign or malign). Figure 3.2 shows the various classes of breast cancer prevails among women.
Figure 3.2 Different classes of breast cancer a) Clustered microcalcifications b) Spiculated lesion c) Circumscribed mass d) Ill defined mass e) Architectural distortion.

3.2.2 Region of Interest (ROI) Selection

An important characteristic of the MIAS database is that each abnormal image comes with a consultant radiologist’s truth data. The neighborhood of the abnormality is given as the coordinate of its center and an approximate radius (in pixels) of a circle enclosing the abnormality. From
this truth data, it is doable to extract sub-image manually. The sub-images contain all biopsy-truthed regions of interests, and a physician annotated abnormalities identified in corresponding pathology reports. These ROI images considerably reduce the algorithmic execution time, and also reduce computer storage necessities. The Figure 3.3(a) and 3.3(b) shows the original image \((mdb209, mdb110)\), and corresponding Region of Interest shows in Figure 3.3(c) and 3.3(d).

![Figure 3.3](image)

**Figure 3.3** Original and ROI images. (a) and (b) Original mammogram obtained from MIAS database \((mdb209, mdb110)\). (c) and (d) – ROI extracted from original images.

### 3.2.3 Preprocessing and Enhancement

In image pre-processing, it is often desirable to perform some kind of noise reduction on an image in order to improve the results of later processing. Median filter technique is widely utilized in digital image process. As a result of beneath sure conditions, it preserves edges whereas
removing noise in images. The principal objective of enhancement is to process an image so that the result is more suitable than the original image for a segmentation of suspicious region. Unsharp masking technique is more suitable for enhancing the quality of the image.

**Median Filtering**

Median filter is found to be very powerful in removing noise from two-dimensional signals without blurring edges. This method is particularly suitable for enhancing mammogram images (Wei Qian, 1994). 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. It is performed in order to eliminate small signals caused by random noise in mammogram. (Thor Ole Gulsrud 2001).

Figure 3.4 illustrates different support regions which are applied in the mammogram mdb245. The Figures 3.4(c)–(e) show the feature images produced when a median filter with a support region of size 5x5, 7x7 and 11x11 respectively, have been applied on the original-image. Clearly, by increasing the size of the support region, both noise signals and signals from lesions are being suppressed. Based on the experimental study, it is observed that when applied to the whole dataset the 7x7 and 11x11 support region it eliminates more of the distinct lesions than the 5x5 support. Thus, the 5x5 support region is best for noise removal and preservation of signals from lesions.
Figure 3.4  Different support region applied on the mammogram (mdb245) a) Original mammogram b) Zoom of the region of interest c) Resulting image – 5x5 Median filter d) Resulting image – 7x7 Median filter e) Resulting image – 11x11 Median filter.
Image Enhancement

Enhancement is aimed at realizing improvement in the quality of a given image (Rafael Gonzalez 2007). It can be accomplished by enhancing contrast and edges. By applying unsharp masking filters, contrast 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 neighborhoods of input pixels (Barba Leiner 2008, Richard Aufrichtig 1995). Often the image is convolved with a finite impulse response filter called spatial mask. Chan et al (1987) investigated the application of unsharp masking on diagnostic procedure. The ROC studies are conducted and they show that unsharp masking improved the detectability of calcifications in digital mammograms.

The unsharp masking filter is a simple sharpening operator. It enhances edges (and other high frequency components in an image) via a procedure which subtracts an unsharp or smoothened version of an image from the original image. The image \(f(x,y)\) obtained through the median filter under sub-section is used as an input image.

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

\[
g(x,y) = f(x,y) - f_{\text{smooth}}(x,y)
\]

(3.1)

where \(f_{\text{smooth}}\) 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 \(f_{\text{sharp}}(x,y)\) is obtained from the model.

\[
f_{\text{sharp}}(x,y) = f(x,y) + \lambda g(x,y)
\]

(3.2)
where $\lambda$ 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.

$$g(x,y) = f(x,y) - \frac{1}{4}[f(x-1,y) + f(x,y-1) + f(x+1,y) + f(x,y+1)]$$

3.2.4 Suspicious Region Segmentation

After the image has been preprocessed, the suspicious regions are segmented from the enhanced image. The Bootstrap Subgroup and Classical Segmentation approaches are used in detection of suspicious masses. The novel combination of Bootstrap method and classical segmentation approach are used in detection of cluster in microcalcifications. In these methods, each pixel is taken into consideration to detect lesions.

Bootstrap Method

The Bootstrap method originally proposed and named by Efron (1979). This is a probabilistic technique that can be used effectively to estimate the sampling distribution of any statistics. In particular, one can use the non-parametric Bootstrap to estimate the sampling distribution of a statistics. The sample is one of the representatives of the population from which it is drawn. The observations are independent and identically distributed. In its simplest form, the non-parametric 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, $F$. 
\[ F_n(x) = \frac{1}{n} \cdot \text{(Number of } X_i \leq x) \]  

(3.4)

A simple random K sample of size n can be drawn from the unspecified distribution F and denoted by \( x_1^*, x_2^*, \ldots, x_k^* \). This sample is called the Bootstrap sample. An estimate of \( (\bar{x}^*) \) can be computed from the Bootstrap sample. This estimate is called the Bootstrap estimate. The K Bootstrap estimates \( \bar{x}_1^*, \bar{x}_2^*, \ldots, \bar{x}_k^* \) can be computed from the Bootstrap samples. Finally, the smallest ordered \( \bar{x}^* \) is found out using \((1 - \alpha / 2)k\). Where \( \alpha \) is the false alarm rate.

**Bootstrap Subgroup Method**

Seppala et al (1995) have proposed a technique called the Bootstrap Subgroup chart that is used to monitor the quality of the process or system. The Bootstrap subgroup model is given below.

\[ X_{ij} = \mu_i + \epsilon_{ij} \]  

(3.5)

where \( i = 1, 2, \ldots, k \) and \( j = 1, 2, \ldots, n \). Here, \( \mu_i \) is the true mean of the \( i^{th} \) subgroup, and \( \epsilon_{ij} \) is a random error term. The K-means and Expectation Maximization algorithms are embedded in the Bootstrap and Bootstrap Subgroup probabilistic methods for drawing K samples from the image data.

### 3.2.5 Reconstruction of Suspicious Region

Mathematical morphology is a powerful tool for image shape analysis. Binary morphological techniques are employed to reconstruct the shapes of the lesions. The isolated pixels can also be removed. Heng-Da Cheng et al (1998) have used a novel approach to detect microcalcification based on fuzzy logic technique. In this approach, the shapes of
microcalcifications are reconstructed and the isolated pixels are removed by employing the mathematical morphology technique. Stelios Halkiotis et al (2007) have developed a new algorithm for the detection of clustered microcalcifications using mathematical morphology and artificial neural networks. Joachim Dengler et al (1993) have developed a systematic method for the detection and segmentation of microcalcifications in mammograms. The method uses a two-stage algorithm for spot detection and shape extraction. A morphological operator reproduces the shape of the spots.

The reconstruction of suspicious regions algorithm is given below.

**Step 1:** Read binary image and store it in a two dimensional matrix $R(x,y)$.

**Step 2:** Apply dilation operator to fill in holes and broken areas and to connect areas that are separated by spaces smaller than the size of the structuring element.

$$R_i(x,y) = \left\{ x \mid (\hat{B})_x \cap R(x,y) \neq \emptyset \right\}$$  

(3.6)

**Step 3:** Apply erosion operator to remove small anomalies by subtracting objects with a radius smaller than the structuring element.

$$R_2(x,y) = \left\{ x \mid (\hat{B})_x \subseteq R_i(x,y) \right\}$$  

(3.7)

**Step 4:** Opening operator is applied to smooth contours of a region to remove bridges between regions and to remove sharp peaks sticking out of regions.

$$R_3(x,y) = (R_2(x,y) \square B) \oplus B$$  

(3.8)

**Step 5:** Apply closing operator to fuse nearby regions, to fill gaps and holes in a region.

$$R_4(x,y) = (R_3(x,y) \oplus B) \square B$$  

(3.9)
Step 6: The suspicious region reconstruction is performed by the combined operator called close-opening.

\[ R_z(x, y) = (R_4(x, y) \bullet B) \circ B \]  \hspace{1cm} (3.10)

Here, B is a small image called structure element. In the experiments, a 2x2 square structure element is employed in step (2) and (3). The size of the structuring element is chosen experimentally in step (4)-(6), so as not to eliminate any lesions.

3.2.6 Separation of Suspicious Region

A lesion is an area which shows an abnormality or alteration in the tissue’s integrity. Breast lesions usually appear in the form of lumps or swellings in and around the breast area. The image must be partitioned into regions that correspond to objects or parts of an object for correct interpretation. 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 forms a connected group is known to be the suspicious region. With the aim of detection of masses, all regions with a pixel count of less than 600 pixels are eliminated, because they are well below the size of the smallest mass in the MIAS database. The features of remaining suspicious region will be generated in the classification step for mass detection.

3.2.7 Hybrid Feature Extraction

The three main lesion features in mammogram are texture, shape and gray level. Several lesion feature based schemes for mass detection and segmentation have been developed in recent years. Cao et al (2008) have proposed a method called robust information clustering incorporating texture features for detection of lesions in mammograms. Hidefumi et al (1999) have
used iris filter method to segment the suspicious regions and shape features are used to detect the malignant tumor. Mencattini et al (2008) have proposed a novel algorithm for image denoising enhancement based on wavelet processing. They developed a new segmentation method combining wavelet information with shape features. Zhang and Desai (2001) have developed a systematic method for the detection and segmentation of bright targets by using gray level features. Zhang and Desai (2001) have developed a method to segment suspicious tumor areas with exact boundaries in mammograms based on multiscale analysis in both grayscale and space.

To locate the regions that are suspicious of tumors, certain features are selected to use in the CAD system. A set of 19 hybrid features are 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 3.2. Based on the feature extraction, Support Vector Machine (SVM) is used as classifier to classify, further, the suspicious region into lesion or normal.

**Table 3.2 Hybrid features**

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texture</td>
<td>Mean gradient, Variance of gradient, Inertia, Correlation, Energy, Homogeneity, Entropy</td>
</tr>
<tr>
<td>Shape</td>
<td>Area of the lesion, Perimeter, Orientation, Extent of lesion, Solidity, lesion eccentricity, lesion’s equivalent diameter</td>
</tr>
<tr>
<td>Gray Level</td>
<td>Mean intensity, Variance intensity, Standard deviation of intensity, Minimum intensity, Maximum intensity</td>
</tr>
</tbody>
</table>
Texture Features

This feature studies the spatial relationship among pixels in the suspicious region. The Gray Level Co-occurrence Matrix (GLCM) is created from the suspicious region. It is calculated by using how often a pixel with gray-level value $i$ occurs horizontally adjacent to a pixel with the value $j$. Each element $(i, j)$ in GLCM specifies the number of times that the pixel with value $i$ occurred horizontally adjacent to a pixel with value $j$. $p(i, j)$ is the probability of element $(i, j)$ occurring in GLCM. The features inertia, correlation, energy, homogeneity and entropy are calculated using GLCM.

Feature (1): Inertia

It measures the intensity contrast between a pixel and its neighbor over the suspicious region.

\[
Inertia = \sum_{i,j} |i - j|^2 p(i, j)
\]  

(3.11)

Feature (2): Correlation

A measure of how interrelated a pixel is to its neighbor over the whole suspicious region.

\[
Correlation = \sum_{i,j} \frac{(i - \mu_i)(j - \mu_j)p(i, j)}{\sigma_i\sigma_j}
\]  

(3.12)

Feature (3): Energy

The sum of squared elements in the Gray Level Co-occurrence Matrix (GLCM)

\[
Energy = \sum_{i,j} p(i, j)^2
\]  

(3.13)
Feature (4): Homogeneity

It measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.

\[
\text{Homogeneity} = \frac{\sum_{i,j} p(i,j)}{1 + |i - j|}
\]  
(3.14)

Feature (5): Entropy

It is the statistical measure of randomness that can be used to characterize the texture of the suspicious region.

\[
\text{Entropy} = -\sum_{i,j} p(i,j) \log_2[p(i,j)]
\]  
(3.15)

Feature (6): Mean gradient

This feature measures the mean gradient of each pixel in the region.

\[
mg = \frac{1}{N} \sum_{k} g_k
\]  
(3.16)

where \(N\) is the number of pixels within the region, \(g_k\) is the gradient at each pixel \(k\).

Feature (7): Variance of gradient

It refers to the variance in the gradient of the pixels in the suspicious region.

\[
vg = \frac{\sum_{k=1}^{n} (g_k - mg)^2}{n}
\]  
(3.17)
where \( g_k \) is the gradient at each pixel \( k \), \( m_g \) is the mean gradient of the region, \( n \) is the number of pixels in the region.

**Shape Features:**

These features are used to measure the properties of the suspicious region.

**Feature (8): Area**

It returns the actual number of pixels in the suspicious region.

\[
f(x, y) = \begin{cases} 
1 & \text{if}(x, y) \in \text{object} \\
0 & \text{otherwise}
\end{cases} \tag{3.18}
\]

\[
\text{Area} = \sum_{x} \sum_{y} f(x, y) \tag{3.19}
\]

where, \( f(x,y) \) – suspicious region, \( x,y \) – Index number of the suspicious region

**Feature (9): Perimeter**

It gives the number of pixels on the contour of the suspicious region. This perimeter is used for the parametric boundary representation. If \( x_1, x_2, \ldots, x_n \) is a boundary coordinate list, \( N \) is the number of pixels on the boundary, the region perimeter is given by:

\[
P = \sum_{i=1}^{N} |x_i - x_{i+1}| \tag{3.20}
\]
Feature (10): Solidity

It returns the proportion of the pixels in the convex hull that are also in the region.

\[ Solidity = \frac{\text{Area}}{\text{ConvexArea}} \]  \hspace{1cm} (3.21)

Feature (11): Eccentricity

Eccentricity is the measure of aspect ratio. It is calculated by using minimum bounding rectangular box. It is the smallest rectangle that contains every point in the shape. For an uninformed shape, eccentricity is the ratio of the length \( L \) and width \( W \) of minimal bounding rectangle of the shape at some set of orientations.

\[ E = \frac{L}{W} \]  \hspace{1cm} (3.22)

Feature (12): Extent

It refers to the proportion of pixels in the bounding rectangular box of a region. It is computed as the area divided by the area of the bounding box.

\[ Extent = \frac{\text{Area}}{L \times W} \]  \hspace{1cm} (3.23)

where \( L \) - length of the bounding rectangular box.

\( W \) - Width of the bounding rectangular box.
Feature (13): Orientation

Orientation is referred as the angle (relative to the x-axis) of the axis through the center of mass that gives the lowest moment of inertia.

Orientation, $\theta$, with respect to the x-axis is found by minimizing the sum

$$I(\theta) = \sum_{\alpha} \sum_{\beta} (\beta - \bar{\beta})^2 f(\alpha, \beta)$$

where $\alpha$ and $\beta$ are the rotated co-ordinates.

$$\alpha = x\cos \theta + y\sin \theta$$
$$\beta = y\cos \theta - x\sin \theta$$

Substituting $\alpha$ and $\beta$ in Equation (3.24)

$$I(\theta) = \sum_{x} \sum_{y} [(y - \bar{y})\cos \theta - (x - \bar{x})\sin \theta]^2 f(x, y)$$

Feature (14): Equivalent Diameter

It refers to the diameter of a circle with the same area as the region.

$$Eq.diameter = \sqrt{\frac{4 \cdot Area}{\pi}}$$

where, Pi=3.14

Gray Level Features:

These features are based on the gray level pixel value of the enhanced image. For each of the formula, n is the number of pixels in the region, $I_g$ is the gray level value of g in region I, $\mu$ is the mean of the region.
Feature (15): Mean

It gives the mean of gray level values in the suspicious region.

\[
Mean = \frac{1}{n} \sum_{g=1}^{n} I_g 
\]  

(3.27)

Feature (16): Variance

It measures the smoothness of the region.

\[
\text{var} = \frac{\sum_{g=1}^{n} (I_g - \mu)^2}{n} 
\]  

(3.28)

Feature (17): Standard Deviation

It measures how far the gray values are spread out in the suspicious region.

\[
\text{std.deviation} = \sqrt{\frac{1}{n} \sum_{g=1}^{n} (I_g - \mu)^2} 
\]  

(3.29)

Feature (18): Minimum Intensity

It refers to the value of the pixel with the lowest gray value in the region.

\[
I_{\min} = \min_{g} I_g 
\]  

(3.30)
Feature (19): Maximum Intensity

It refers to the value of the pixel with the greatest gray value in the region.

\[ I_{\text{max}} = \max_g I_g \]  \hspace{1cm} (3.31)

3.2.8 SVM Classifier towards Mammographic Abnormality Detection

Image classification analyses the numerical properties of various image features and organizes data into different classes. Compared with other classification algorithm, Support Vector Machine (SVM) provides good generalization accuracy and is fast to learn. SVM is a machine learning method for creating a classification function from a set of labeled training data (Issam El-Naqa 2002). According to literature review, many researchers (Papadopoulos et al 2005, Fu et al 2005, Tomasz Arodz et al 2005, Cao et al 2008, Wener et al 2011) have used SVM classifier for classification of lesion in digitized Mammograms.

Classification of breast abnormality has been performed as a two class problem, and the two classes are lesion and normal. To begin with, let vector \( x \in R^n \) denote a pattern to be classified, and let scalar \( y \) denote its category label (i.e., \( y \in \{\pm 1\} \)). Additionally, let \( \{(x_i, y_i), i = 1, 2, ..., n\} \) denote a given set of \( n \) training examples, where every sample \( x_i \) has a known category label \( y_i \). The matter is to work out a classifier \( f(x) \) (i.e., a decision function) that may properly classify an input pattern.
An SVM classifier is used in the classification step. First, it maps the input data vector \( x \) into a higher dimensional space \( H \) through an underlying nonlinear operator \( \phi(\cdot) \). The non-linear SVM classifier so obtained is defined as

\[
\begin{align*}
  f(x) &= w^T \phi(x) + b \\
  &\quad \text{(3.32)}
\end{align*}
\]

which is linear in terms of the transformed data \( \phi(x) \), but non-linear in terms of the original data \( x \in R^p \). The parameters \( w, b \) are determined from the training data samples.

Following nonlinear transformation, the parameters of the decision function \( f(x) \) are determined by the following minimization:

\[
\begin{align*}
  \min_{w, \xi} \ J(w, \xi) &= \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i \\
  &\quad \text{subject to } y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0; i = 1, 2, \ldots, n. \\
  &\quad \text{(3.33)}
\end{align*}
\]

where, \( C \) is a user specified positive parameter, \( \xi_i \) are slack variables.

Using the technique of Lagrange multipliers, one will show that a necessary condition for minimizing \( J(w, \xi) \) in Equation (3.33) is that vector \( w \) is created by a linear combination of the mapped vector \( \phi(x_i) \), i.e.,

\[
\begin{align*}
  w &= \sum_{i=1}^{n} \alpha_i y_i \phi(x_i) \\
  &\quad \text{(3.35)}
\end{align*}
\]

where \( \alpha_i \geq 0, i = 1, 2, \ldots, n \), are the Lagrange multipliers associated with the constraints in Equation (3.34).
Substituting Equation (3.35) into Equation (3.32) yields

\[
f(x) = \sum_{i=1}^{n} \alpha_i y_i \phi^T (x_i) \phi(x) + b \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b
\]  

(3.36)

where the function \( K(\cdot, \cdot) \) is defined as

\[
K(x, z) \equiv \phi^T (x) \phi(z)
\]  

(3.37)

The Lagrange multipliers \( \alpha_i \geq 0, \ i=1,2,...,n \) are solved from the dual form of Equation (3.33), which is expressed as

\[
\max w(\alpha_1, \alpha_2, \ldots, \alpha_n) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i, x_j)
\]

subject to

\[
0 \leq \alpha_i \leq C, \quad i=1,2,...,n, \quad C > 0
\]

(3.39)

\[
\sum_{i=1}^{n} \alpha_i y_i = 0
\]

(3.40)

The cost function \( w(\alpha_1, \alpha_2, \ldots, \alpha_n) \) is convex and quadratic in terms of the unknown parameters \( \alpha_i \). In practice, this problem is solved numerically through quadratic programming.

The training examples in a typical problem are correctly classified by the trained classifier, i.e., only a few training examples will support vectors. To simplify, let \( S_j, \ \alpha_j^*, \ j=1,2,...,n_s \), denote these support vectors and their corresponding nonzero Lagrange multipliers, respectively, and let \( y_i \)
denote their category labels. The decision function in Equation (3.36) can, thus, be simplified as

\[
f(x) = \sum_{j=1}^{n_s} \alpha_j^* y_j \phi^T(S_j)\phi(x) + b \sum_{j=1}^{n_s} \alpha_j^* y_j K(S_j, x) + b \quad \text{(3.41)}
\]

where the function \( K(S_j, x) \) is defined as radial basis function kernel:

\[
K(S_j, x) = \exp \left( -\frac{\|S_j - x\|^2}{2\sigma^2} \right), \quad \sigma > 0 \text{ may be a constant that defines the kernel width.}
\]

### 3.3 EXPERIMENTAL SETUP

#### 3.3.1 Training and Testing Data Set

The training data set comprising tumor cases of 40 images and 40 non tumor cases is used for detection of masses. The 50 single microcalcifications and 50 normal cases are used for detection of microcalcification in training data set.

With the aim of detecting the masses, the test data set comprising tumor cases of 89 images and 75 non tumor cases is used. Among the 89 tumor cases, the set comprised 19 spiculated masses (SPIC), 22 circumscribed masses (CIRC), 14 irregular masses (MISC), 15 cases with asymmetric densities (ASYM), and 19 cases with architectural distortion (ARCH). To evaluate the computer aided detection results, the findings of the proposed method are considered True Positive (TP), if its area is overlapped by, at least, 50% of a true lesion.
With the aim of detecting the cluster of microcalcification, the test data set of 50 mammograms is selected from the databases. Among the 50 mammograms, 25 mammograms are normal and 25 are abnormal. Before applying the proposed approach, region of interest is located using true information provided by the database. In MIAS database, centre location and radius are not provided for the images mdb216, mdb233 and mdb245. So, it considered these three images contain one cluster. To evaluate the computer aided detection results, a group of microcalcification pixels (≥2x2 pixels) in a 15x15 pixels box is considered a cluster. A true positive cluster is a detected cluster which has overlapped more than 50% with the suspicious area provided by the MIAS database. All other clusters are considered to be False Positive (FP) clusters.

3.4 PERFORMANCE EVALUATION

3.4.1 Radiologist Rating

The results obtained from the proposed methods are evaluated by a radiologist who is an expert in mammography analysis. The original and resulting images are simultaneously presented on a computer monitor for subjective assessment. By visual comparison, the radiologist’s assigned one of four ranking options (Excellent, Good, Average and Poor) to the segmentation results, such that the detected suspicious region overlapped by 90% - 100 % is considered as Excellent, 70% - 90% is Good, 50% - 70% is Average and less than 50% is Poor. According to the radiologist’s rating, 89 tumor cases which contain 92 lesions are analyzed. Similarly, for detection of cluster of microcalcification, 25 tumor cases which contain 30 microcalcification clusters are analyzed. True positive rate and false positive
per image are calculated based on the radiologist’s rating to generate a Free Receiver Operating Characteristics (FROC) curve. The FROC gives performance evaluation of the proposed CAD system.

### 3.4.2 FROC Analysis

FROC analysis is performed to evaluate the performance of the proposed system. The metrics used to report the performance of detection in algorithms are True Positive Fraction (TPF) and the number of False Positives per Image (FP/I). The sensitivity of CAD system is measured as the ratio between TPF and FP/I. A true positive mark is an indication made by the CAD system that corresponds to the location of a lesion. A false-positive mark is an indication made by the CAD system that does not correspond to the location of a lesion. A plot of TPF versus FP/I is called a Free Receiver Operating Characteristic (FROC) curve, and this is generally used to report the performance of the detection in algorithm. In the FROC curve, $x$-axis represents the average number of false positive per image, and $y$-axis represents the true positive fraction.

\[
TPF = \frac{\text{Number of True Positive marks}}{\text{Number of lesions}}
\]  \hfill (3.42)

\[
FPI = \frac{\text{Number of False Positive marks}}{\text{Number of Images}}
\]  \hfill (3.43)

Li et al (2001) have proposed a statistical model for enhanced segmentation and extraction of suspicious masses from mammographic images. They have implemented the experiment with different number of clusters (K). The experimental results indicate that the selection of suspected masses can be affected by different values of K. They have proposed the EM
algorithm with the information theoretic criteria to determine the optimal value of K. Aylward et al (1998) have proposed a mixture model for mammogram analysis by taking into account the background, uncompressed-fat, fat, dense and muscle region. The number of cluster (K) found by Li et al and Aylward et al is optimal. Hence K=5 and K=8 are applied to analyse the FROC in the proposed methods.

3.5 SUMMARY

This chapter elucidates the general CAD framework and also discusses a specific framework on which the methods proposed in the thesis are built. The step of the proposed framework and the dataset used in the experiment are specified distinctly. An elaborate description on the selection of region of interest, preprocessing methods and enhancement techniques is given. It exhibits the segmentation of suspicious region by combination of classical segmentation approach with Bootstrap technique. Reconstruction of features, classification of lesions and performance evaluation are also highlighted.