CHAPTER 4

FUZZY C MEANS CLUSTERING AND STATISTICAL MOMENTS CALCULATION BASED DETECTION

4.1 INTRODUCTION

Detection of cancer information from RoI is the third and final stage of a CAD system. Microcalcifications, masses and statistical distortions are the most important indications of breast cancer. Several algorithms have been developed for the detection of those indications and follow any one of the three major detection techniques. They are multiresolution analysis, filtering methods and statistical methods. The ultimate aim of CADi is to help the radiologist in making decisions to find the malignancy (cancerous) and benign (non cancerous), giving suggestions and recommendations for the cancer patients. Information about the CAD, CADi, classification of suspected features and detection of cancer has been discussed in this chapter.

4.2 LITERATURE REVIEW ON COMPUTER AIDED DETECTION AND DIAGNOSIS

Support Vector Machine (SVM) used to separate the false signals from microcalcifications for the reduction of false positives tested by Papadopoulos et al (2008), it combines a multiresolution analysis based on wavelet transform with a difference image method and Gaussianity statistical test to perform a logical OR operation on the detected microcalcifications. Armando Bazzani et al (2000) tested the filtering method and got the
sensitivity of 94.6% with 0.6 False Positive (FP) per image on 40 sample images. A back propagation neural network combined with Kalman filter have been tested by Zhang et al (2002) and provided the result for the detection of microcalcification with best false positive reduction of 3.15 per image with an accuracy of 88%. Discrete Fourier Transform and Discrete Wavelet Transform have been used by Harris Georgiou et al (2007) for the detection of masses from mammogram images with 91.54% success rate.

Brake & Karssemeijer (2001) studied that the textures of masses generally present distinct features from that of normal breast tissues. Two methods have been developed by Sheng-Chih-Yang et al (2005) for texture feature extraction. They are Spatial Gray Level Difference Matrix (SGLDM) and fractal dimension. The SGLDM has been widely used for texture analysis and calcifications separated from the background by exploiting the evaluation of Renyi’s information at different decomposition levels of the wavelet transform. ON wavelet has also been applied for input image compression and a’trous algorithm have been tested for the good computational efficiency. Giuseppe Boccignone et al (2000) proposed Mallat’s decomposition for the nonlinear estimation with thresholding for the exact detection of masses from the mammogram. Hierarchical Image Property (HIP) and Hidden Markov Tree (HMT) techniques were been proposed by Paul Sajda et al (2003) for the detection of masses with the detection accuracy of 86%. Among the two methods they have concluded that the HIP performs well by producing 78% sensitivity with 16% reduction in FP. Polakowski et al (1997) used a single Difference of Gaussian (DoG) filter to detect masses. The DoG filter has been designed to match masses which were approximately 1 cm in diameter and RoI selection from the filtered image. They used nine features based on size, contrast, circularity and texture features to reduce the number of FP and to then classify RoI as malignant or normal (benign). The DoG filter, which is a band pass filter, has been used by several researchers for the preliminary task of detecting potential masses in an image.
Brzakovic et al (1990) used a two stage multiresolution approach for the detection of masses. Qian et al (1999) developed a multiresolution and multiorientation wavelet transform for the detection of masses and spiculation analysis. And observed that, traditional wavelet transforms cannot extract directional information which is crucial for a spiculation detection task and thus, they introduced a Directional Wavelet Transform (DWT) and found 97% accuracy. Petrick et al (1999) studied about the methods have been developed for the detection of features such as central mass region called pixel based method and methods developed for the detection of masses called region based method. Lai et al (1989) developed a simple template matching algorithm to detect circumscribed masses only. Neural and Fuzzy based classifiers have been dealt by Cheng et al (2006), Sajda et al (2002) and Floyd et al (1994) for the classification of cancers with less FP rate. Robust Information Clustering (RIC) algorithm incorporating spatial information for breast mass detection has been adopted by Aize Cao et al (2008). The detection system employs RIC algorithm based on the RoI extracted from global mammogram by two steps of adaptive thresholding. Pixels on the Fuzzy margin of a mass and noisy data were identified by RIC through the minimax optimization of mutual information.

Brzakovic et al (1990) presented in their literature about most diagnosis algorithms begin with RoI containing the abnormality. Rangayyan et al (2007) studied and included the possible methods involved in the CADI for the diagnosis of cancer presence. If a mass is suspected to be malignant
Figure 4.1 Simple CAD and CADi system

Segmentation stage from the Figure 4.1 is about the RoI selection. Then from the segmented RoI, features will be selected, classified and the diagnosis output will be produced in the output stage. The output of CADi system may be the likelihood of malignancy or a management recommendation. Features are extracted from RoI containing the abnormality and each RoI is classified using a classification algorithms. One of the main points that should be taken under serious consideration when implementing a robust classifier for recognizing breast tissue is the selection of the appropriate features that describe and highlight the differences between the abnormal and the normal tissue in an ample way.

Numerous classification methods have been reported are based on the Fuzzy set theory discussed by Brzakovic & Neskovic (1993). Neural network based classification algorithms have been discussed by Floyd et al (1994) and Sajda et al (2002). Markov random field models for mass classification have been tested by Li et al (1995). Artificial neural networks
Most of the classification algorithm used by the researchers is supervised method that is first trained on a set of sample cases called the training set. Huo et al (1998) dealt supervised method in their study for the classification of masses. Statistical based classification has been adopted by Christoyianni et al (1999). Classification of all kinds of abnormalities in RoS has been achieved by Radial Basis Function Neural Network (RBFNN). This network has shown impressive accuracy in detection and recognition because of its generalization capabilities and the fast learning rates has been studied by Christoyianni et al (2000). The feature descriptions utilized for classification are the Gray Level Histogram Moments (GLHM) and the Spatial Gray Level Dependence (SGLD) matrix features, which result from statistical texture analysis and the descriptors extracted from RoS using a technique employing Independent Component Analysis (ICA), capable of distinguishing between timorous and healthy tissue among various parenchymal tissue patterns has been studied by Christoyianni et al (2002). The metrics used to report the accuracy of these algorithms are sensitivity and specificity. Sensitivity is a TP classification defined as s lesion for which the CAD predicts that it is cancerous and it is actually found to be cancerous (malignant). Specificity is the fraction of benign lesions that are correctly identified by the CAD as being benign. From the Figure 4.2, it is easy to understand the shape of the suspected area and the nature of the cancer. It starts at round shape as benign and ends at stellate shape as malignant.

**Figure 4.2  Flowchart showing the stages involved in the diagnosis of abnormalities**
Receiver Operating Characteristics (ROC) and Free-response Receiver Operating Characteristics (FROC) are the two important graphical methods to evaluate the results of breast cancer from the collected information through various databases and plotted for the detection and diagnosis respectively. These plots help to find the detection accuracy of the methods proposed. A plot of sensitivity versus specificity is called a ROC curve and this is generally used to report the performance of the diagnosis algorithm. An example of an ROC curve is shown in figure 4.3. The FROC curve can be used to report the performance of the detection algorithms. A plot of sensitivity versus False Positive Index (FPI) is called a PROC curve. It is also shown in Figure 4.3.

Sensitivity = (Number of TP Classifications/Number of Malignant Lesions), Specificity = (Number of TN Classifications/Number of Benign Lesions) and FPI (False Positive Index) = (Number of FP marks/Number of images. These are the factors involved in the construction of ROC and FROC curves.

The sensitivity is also called as True Positive Fraction (TPF) and the specificity is called as False Positive Fraction (FPF). The four possible

Figure 4.3 ROC and FROC curves
diagnosis results are True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN). The diagnosis results are tabulated in Table 4.1.

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Patient is Abnormal</th>
<th>Patient is Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abnormal</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>Normal</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

From the Table 4.1, diagnosis can be made based on four different possibilities. If the diagnosis is abnormal and the patient is also abnormal then it is called as TP, if the diagnosis is abnormal and the patient is normal then it is called as FP, if the diagnosis is normal and the patient is abnormal then it is called as FN and if the diagnosis is normal and the patient is also normal then it is called as TN. Detection and diagnosis of cancer is purely based on the evaluation done by the medical expert or radiologist.

After the completion of preprocessing and segmentation, the segmented RoI is further processed by the classification methods for texture feature detection. Many methods have been proposed to achieve this task. The texture feature and the shape feature are two important features involved in identifying the benign and malignant masses. Detection of those features has been grouped into two methods. They are clustering based methods and Histogram based textured feature detection methods. Calculation of many parameters has been discussed in this chapter under those methods. Detailed description of those methods is as follows.
4.3 CLUSTERING METHODS FOR FEATURE DETECTION AND CLASSIFICATION

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group is more similar to each other than to those in other groups called clusters. It is one of the image analysis techniques shown in Figure 4.4 and it involves in finding the presence of breast cancer. Microcalcification and mass are the group of same like cells (pixels) presented in digital mammogram image called suspicious areas and they can indicate the possible presence of breast cancer. Those areas can be segmented and classified by using this clustering technique. That is the reason for many researchers have used clustering is a main and common task for the techniques used in analysis of mammogram image. Radiologists usually use clusters to classify the TP and FP calcifications.

![Figure 4.4 Techniques involved in mammogram image analysis](image)

- Spatial features
- Transform Features
- Edges
  - And boundaries
- Shape features
- Moments
- Texture

- Template Matching
- Thresholding
- Boundary Detection
- Clustering
- Quad-trees
- Texture Matching
- Clustering
- Statistical
- Decision Trees
- Similarity Measures
- Minimum Spanning Trees
Cluster analysis can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with low distances among the cluster members, dense areas of the breast space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multiobjective optimization problem.

The appropriate clustering algorithm and parameter settings include values such as the distance function to use and density threshold or the number of expected clusters. It depends on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multiobjective optimization that involves trial and failure. Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy and typological analysis. The subtle differences are often in the usage of the results, while in mammogram image processing, the resulting groups are the matter of interest, in automatic classification primarily their discriminative power is of interest. The notion of a cluster cannot be precisely, which is one of the reasons why there are so many clustering algorithms. However, different researchers employ different cluster models and for each of these cluster models again different algorithms can be given. The notion of a cluster found by different algorithms varies significantly in their properties and understanding these cluster models is key to understanding the differences between the various algorithms. Typical cluster models include connectivity models, centroid models, distribution models, subspace models, group models and graph based models. A clustering is essentially a set of such clusters, usually containing all objects in the image. Additionally, it may specify the relationship of the clusters to each other. Clusterings can be roughly distinguished in hard clustering that is each
objects belongs to a cluster or not and soft clustering that is each object belongs to each cluster to a certain degree.

4.4 TYPES OF CLUSTERING

Six clustering methods have been discussed in this chapter. They are connectivity based clustering, centroid based clustering, distribution based clustering, density based clustering, mean shift clustering and Fuzzy C Means Clustering (FCMC). The detailed description of the clustering methods is as follows.

4.4.1 Connectivity Based Clustering

Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. As such, these algorithms connect objects to form clusters based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. Connectivity based clustering is a whole family of methods that differ by the way distances are computed. At different distances, different clusters will form. These algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. Apart from the usual choice of distance functions, the user also needs to decide on the linkage criterion. Since a cluster consists of multiple objects, there are multiple candidates to compute the distance use. Popular choices are known as single linkage clustering for minimum object distances, complete linkage clustering for the maximum of object distances or average linking. Furthermore, hierarchical clustering can be starting with single elements and aggregating them into clusters or starting with the complete data set and dividing it into partitions. While these methods are fairly easy to understand, the results are not always easy to use, as they will
not produce a unique partitioning of the image, but a hierarchy the user still needs to choose appropriate clusters form. The methods are not very robust towards outliers, which will either show up as additional clusters or even cause other clusters to merge especially in single linkage clustering. The complexity made the algorithm too slow for large image size. For more special cases, optimal efficient methods with less complexity can be used. They are Single LINKage (SLINK) by (Sibson 1973) and Complete LINKage (CLINK) by (Defays 1977). However, complexity is not considered if the clustering result is capable of grouping the required information.

4.4.2 Centroid Based Clustering

This method is otherwise called as K means clustering. Here, clusters are represented by a central vector, which may not necessarily be a member of the data set chosen for the mammogram image. When the number of clusters is fixed to K, K-means clustering gives a formal definition as an optimization problem which includes the K cluster centers and assigns the objects to the nearest center, such that the squared distances from the cluster are minimized. The optimization algorithm is the common approach that searches only for approximate solutions. A particularly well known approximative method is Lloyd’s algorithm (Lloyd 1982) also referred to as K means algorithm. It does however only find a local optimum and is commonly run multiple times with different random initializations. Variations of K means often include such optimixations as choosing the best of multiple runs, but also restricting the centroids to members of the data set called K medoids, choosing medians called K medians clustering, choosing the initial centers less randomly or allowing a Fuzzy cluster assignment. Most K means type algorithms require the number of clusters K to be specified in advance, which is considered to be one of the biggest drawbacks of these algorithms except for FCMC method.
Furthermore, the algorithms prefer clusters of approximately similar size, as they will always assign an object to the nearest centroid. This often leads to incorrectly cut boarders in between of clusters. This algorithm optimizes cluster centers and not clusters border. K means has a number of interesting theoretical properties and tends to find clusters of comparable spatial extent to have different shapes of clusters. Given a set of observations \((x_1, x_2, \ldots, x_n)\), where each observation is n dimensional real vector, K means clustering aims to partition the n observations into k sets \((k \leq n) \ s = \{s_1, s_2, \ldots, s_k\}\) to minimize the Within Cluster Sum of Squares (WCSS) = \(\arg\min_s \sum_{i=1}^{k} \sum_{j \in s_i} \|x_j - \mu_i\|^2\). Where \(\mu_i\) is the mean of points in \(s_i\). A more efficient version of K means method has been studied by Hartigan & Wong (1979). The most common algorithm uses an iterative refinement technique. Due to its ubiquity, it is often called the K means algorithm. It is also referred to as Lloyd’s algorithm, particularly in the image processing area. And it proceeds by alternating between two steps.

**Assignment Step:** Assignment of each observation is done to the cluster whose mean is closest to it. (That is, partition the observations according to the means value).

**Update Step:** Calculation is done for the new means to be the centroids of the observations in the new clusters.

The algorithm has converged when the assignments no longer change. The two key features of K means which make it efficient are often regarded as its biggest drawbacks. Euclidean distance is used as a metric and variance is used as a measure of cluster scatter. The number of clusters k is input parameter, an inappropriate choice of k may yield poor results. That is why, when performing K means, it is important to run diagnostic checks for determining the number of clusters in the data set. Convergence to a local minimum may produce wrong results.
A key limitation of K means is its cluster model. The concept is based on spherical clusters that are separable in a way that the mean value converges towards the cluster center. The clusters are expected to be of similar size, so that the assignment to the nearest cluster center is the correct assignment. It works well on some image data sets, while failing on others. K means implicitly assumes that the ordering of the input data set does not matter. The K means algorithm is an iterative technique that is used to partition mammogram images into K clusters. Steps of the basic algorithm is as follows,

1. Pick K cluster centers, either randomly or based on some heuristic.

2. Assign each pixel in the mammogram image to the cluster that minimizes the distance between the pixel and the cluster center.

3. Recompute the cluster centers by averaging all of the pixels in the center.

4. Repeat steps 2 and 3 until convergence is attained.

In this case, distance is the squared or absolute difference between a pixel and a cluster center. The difference is typically based on pixel intensity, texture and location, or a weighted combination of these factors. K can be selected manually, randomly or by a heuristic. This algorithm is guaranteed to converge, but it may not return the optimal solution. The quality of the solution depends on the initial set of clusters and the value of K.
4.4.3 Distribution Based Clustering

The clustering model most closely related to statistics is based on distribution models. Clusters can then easily be defined as objects belonging most likely to the same distribution. A nice property of this approach is that this closely resembles the way artificial data sets are generated by sampling random objects from a distribution. While the theoretical foundation of these methods is excellent, they suffer from one key problem known as over fitting, unless constraints are put on the model complexity. A more complex model will usually always be able explain the image data better, which makes choosing the appropriate model complexity inherently difficult. The most prominent method is known as Expectation Maximization (EM) clustering algorithm. Here, the data set is usually modelled with a fixed number of Gaussian distributions that are initialized randomly and whose parameters are iteratively optimized to fit better to the mammogram image data set to avoid overfitting. This will converge to a local optimum, so multiple runs may produce different results. In order to obtain a hard clustering, objects are often then assigned to the Gaussian distribution they most likely belong to and for soft clustering this is not necessary. Distribution based clustering is a semantically strong method, as it not only provides the clusters, but also produces complex models for the clusters that can also capture correlation and dependence of attributes.

4.4.4 Density Based Clustering

In density based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas that are required to separate clusters are usually considered to be noise and border points. This can be applied to mammogram image and works similar to that of linkage based clustering based on connecting points within certain distance thresholds. However, it only connects points that satisfy a density criterion, in
the original variant defined as a minimum number of other objects within this radius. A cluster consists of all density connected objects can form a cluster of an arbitrary shape plus all objects that are within these objects range. Another interesting property is that its complexity is fairly low and it requires a linear number of range queries on the image data base and that it will discover essentially the same results. It is deterministic for core and noise points, but not for border points in each run, therefore, there is no need to run it multiple times. The key drawback of this method is that, it expects some kind of density drop to detect cluster borders. Moreover this cannot detect intrinsic cluster structures in mammogram images and also in some real life images. A variation of this method has been discussed by Roy & Bhattacharyya (2005) for the efficient detection of such structures. The cluster border produced by these algorithms overlapping Gaussian distributions applied to artificial data produced the cluster border and it looks arbitrary because of the continuous decrease in cluster density. On a data set consisting of mixtures of Gaussians, these algorithms are nearly always outperformed by the methods such as EM clustering that are able to precisely model this kind of data. In recent years considerable effort has been put into improving algorithm performance of the existing algorithms. Various other approaches to clustering have been tried such as seed based clustering by Can & Ozkarahan (1990). For high dimensional data, many of the existing methods fail due to the curse of dimensionality, which renders particular distance functions problematic in high dimensional spaces.

4.4.5 Mean Shift Clustering

Basic mean shift clustering algorithm has been studied and investigated that, it maintains a set of data points the same size as the input mammogram image data set. Initially, this set is copied from the input set. Then this set is iteratively replaced by the mean of those points in the set that
are within a given distance of that point. By contrast, K means restricts this updated set to k points usually much less than the number of points in the input data set and replaces each point in this set by the mean of all points in the input set that are closer to that point than any other. A mean shift algorithm that is similar then to K means, called likelihood mean shift, replaces the set of points undergoing replacement by the mean of all points in the input set that are within a given distance of the changing set. One of the advantages of mean shift over K means is that there is no need to choose the number of clusters, because mean shift is likely to find only a few clusters if indeed only a small number exist. However, mean shift can much slower than K means and still requires selection of a bandwidth parameter. Mean shift has soft variants much as K means does.

4.4.6 Fuzzy C Means Clustering

Among the clustering algorithms discussed in this section, the K - means clustering algorithm is the best for grouping the information as clusters. But it cannot be able to find non convex clusters. Hence, it is suggested not to use this algorithm for the mammogram image processing for the selection of features. Fuzzy C Means Clustering (FCMC) algorithm can be able to find all type of clusters because of its ability in finding both convex and non convex clusters. It works based on the principle of, objects within a given cluster have a high degree of similarity and objects belonging to different clusters have a high degree of dissimilarity. Based on those criteria, the FCMC groups the features in different category as clusters. Objective function J of FCMC is given in Equation (4.1).

\[
J = \sum_{i=1}^{c} \sum_{j=1}^{N} U_{i}^{m} \left\| x_{j} - C_{i} \right\|^{1/m}, 1 \leq m \leq \alpha \tag{4.1}
\]
where \( m \) is any real number greater than 1, \( c \) is number of cluster, \( X \) is the \( i \) object of given \( N \) objects, \( U \) is the degree of membership of \( x \) in the cluster \( j \), \( C \) is centroid of cluster \( j \), \( \| \* \| \) is euclidean distance between any data object and the centroid. The parameter \( m \ (\geq 1) \) is called fuzzifier and signifies the amount of fuzziness in the solution set. And the algorithm is as follows.

Input: Dataset \( X \) of \( n \) objects with \( d \) features, value of \( K \) and fuzzification value \( m>1 \)

Output: Membership matrix \( U \) for \( n \) objects and \( K \) clusters

**Step-1:** Declare a membership matrix \( U \) of size \( n*K \).

**Step-2:** Generate \( K \) cluster centroids randomly within the range of the data or select \( K \) objects randomly as initial cluster centroids. Let the centroids be \( c \), \( c, \ldots, c \).

**Step-3:** Calculate the distance measure \( d=|x-c| \) using Euclidean distance, for all cluster centroids and data objects \( x,i=1,2,\ldots,n \).

**Step-4:** Compute the Fuzzy membership matrix \( U \) using equation (4.1)

**Step-5:** Compute new cluster centroids \( c \),

**Step-6:** Repeat steps 3 to 5 until convergence.

The resulted clusters have been grouped in to four different category based on the centroid levels. They are cluster 1, cluster 2, cluster 3 and cluster 4. The conclusion has been made from the resulted clusters based on the features appearance on the number of clusters. The results have been discussed and included in the results section.
4.5 RESULTS OF FUZZY C MEANS CLUSTERING METHOD

After the preprocessing of mammogram image and segmentation of RoI, the FCM clustering is applied to the RoI for the selection of textured features as clustered calcification or mass. Many clustering method has been discussed in this work. But FCM clustering is found to be more suitable for mammogram image analysis. Because, it process the RoI and divides it in to four sub images such a way that, each having its own clusters. If any bright spot found in it then it is going to be mass or tiny deposit of microcalcification which leads to be benign. This method has been tested for seven mammogram images. The resulted images are as follows.

![Image of RoI and Clusters]

(a) RoI  (b) Cluster 1  (c) Cluster 2  (d) Cluster 3  (e) Cluster 4

**Figure 4.5 Results of FCM clustering method for mdb001**

From Figure 4.5, (a) is the RoI input image, (b) is the cluster 1 output, (c) is the cluster 2 output, (d) is the cluster 3 output and (e) is the cluster 4 output. From the segmented RoI, it is found a bright spot in the centre of image mdb001. And it is presented only in cluster 3. Hence, it is concluded that the image may be having the chance of benign. The same result is available in the MIAS database. Similar to this, seven mammogram images have been tested in this thesis and the output of another six different images have also been given in the following figures.
From Figure 4.6, (a) is the RoI input image, (b) is the cluster 1 output, (c) is the cluster 2 output, (d) is the cluster 3 output and (e) is the cluster 4 output. From the segmented RoI, it is found a small bright spot in the centre of image mdb004. But it is not presented in all the four cluster outputs. Hence, it is concluded that the image does not have possibility of presence of cancer. The same result is available in the MIAS database.

From Figure 4.7, (a) is the RoI input image, (b) is the cluster 1 output, (c) is the cluster 2 output, (d) is the cluster 3 output and (e) is the cluster 4 output. From the segmented RoI, it is found a bright spot in the image mdb072. The same is presented in three cluster outputs with different brightness except cluster 1 output. Hence, it is concluded that the image has possibility of presence of malignancy. The same result is available in the MIAS database.
From Figure 4.8, (a) is the RoI input image, (b) is the cluster 1 output, (c) is the cluster 2 output, (d) is the cluster 3 output and (e) is the cluster 4 output. From the segmented RoI, it is found a bright spot in the image mdb075. The same is presented in cluster 3 and cluster 4 outputs with different brightness levels. Hence, it is concluded that the image has possibility of presence of malignancy. The same result is available in the MIAS database.

From Figure 4.9, (a) is the RoI input image, (b) is the cluster 1 output, (c) is the cluster 2 output, (d) is the cluster 3 output and (e) is the cluster 4 output. From the segmented RoI, it is found a bright spot in the top centre of image mdb218. The same is presented in cluster 1 output as small bright spot and
not presented in remaining three cluster outputs. Hence, it is concluded that the image does not have possibility of presence of malignancy but possibility of benign. The same result is available in the MIAS database.

![Image](image1)

(a) RoI           (b) Cluster 1       (c) Cluster 2      (d) Cluster 3   (e) Cluster 4

Figure 4.10  Results of FCM clustering method for mdb219

From Figure 4.10, (a) is the RoI input image, (b) is the cluster 1 output, (c) is the cluster 2 output, (d) is the cluster 3 output and (e) is the cluster 4 output. From the segmented RoI, it is found a bright spot in the top centre of image mdb219. The same is presented in cluster 4 output as small bright spot and not presented in remaining three cluster outputs. Hence, it is concluded that the image does not have possibility of presence of malignancy but possibility of benign. The same result is available in the MIAS database.

![Image](image2)

(a) RoI            (b) Cluster 1       (c) Cluster 2      (d) Cluster 3   (e) Cluster 4

Figure 4.11  Results of FCM clustering method for mdb322
Figure 4.11, (a) is the RoI input image, (b) is the cluster 1 output, (c) is the cluster 2 output, (d) is the cluster 3 output and (e) is the cluster 4 output. From the segmented RoI, there is no bright spot in the image mdb322. And there is no bright spot in all the four cluster outputs. Hence, it is concluded that the image does not have possibility of presence of cancer. The same result is available in the MIAS database.

4.6 HISTOGRAM BASED TEXTURE FEATURE DETECTION

Intensity transformation function based on information extracted from image intensity histograms plays vital role in mammogram image processing, especially in enhancement, segmentation and classification. Histogram of a digital image with N possible intensity levels in the range \([0, G]\) is defined as the discrete function \(h(r_k) = n_k\) where \(r_k\) is the k\textsuperscript{th} intensity level in the interval \([0, G]\) and \(n_k\) is the number pixels in the image whose intensity level is \(r_k\). Normalized histogram can also be used and it has been obtained by dividing all elements of \(h(r_k)\) by the total number of pixels \(n\) in the image \(p(r_k) = \frac{h(r_k)}{n} = \frac{n_k}{u}\). Where \(k = 0, 1, 2, ..., L-1\). And from basic probability, it is recognized that \(p(r_k)\) as an estimate of the probability of occurrence of intensity level \(r_k\).

The mass and microcalcification are detected from the mammogram image after the RoI extraction. Possible features to be considered from the detected mass are regenerative features and syntactic features. Boundaries, regions, moments and structures are the regenerative features. In mammogram image analysis for classification needs to measure any one of geometric attributes of the object such as geometry (perimeter, area, max-min radii or eccentricity, corners, roundness, bending energy, holes, Euler number and symmetry), moments (center of mass, orientation, bounding rectangle, best fit ellipse and eccentricity) and statistical moments (mean, standard deviation, smoothness, third moment, uniformity and
entropy). Calculation of statistical moments has been utilized for classification because of its ability to classify the textured features from masses or microcalcifications. Moments calculation has been achieved from the intensity histogram.

The expression for the \( n^{th} \) moment about the mean is

\[
\mu_n = \sum_{i=0}^{L-1} (z_i - m)^n p(z_i),
\]

where \( z \) is a random variable indicating intensity, \( p(z) \) is the histogram of the intensity levels in a region, \( L \) is the number of possible intensity levels and \( m = \sum_{i=0}^{L-1} z_i p(z_i) \) is the average mean intensity.

Uniformity and entropy are also the statistical approaches. Second moment \( \mu_2 \) is the variance \( \sigma^2 \). Average intensity, average contrast, smoothness, third moment, uniformity and entropy are the features have been calculated as moments from the intensity histogram. Expressions for the moments calculated using intensity histograms are tabulated in Table 4.2. The calculated moments and detected features have been presented in the results section of this chapter.

**Table 4.2 Descriptors of texture based on intensity histograms**

<table>
<thead>
<tr>
<th>S.No</th>
<th>Moment</th>
<th>Expression</th>
<th>Measure of Texture</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mean</td>
<td>( m = \sum_{i=0}^{L-1} z_i p(z_i) )</td>
<td>A measure of average intensity.</td>
</tr>
<tr>
<td>2</td>
<td>Standard</td>
<td>( \sigma = \sqrt{\mu_2} = \sqrt{\sigma^2} )</td>
<td>A measure of average contrast.</td>
</tr>
<tr>
<td></td>
<td>Deviation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Smoothness</td>
<td>( R = 1 - 1/(1 + \sigma^2) )</td>
<td>Measures the relative smoothness of the intensity in a region. R is 0 for a region of contrast intensity and approaches 1 for regions with large excursions in the values of its intensity levels.</td>
</tr>
</tbody>
</table>
Table 4.2 (Continued)

<table>
<thead>
<tr>
<th></th>
<th>Third Moment</th>
<th>Measure: (\mu_3 = \sum_{i=0}^{L-1} (z_i - m)^3 p(z_i))</th>
<th>Measures the skewness of a histogram. This measure is 0 for symmetric histograms, positive for histograms skewed to the right about the mean and negative for histograms skewed to the left.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Uniformity</td>
<td>(U = \sum_{i=0}^{L-1} p^2(z_i))</td>
<td>Measures uniformity. This measure is maximum when all intensity values are equal and decreases from there.</td>
</tr>
<tr>
<td>6</td>
<td>Entropy</td>
<td>(e = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i))</td>
<td>Measure of Randomness.</td>
</tr>
</tbody>
</table>

From the intensity histogram of the segmented outputs, statistical moments have been calculated and the values found for all the sample mammogram images tested in this work are tabulated in table 4.3.

Table 4.3 Measurement of texture features for the sample mammogram images

<table>
<thead>
<tr>
<th>Image Samples</th>
<th>Average Intensity</th>
<th>Average Contrast</th>
<th>Smoothness</th>
<th>Third Moment</th>
<th>Uniformity</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>mdb001</td>
<td>41.4323</td>
<td>44.8901</td>
<td>0.0812</td>
<td>0.3425</td>
<td>0.1655</td>
<td>4.6032</td>
</tr>
<tr>
<td>mdb004</td>
<td>65.0678</td>
<td>65.0032</td>
<td>0.0179</td>
<td>1.5012</td>
<td>0.1907</td>
<td>4.5219</td>
</tr>
<tr>
<td>mdb72</td>
<td>38.2165</td>
<td>42.3122</td>
<td>0.0542</td>
<td>5.4520</td>
<td>0.1005</td>
<td>3.9213</td>
</tr>
<tr>
<td>mdb75</td>
<td>45.3562</td>
<td>45.1204</td>
<td>0.0249</td>
<td>5.7034</td>
<td>0.1723</td>
<td>3.0476</td>
</tr>
<tr>
<td>mdb218</td>
<td>43.4235</td>
<td>73.5089</td>
<td>0.0624</td>
<td>0.2311</td>
<td>0.2564</td>
<td>4.1023</td>
</tr>
<tr>
<td>mdb219</td>
<td>50.5643</td>
<td>79.0076</td>
<td>0.0741</td>
<td>7.7022</td>
<td>0.4263</td>
<td>5.7643</td>
</tr>
<tr>
<td>mdb322</td>
<td>58.2365</td>
<td>68.3205</td>
<td>0.0235</td>
<td>4.5012</td>
<td>0.4303</td>
<td>3.7408</td>
</tr>
</tbody>
</table>
Based on the above measured parameters, the feature classification has been made and the classified results have been presented in the results chapter.

4.7 PROPOSED COMPUTER AIDED DETECTION METHOD

The main methods required to construct a CAD system for breast cancer detection has been discussed in the previous sections. All the proposed methods has been grouped and combined with the existing detection methods for the development of a CAD system for the breast cancer detection. The proposed algorithm consists of noise removal, contrast enhancement, segmentation and detection. Steps involved in the proposed CAD method are as follows.

Steps:

1. Input the digital mammogram image
2. Apply the curvelet based filtering for noise reduction.
3. Apply MLRM to noise removed image for the contrast enhancement.
4. Apply the Gaussian function to the MLRM enhanced image.
5. Apply Laplacian function to the step 4 gives the LoG of MLRM for the segmentation of RoI.
6. Apply the FCM clustering to the segmented RoI and find the clusters.
7. Classify the features selected from clusters as malignancy or benign based on the statistical moments calculation.
8. Stop the process, once the detection is over and input the next mammogram image.
4.8 FLOWCHART OF THE PROPOSED COMPUTER AIDED DETECTION METHOD

The algorithm discussed in the previous section has been given as a flowchart in Figure 4.12. It consists of all the methods discussed for CAD of breast cancer.

Figure 4.12 Flowchart of the CAD process adopted in the research
4.9 SUMMARY

Several clustering methods have been discussed for the feature selection from the RoI. FCMC has been chosen among the methods discussed because of its ability to process both convex and non convex clusters. Measurement of shape features, texture features and statistical moments have been studied and included in this section. Values for six features have been calculated and classification is performed based on the calculated values. Algorithm and flowchart for the methods developed to construct a CAD system has been presented.

The detection accuracy of the CAD has been found from the 105 sample mammogram images tested in this research work. The proposed CAD system works with 99% detection accuracy.