CHAPTER 5

MASS DIAGNOSIS USING REGION GROWING METHOD WITH
GLCM FEATURES AND SVM CLASSIFIERS

5.1 INTRODUCTION

This is the second approach developed for the mass diagnosis. Pitfalls of the first approach were kept in mind while developing this approach. So it can overcome the drawbacks of the previous system. In addition to the previous processes a separate Classification process was included in the implementation of the system. This approach has the following processes such as,

- Preprocessing
- Segmentation
- Feature Extraction
- Classification

For preprocessing the same median filter was used to remove the noise, in addition to that morphological operations are used to remove the labels and film artifacts in the mammogram. As in the previous approach basic normalization is performed to maintain the intensity values of the mammogram. But in addition to that Adaptive Histogram Equalization is implemented to increase the contrast of the mammogram. In the Previous approach combination of Gaussian and Gabor filter were used for segmentation. Since Segmentation has strong dominancy in improving the accuracy, after some fine evaluations region growing method was selected to be optimal to segment the mammogram. Same GLCM method was used for feature extraction. But unlike in the previous process GLCM features of
whole mammogram is extracted. Separate classification process was implemented in order to classify the mammogram in normal and cancerous according to the extracted feature values. Due to the inclusion of Classification process this approach is named as Malignancy Detection Technique with Classification (MDTC).

5.2 DATA COLLECTION

The data used in the experiments of the proposed work was taken from DDSM (Digital Database for Screening Mammography). DDSM database provides two different views such as Crasino Caudal view (CC) and Medio Lateral Oblique (MLO) view of left and right breast images. It contains 2620 cases acquired from Massachusetts general hospital wake forest University.

5.3 PREPROCESSING

Mammograms are medical images that are difficult to interpret. Hence preprocessing is essential to improve the quality. DDSM has digital mammogram database hence separate digitization process in not required. Digitization noise and high frequency components in the mammography images are removed by using median filter. The film artifacts such as label and x-ray marks are there in the mammograms of DDSM database which are unnecessary region in mass diagnosis. Sometimes the presence of these artifacts will affect the performance of the segmentation process. Hence removing these artifacts is the essential to improve the performance of the mass diagnosis system. These artifacts are removed using morphological operation and thresholding method with the use of MATLAB functions. In order to reduce the variation in brightness and to achieve computational consistency images are normalized by mapping all mammograms in to fixed intensity range. The original raw mammograms are shown in Fig 5.1.
Fig. 5.1 Original Raw Mammogram (MDTC)

The original raw mammogram is preprocessed with median filtering for noise removal, morphological operators for artifacts removal, Normalization to maintain the intensity values. The detailed explanation of Median filtering is given in 3.2.2.3. After preprocessing the original raw mammogram will be optimal for segmentation. The image of the original raw mammograms is given in Fig. 5.1. Image of the preprocessed mammograms is shown in Fig 5.2.

5.3.1 Contrast enhancement

Contrast enhancement can be performed by increasing the brightness. It adds intensity values by using adaptive histogram equalization over different segments. It adaptively enhances the contrast of each pixel relative to its local neighborhood which produces improved contrast for all levels in the image. Adaptive histogram equalization also helps to reduce the noise produced in homogenous area. The detailed explanation of AHE process is
given in 3.2.3.2. Mammogram after contrast enhancement process is shown in Fig 5.3.

**Fig.5.2.** Preprocessed Mammograms (MDTC)

**Fig.5.3.** Mammogram after Contrast Enhancement process
5.4 SEGMENTATION

The goal of segmentation is to find out the entire suspicious mass region from mammogram. A mass is space occupying lesion and usually appears as a bright region on a mammogram. So contrast enhancement is implemented in order to extract the brighter region.

5.4.1 Alarm pixel generation

Alarm pixels are produced by thresholding the contrast enhanced image. Alarm threshold is determined by histogram analysis. Segmentation through alarm pixel generation contains the following steps,

i. Histogram and accumulated histogram should be computed. \( H_{Fm} \) and \( AH_{Fm} \).

ii. Using histogram gradient changes location of peaks in histogram should be found out. \( (A_{g1}, A_{g2}, \ldots, A_{gi}) \) where \( A_{gi} \) are the gray levels.

iii. Candidate of alarm threshold is chosen by following condition,
\[
T_K = \{ A_{gi} \mid \text{When the selected alarm area < 10% of the entire region of interest}, k=p, p+1 \ldots q \},
\]
\( AH_{Fm} \) can be used to calculate the selected alarm area.

iv. Alarm threshold should be one of \{\( T_K; k=p \sim q \)\} \( i.e \ T_{Am}= T_1 \ p \leq l \leq q \), such that |\( A_{gi} - A_{gi-1} \) | is maximum among \{\( |A_{gk} - A_{gi-1}|;k=p \sim q \)\}.

v. Mark pixel at \( (x, y) \) as a candidate of alarm pixel if \( I_{Fm}(X,Y) > T_{Am} \) (m=1,2,3,4).

vi. A pixel at \( (x,y) \) is considered as alarm pixel if
\[
\sum_{m=1}^{4} I_{Am}(X,Y) \geq 4.
\]
Mammogram after alarm region generation process is shown in Fig 5.4.
Fig. 5.4 Mammogram after alarm Region generation process

5.4.2 Region growing

Region growing method seeks group of pixels with uniform intensities. Seeded region growing performs a segmentation of an image with respect to set of points known as seed. Alarm pixel generated from the above process can be considered as seed point. For the given seed points region growing method finds the tessellation of the image into regions with the property that each connected component of region meets exactly one of $A_i$. Each step of algorithm involves the addition of one pixel into above set. Let $z$ be the unallocated pixel as follows,

$$Z = \{x \in \bigcap_{i=1}^{n} A_i \mid N(x) \cap \bigcup_{i=1}^{n} A_i \neq \emptyset\} \quad (5.1)$$

Where, $N(x)$ is set of immediate neighbors of the pixel $x$. Consider the rectangular grid with immediate neighbors of 8 connected pixel $x$. if $x \in z$ then $N(x)$ meets just one of $A_i$. Hence $i(x) \in \{1, 2, \ldots, n\}$ to be the index such that
N(X) ∩ A_i(X) ≠ Φ. \hspace{1cm} (5.2)

δ(X) = |g(X) - \text{mean}_{g \in A_i(X)} [g(y)]|. \hspace{1cm} (5.3)

Where δ(X) is measure of how different x is from the region it joins and g(x) is the gray value of the pixel x. If N(x) meets two or more values of A_i then A_i will be selected according to the lowest value

\[ \delta(X) = \min_{x \in T} \{ \delta(X) \} \] \hspace{1cm} (5.4)

The above process is repeated until all the pixels have been allocated.

Mammogram after final segmentation is shown in Figure 5.5. The detailed explanation of Region growing is given in the section 3.3.4.

![Mammogram after final segmentation](image)

Fig.5.5. Mammogram after region growing segmentation

### 5.5 FEATURE EXTRACTION

Texture feature is useful in differentiating normal and abnormal pattern. Texture is an alteration and variation of surface of the image. Texture is characterized as the space distribution of gray levels in
neighborhood. There are two types of texture measures first order and second order. In the first order texture measure are statistics calculated from individual pixel. In second order relationship between neighbor pixels is considered. In the proposed method Gray Level Co occurrence (GLCM) matrix is used for feature extraction which comes under second order texture measure.

Second order statics can be used to model the relationship between pixels within the breast region by constructing GLCM matrix. A GLCM matrix is the joint probability of occurrences of gray levels $i$ and $j$ for the two pixels with a defined spatial relationship in an image. Spatial relationship is defined in terms of distance $d$ and angle $\theta$. GLCM matrix is constructed at a distance $d=1, 2, 3, 4$ and for angles $\theta=0^\circ, 45^\circ, 90^\circ$ and $135^\circ$. If the texture is course and distance $d$ is small then pair of points at distance $d$ should have similar gray levels. If the texture is fine and distance $d$ is comparable to the texture size then gray level of the two points would be different. Hence texture coarseness should be analyzed with various values of distance $d$. From GLCM matrices a variety of features may be extracted. Texture descriptors derived from GLCM are contrast, Energy, Homogeneity and Correlation which are calculated by using the following equations,

$$\text{Contrast} = \sum_{i,j=0}^{n-1} pij(i - j)^2$$  \hspace{1cm} (5.5)  

$$\text{Energy} = \sum_{i,j=0}^{n-1} (pij)^2$$  \hspace{1cm} (5.6)  

$$\text{Homogeneity} = \sum_{i,j=0}^{n-1} \frac{pij}{1+(i-j)}^2$$  \hspace{1cm} (5.7)  

$$\text{Correlation} = \sum_{i,j=0}^{n-1} (Pij) \frac{(i-\mu)(j-\mu)}{\sigma_1 \sigma_2}$$  \hspace{1cm} (5.8)  

$P_{ij}$ = Element $i, j$ of the normalized symmetrical GLCM
N is number of gray levels in the image

The GLCM mean is calculated as: \(\mu = \sum_{i,j=0}^{N-1} ij \)  \( P_{ij} \)  (5.9)

The variance of the intensities is calculated as: \(\sigma^2 = \sum_{i,j=0}^{N-1} P_{ij}(i - \mu)\)  (5.10)

Where,

Contrast is the contrast between a pixel and its neighbor.

Energy is the sum of squared elements in GLCM or uniformity.

Homogeneity is closeness of the distribution of elements in GLCM.

Correlation shows how correlated a pixel is to its neighbor over the whole image.

**Table 5.1 Result of feature extraction process (MDTC)**

<table>
<thead>
<tr>
<th>Image id</th>
<th>Image class</th>
<th>Homogeneity</th>
<th>Energy</th>
<th>Correlation</th>
<th>Contrast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mam1</td>
<td>Cancer</td>
<td>0.990</td>
<td>0.902</td>
<td>0.928</td>
<td>0.190</td>
</tr>
<tr>
<td>Mam2</td>
<td>Cancer</td>
<td>0.991</td>
<td>0.753</td>
<td>0.973</td>
<td>0.169</td>
</tr>
<tr>
<td>Mam3</td>
<td>Cancer</td>
<td>0.990</td>
<td>0.906</td>
<td>0.889</td>
<td>0.166</td>
</tr>
<tr>
<td>Mam4</td>
<td>Cancer</td>
<td>0.991</td>
<td>0.945</td>
<td>0.987</td>
<td>0.160</td>
</tr>
<tr>
<td>Mam5</td>
<td>Cancer</td>
<td>0.99</td>
<td>0.888</td>
<td>0.892</td>
<td>0.170</td>
</tr>
<tr>
<td>Mam6</td>
<td>Cancer</td>
<td>0.994</td>
<td>0.890</td>
<td>0.971</td>
<td>0.188</td>
</tr>
<tr>
<td>Mam7</td>
<td>Cancer</td>
<td>0.996</td>
<td>0.849</td>
<td>0.982</td>
<td>0.139</td>
</tr>
<tr>
<td>Mam8</td>
<td>Cancer</td>
<td>0.997</td>
<td>0.756</td>
<td>0.971</td>
<td>0.130</td>
</tr>
<tr>
<td>Mam9</td>
<td>Cancer</td>
<td>0.989</td>
<td>0.968</td>
<td>0.899</td>
<td>0.188</td>
</tr>
<tr>
<td>Mam10</td>
<td>Cancer</td>
<td>0.999</td>
<td>0.920</td>
<td>0.986</td>
<td>0.128</td>
</tr>
<tr>
<td>Mam11</td>
<td>Normal</td>
<td>0.888</td>
<td>0.466</td>
<td>0.750</td>
<td>0.190</td>
</tr>
<tr>
<td>Mam12</td>
<td>Normal</td>
<td>0.899</td>
<td>0.546</td>
<td>0.802</td>
<td>0.178</td>
</tr>
<tr>
<td>Mam13</td>
<td>Normal</td>
<td>0.854</td>
<td>0.396</td>
<td>0.869</td>
<td>0.194</td>
</tr>
<tr>
<td>Mam14</td>
<td>Normal</td>
<td>0.848</td>
<td>0.498</td>
<td>0.833</td>
<td>0.021</td>
</tr>
<tr>
<td>Mam15</td>
<td>Normal</td>
<td>0.798</td>
<td>0.591</td>
<td>0.840</td>
<td>0.110</td>
</tr>
<tr>
<td>Mam16</td>
<td>Normal</td>
<td>0.792</td>
<td>0.897</td>
<td>0.747</td>
<td>0.270</td>
</tr>
<tr>
<td>Mam17</td>
<td>Normal</td>
<td>0.784</td>
<td>0.919</td>
<td>0.880</td>
<td>0.131</td>
</tr>
<tr>
<td>Mam18</td>
<td>Normal</td>
<td>0.770</td>
<td>0.19</td>
<td>0.884</td>
<td>0.121</td>
</tr>
<tr>
<td>Mam19</td>
<td>Normal</td>
<td>0.875</td>
<td>0.786</td>
<td>0.869</td>
<td>0.194</td>
</tr>
<tr>
<td>Mam20</td>
<td>Normal</td>
<td>0.846</td>
<td>0.775</td>
<td>0.946</td>
<td>0.197</td>
</tr>
</tbody>
</table>
The detailed explanation about GLCM method is given in 3.4.1. Sample Results of feature extraction process for 20 mammograms are given in Table 5.1.

5.6 CLASSIFICATION

Classifiers are used in wider range for medical diagnosis. It helps to examine the medical data in shorter time and in more detailed manner. Over different type of classifiers Support Vector Machine produces perfect classification result in breast cancer diagnosis. Result obtained from SGLD matrix is given as input data to SVM classifier. SVM is a reliable classification technique based on statistical learning theory. SVM can classify the given data set into two separable classes \( \{1, -1\} \). SVM uses separating hyper plane to classify the classes. Training data is given as input to SVM classifier which consists of \( n \) datum \((x_1, y_1), .., (x^n, y^n)\), \( x \)

Separating hyper planes are performed as follows,

\[
D(x) = (w \cdot x) + w_0
\]  
\( (5.11) \)

The inequality \( y_i (w \cdot x_i) + w_0 \geq 1 \) is produced for both \( y=1 \) and \( y=-1 \) \( \mathbb{R}^n \), \( y \in \{1, -1\} \).

\[
Y_i [ (w \cdot x_i) + w_0 ] \geq 1 , i=1,..,n
\]  
\( (5.12) \)

If data points satisfy the above inequality condition then they form support vectors. Classification process is performed based on the support vectors. Margins of hyper plane obey the following inequality,

\[
yk * D(xk) / ||y|| \geq \Gamma, \ k=1,2,...,n
\]  
\( (5.13) \)

We can maximize the margin by minimizing \( w \) using eqn.(5.14),

\[
\Gamma \times w = 1.
\]  
\( (5.14) \)

In the case of non separable data slack variable \( \xi_i \) is added as follows,
\[ Y_i \left( w^* x_i + w_0 \right) \geq 1 - \xi_i \] 

(5.15)

In the case of non linear data, non linear input should be converted to high dimensional linear feature via kernels. In the proposed method RBF kernels are used

\[ K(x, x') = \exp\left(-\frac{\|x-x'\|^2}{\sigma^2}\right) \] 

(5.16)

Where \( \sigma \) is positive real number.

The detailed explanation about SVM classification method is given in 3.5.4. Experiments are conducted on the image taken DDSM database. 250 mammograms have taken for experiments in which 125 are normal and 125 are abnormal. 75% of images are used for training and 25% of the images are used for testing phase. Classification results for 250 mammograms are given in Table 5.2.

**Table 5.2 Classification result (MDTC)**

<table>
<thead>
<tr>
<th>Number of cases</th>
<th>True positive</th>
<th>True Negative</th>
<th>False Positive</th>
<th>False Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>108/125</td>
<td>120/125</td>
<td>17/125</td>
<td>5/125</td>
</tr>
<tr>
<td>Percentage</td>
<td>100</td>
<td>86.4%</td>
<td>96%</td>
<td>13.6%</td>
</tr>
</tbody>
</table>
5.7 PERFORMANCE EVALUATION

Perfect test method is one of the methods in ROC curve method. Perfect test method is used to evaluate the performance of designed algorithm. The result obtained from the classification process is given as input to partest method. According to that the sensitivity of this approach is 95.9% and Specificity is 86.2%. The accuracy of diagnosis is found to be 94.6%. Efficiency of the algorithm is decided by major factors TP, TN, FP, FN. Which are described as follows,

True Positive (TP): mass region is present and algorithm shows the same result
True Negative (TN): mass region is absent and algorithm shows the same result
False Positive (FP): mass region is absent and algorithm shows that mass region is present
False Negative (FN): mass region is present and algorithm shows that mass region is absent

The result is shown as graphical representation in Figure 5.6.
Fig. 5.6. Graphical representation of the result (MDTC)
Flow chart of the entire processes of this method is shown in Fig.5.7.
MDTC is the second approach developed to implement more accurate CAD system. This method is automatic and does not need any human interruption until getting classification result. Preprocessed image is segmented by using alarm pixel generation process in combination with seeded region growing. Segmented image contains the suspected region which is given for feature extraction process. Extracted features are classified into normal and abnormal region using support vector machine method. The performance of the proposed method is evaluated using perfect test method which gives the sensitivity and specificity of the result with graphical representation. The sensitivity of the proposed method is 95.9% and specificity is 86.2%. The accuracy of mass diagnosis is 94.6%. Hence

The proposed method is comparatively better than the previous approach. But even then the accuracy of diagnosis and efficiency of the system can be improved. The ultimate goal of this research work is getting the accuracy of diagnosis nearer to 100%. So some of the pitfalls found in this approach were used to design the flawless computer Aided Diagnosis as much as possible. Since segmentation is found to be brain of CAD system new segmentation algorithm was designed specifically for mammogram segmentation. Extracting the suspected region alone and processing the Region of Interest will help to get the improvised result. The new CAD system with the Implementation of new segmentation algorithm and inclusion of ROI extraction and it’s outcome is explained in the next chapter.