Chapter 5: Automated and Effective Content-Based Mammogram Retrieval Using Wavelet Based CS-LBP Feature and Self-Organizing Map

In this chapter, an automated, fast and effective content-based mammogram retrieval system is presented. The proposed pre-processing steps include automatic labelling-scratches suppression, automatic pectoral muscle removal, and image enhancement. Further, for segmentation selective thresholds based seeded region growing algorithm is introduced. Furthermore, we apply 2-level DWT on the segmented region and WCS-LBP features are extracted. Then, extracted features are fed to self-organizing map (SOM), which generates clusters of images, having similar visual content. SOM produces different clusters with their centers and query image features are matched with all cluster representatives to find the closest cluster. Finally, images are retrieved from this closest cluster using Euclidean distance similarity measure. So, at the retrieval time, a query image is searched only in a small cluster, reflects a superior response time and good retrieval performances as compared to traditional exhaustive search method. Descriptive experimental and empirical discussions confirm the effectiveness of this chapter.

5.1 Introduction

CBIR systems can assist more reliable diagnosis by classifying the query mammograms and retrieving similar past cases already annotated by diagnostic descriptions and treatment results, which help the radiologist to interpret and analyse the current case with historical cases. But mammograms are difficult images to interpret because it holds

labels, scratches, and pectoral muscles as shown in Fig. 5.1. These artifacts have high-intensity gray values and visible appearance is closer to abnormal mammograms, which misguide the existing segmentation algorithms, and they are unable to segment accurate pathology-bearing regions. Therefore, suppression of these is an essential pre-processing step.

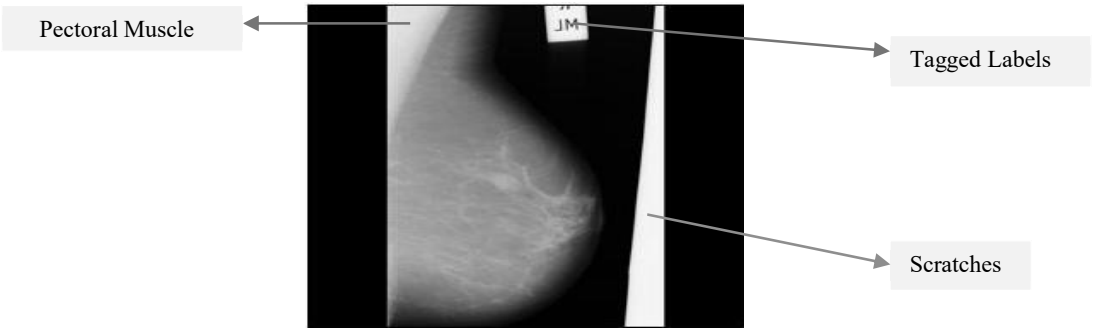


Fig. 5.1: Sample mammogram with artifacts

Generally, human interventions are involved in CBIR process, and for analysing the mammogram, the region of interest is manually cropped from the breast area for avoiding unwanted tags, labels, and pectoral muscles. But this manual pre-processing requires too much time and is expensive to implement. Most methods for mammogram image retrieval are based on this manual cropping and traditional retrieval [118-126], where traditional retrieval uses exhaustive linear search in the entire database. It compares and searches the query image from all images of the database, which slows down the response of retrieval.

Since the visual appearance of all the mammograms is significantly close to each other, therefore breast tumor segmentations are one of the most important and crucial stages in CAD system. Usually, pixels inside masses mammograms have high variant intensity. So, these characteristics tend to a region growing procedure to segment masses in mammograms [87-88]. Region growing method is a region-based

segmentation in which masses are segmented by grouping the pixels or sub-regions to get bigger regions, and this process continues until they satisfy certain intensity based criterion. In region growing, selection of initial seed points and appropriate value for the termination threshold is a more considerable problem because the wrong selection of these can affect the boundary information of the benign and malignancy masses [146-147]. In this work, we have taken highest intensity (pixel) as a seed point and proposed GLCM contrast based selective thresholds for the termination of region growing algorithm.

As we know that, texture is the most suitable descriptor for the mammograms. Being inspired by this fact, various approaches use GLCM, gray level aurora matrix, gray level neighbors matrix, and Gabor features for the retrieval [118-126]. For the capturing of texture representation of mammograms, we have taken our previous contribution of CS-LBP feature in wavelet domain.

Furthermore, as we have already mentioned, the size of image databases has been increasing rapidly, and making an exhaustive search in the whole database for similarity matching is very time consuming [23]. It is very important to increase the accuracy and narrow down the search space of retrieval. Using clustering, searching space is approximately reduced up to the number of images in the closest cluster. So, clustering without known predefined classes can play an excellent role in mammogram image retrieval, where we divide the samples into different groups such that similar images belong to the same cluster and dissimilar images in different clusters. In this work, we introduce SOM clustering approach [149]. SOM is unsupervised learning based artificial neural network (ANN) which map high- dimensional samples items onto a low dimensional grid of neurons. Using SOM, system performances are improved by the learning and searching capability of the neural network.

The working steps of this study are carried out in six steps viz.:

- Remove the artifacts and labels.
- Remove the pectoral muscles, and suppress the noises without human intervention.
- Proposed co-occurrence contrast based thresholds for the termination of region growing algorithm.
- Capture the texture characteristics by using proposed wavelet based CS-LBP features
- Apply self-organizing map (SOM) for clustering
- Retrieve similar mammograms relevant to the query image in less searching time.

This chapter is organized as follows; Section 5.2 presents the proposed methodology, Section 5.3 sheds some light on results analysis, and finally, Section 5.4 concludes the section with some further discussions.

5.2 Methods and Models

The proposed CBIR model is divided into two parts, off-line feature extraction and online image retrieval (shown in Fig. 5.2). In the component of off-line feature extraction, images are automatically pre-processed by removing the artifacts, pectoral muscles, and noises. Further, mammograms are segmented using proposed region growing approach. Then we apply the 2-level decomposition on the segmented mammogram and extract the CS-LBP features. These features are described as feature vectors of the images, constitute a feature dataset stored in the database. These features are fed to SOM for forming the different clusters. In the component of online image

retrieval, the user or the radiologist can submit a query image to the CBIR system to search for desired images having similar content. The system pre-processes and represents this query with another feature vector with same processes.

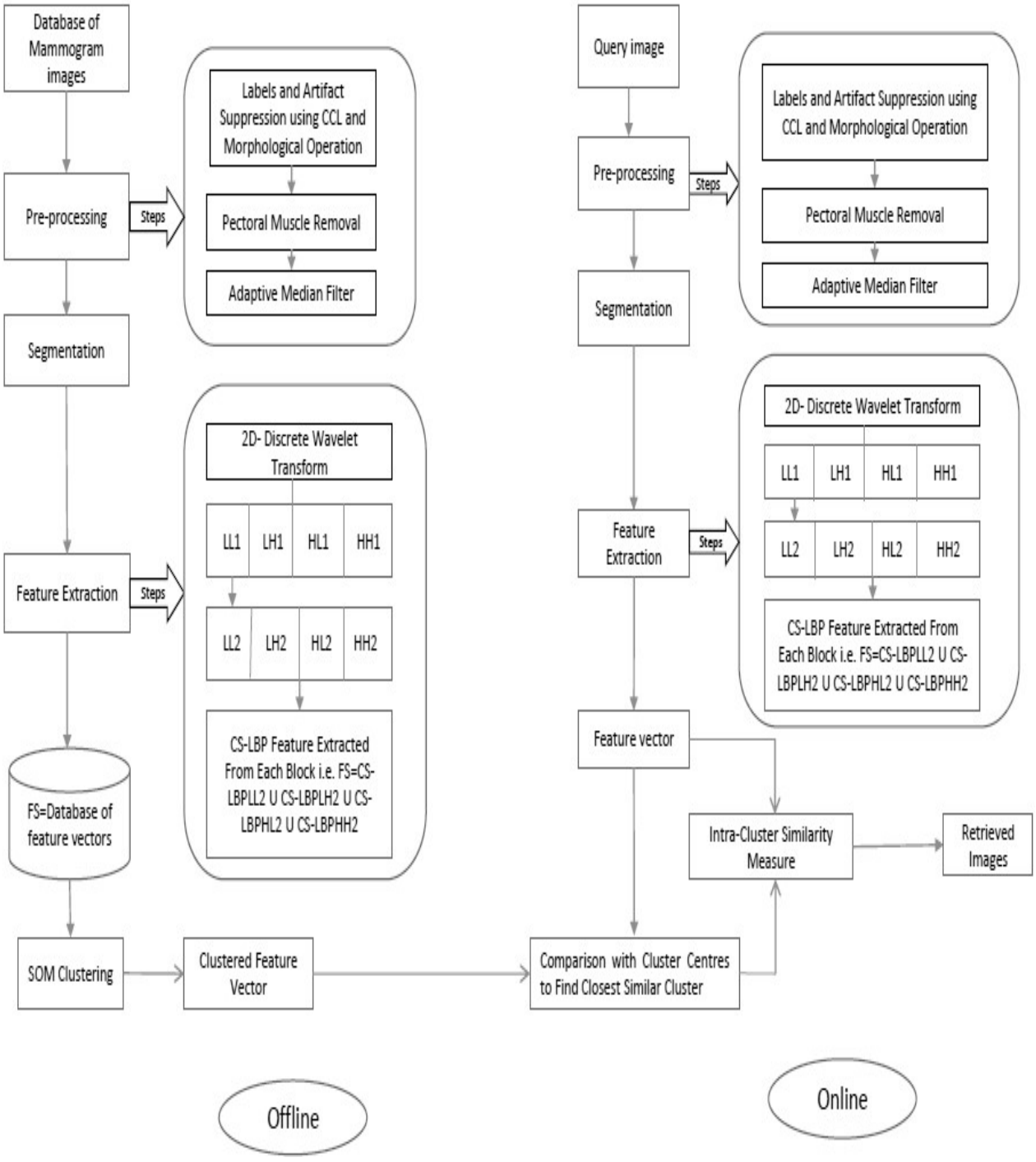


Fig. 5.2: Proposed mammogram retrieval framework

The similarities between the feature vectors of the query and those of the images in the closest cluster dataset are then computed and sorted. This similarity is computed using Euclidean distance similarity measure. After this system ranks the search results in non-decreasing order of the Euclidean distance and returns the results that are most similar to the query. A stepwise contribution of this work is discussed in the next sections.

5.2.1 Pre-processing

Pre-processing is necessary for improving the quality of images. It makes image segmentation and feature extraction phases more reliable. Many mammographic images are affected by artifacts, such as; tags, scratches, opaque marker, labels, and scanning artifact which are necessary to remove. Moreover, the dense region of pectoral muscle is closely related to abnormal density, which may affect the further mammogram analysis. Therefore, detection and elimination of pectoral muscle play an important role on mammogram pre-processing [88].

5.2.1.1 Label and Artefacts Suppression

This work has used morphological operations and connected component labelling for the removal of artefacts and labels. Before applying these, mammogram images are converted into binary images by using global threshold. This threshold is obtained by minimizing the intra-class variance of the black and white pixels, known as Otsu's method [150]. A stepwise solution for the removal of artefacts is given below.

Step-wise procedure for artefacts and label suppression:

1. Convert gray level mammogram images in binary using global threshold.
2. Flip the left Mediolateral Oblique view mammograms by 180º

3. Find the labels matrix that contains all connected regions.
4. Using label matrix, calculate the number of pixels for all regions.
5. Find the region with maximum connected pixels (largest area).
6. Remove all connected regions in the binary mammogram that have less number of pixels as compared to the maximum connected pixels.
7. After step-6, we get a mask for each mammogram.
8. Finally, pixel-wise multiply this mask with binary mammogram and remove all the scratches, labels, and other artifacts.

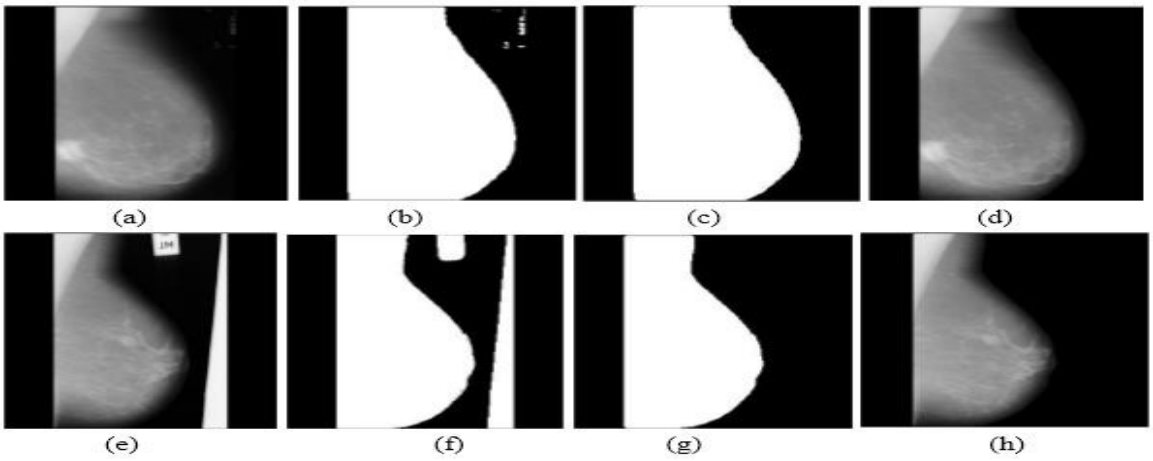


Fig. 5.3 (a-h): Artifacts suppression steps for two different sample images

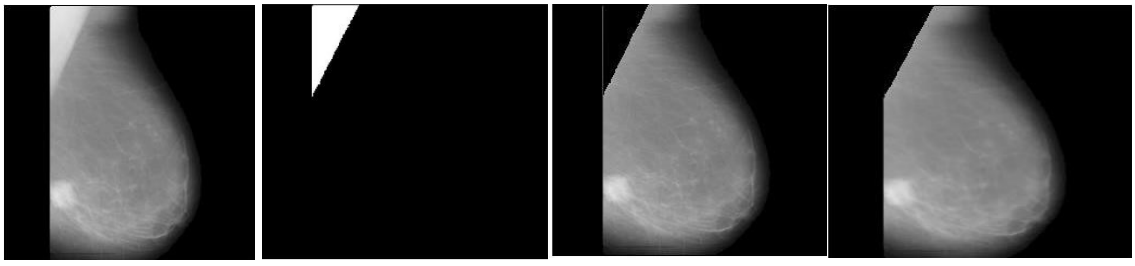
Fig. 5.3 (a-h) show the outcomes in succession of the suggested pre-processing steps on two sample images of MIAS database. Fig.5.3a displays the first sample image with label-marker in upper right corner. Fig. 5.3b presents the output of step-3, connected labels with different regions. Fig. 5.3c indicates the mask (maximum area morphological filtered image, output of step 7), and Fig. 5.3d exhibits the final label-marker and the other artifacts suppressed image. Fig. 5.3e demonstrates another MIAS database image with labels and scratches, which is pre-processed with same procedures and the final result is shown in Fig. 5.3h.

5.2.2.2 Pectoral Muscle Removal & Filtering of Mammograms

The pectoral muscle is a thick muscle, appears as a triangular opacity across the upper posterior margin [151]. In this work, we have used the adaptive k-means clustering for the removal of the pectoral muscle. Further, an adaptive median filter has been found to smooth the non-repulsive noises from 2D signals without blurring edges [141]. This filter also preserves image details hence used by the proposed work. A complete stepwise procedure for the removal of pectoral muscle and noises is given below.

Step-wise procedure for pectoral muscle removal and filtering:

1. Take the suppressed mammogram as an input, and find all the presented regions using adaptive k-means algorithm.
2. Then select the seed point by selecting any occurrence of non-zero intensity value in the pectoral region.
3. Using this seed point, we apply region-growing algorithm with very low threshold.
4. This region growing algorithm returns the segmented pectoral region.
5. Take the complement of segmented pectoral region.
6. Further, extract the segmented pectoral muscle region by pixel-wise multiplying the complement of segmented region with input mammogram.
7. Apply adaptive median filter to smooth the mammogram.



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

Fig.5.4 (a-d): Pectoral muscle removal and smoothing for same sample mammogram

Fig. 5.4 (a-d) show the further outcomes in succession of proposed pectoral muscles removal and smoothing steps. Fig. 5a shows the input image for this phase. Fig. 5.4b shows the segmented pectoral muscle, output of step-4. Further, the complement of this segmented region is pixel-wise multiplied with the input image, and the pectoral region is consecutively removed, output is shown in Fig. 5.4c, and finally Fig. 5.4d shows the filtered image.

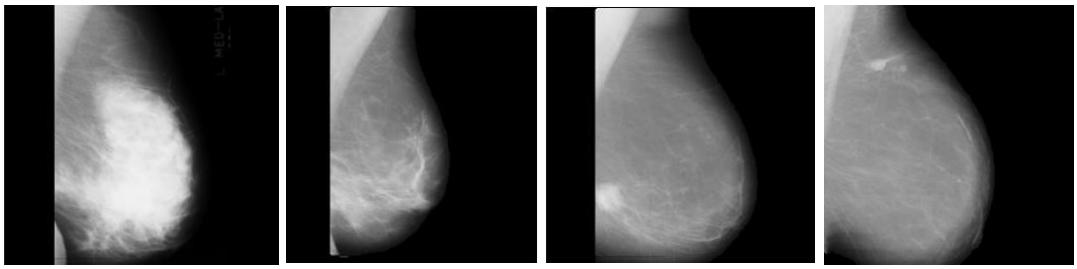
5.2.2.3 Mammogram Segmentation Using New Termintaion Criteria for Region-Growing

In this work, we have modified the seeded region-growing segmentation algorithm by designing new termination criteria. In our algorithm, we choose the brightest pixel in the remaining mammogram as the initial seed point. In many cases, it is found, this seed lies within the suspected lesion region. The similarity criterion used is based on intensity based gray level threshold. The region is grown until all the pixels for which intensity difference with seed is less than the threshold are added. Being motivated by Wei et al [19] where determination of the threshold was done experimentally in which pixel intensities of different mammograms were tested using different thresholds and then the final threshold was set. In this work, we have used another limiting criterion using gray level co-occurrence matrix [37]. Using this matrix, we have calculated the contrast, which measures the intensity between a pixel and its neighbour over the whole image. It reflects variations in local neighbourhood and it is 0 for areas with identical image and is high where there are large differences in gray tone. Contrast values are calculated using Eq. (5.1) help in finding the threshold values for the termination of region growing based segmentation.

$$Contrast = \sum_{i,j} |i - j|^2 \, G(i, j) \tag{5.1}$$

Where i, j are the indices for GLCM matrix and *G*(i, j) is corresponding co-occurrence values.

According to the mammographic society, there are three classes of background tissue mammograms- Dense (D), Glandular-Dense (GD), and Fatty (F). During experimental analysis on MIAS database, it has been noted that Dense and Glandular-Dense mammograms have larger brighter region (*Contrast>1.8*) as compared to Fatty mammograms. Fatty mammograms have limited bright region in specific position and *Contrast* usually lies between 0.5 to 1.8.



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

Fig. 5.5: Sample images based on background tissues

Fig. 5.5 shows the some sample mammograms from MIAS database, where 5.5 (a) shows Dense mammogram having *Contrast*=2.67, Fig. 5.5 (b) shows Glandular-Dense image having *Contrast*=1.85, Fig. 5.5 (c) and Fig. 5.5(d), show Fatty images with Contrast value=1.13 and 0.69. As per the proposed region growing algorithm, we choose the brightest pixel in the mammogram as the initial seed point, and the region is grown until all the pixels for which intensity difference with seed is less than the threshold. In order to prevent the region from growing indefinitely large (as in case of D and GD tissues), we have used limiting criterion by reducing the threshold value as t=0.13. For Fatty images, there are specific brighter region in small limited area, so we have used higher threshold t=0.20 for growing the algorithm in more region.

For the justification of used thresholds, we have taken 5 abnormal mammograms from each class (Dense, Glandular, and Fatty) and compared the segmentation performance for different thresholds. Segmentation performance of these thresholds is evaluated using Random Index [89].

- Random Index (RI)

The random index measure is used for the evaluation of segmentation and clustering algorithms. The random index between ground truth image (GT) and segmented image (S) is calculated by summing the number of pixel pairs with same label and number of pixel pairs having different labels in both S and GT, and then dividing it by total number of pixel pairs. RI values lie between 0-1, where higher value of RI indicates perfect segmentation.

Table 5.1: Segmentation performance for mammograms having *Contrast >1.8*

Image ids	Random Index on different threshold t						
	t=0.07	t=0.09	t=0.11	t = 0.13	t= 0.15	t=0.17	t=0.19
Mdb001(GD)	0.6312	0.5914	0.6912	**0.7921**	0.790	0.7663	0.7320
Mdb002(GD)	0.6918	0.7131	0.7231	**0.7863**	0.7428	0.7195	0.7076
Mdb013(GD)	0.6113	0.5677	0.**7376**	0.7057	0.6923	0.6717	0.6533
Mdb015(GD)	0.7155	0.7507	0.7737	**0.7337**	0.7655	0.7333	0.6912
Mdb017(GD)	0.6112	0.6867	**0.6960**	0.6697	0.6412	0.6345	0.6112
Mdb058(D)	0.6052	**0.6790**	0.6735	0.6576	0.6352	0.6218	0.6052
Mdb063(D)	0.5635	0.5672	0.4784	0.5993	0.5835	**0.6173**	**0.5982**
Mdb099(D)	0.6119	0.5731	0.5476	0.6567	**0.6719**	0.6434	0.5837
Mdb102(D)	0.6533	0.6173	0.6533	**0.6963**	0.6512	0.6323	0.5923
Mdb104(D)	0.5149	0.5518	0.5982	**0.6413**	0.6340	0.6131	0.5732
Average	0. 6209	0. 629	0.6572	**0. 6939**	0.68078	0.66532	0.6348

Table 5.2: Segmentation performance for mammograms having *Contrast <1.8*

Image ids	Random Index on different threshold t						
	t=0.14	t=0.16	t=0.18	t=0.20	t=0.22	t=0.24	t=0.26
Mdb5 (F)	0.6952	0.7199	0.7375	**0.7576**	0.7218	0.4952	0.5173
Mdb10(F)	0.6825	0.6671	**0.6784**	0.6493	0.5773	0.5035	0.4387
Mdb25(F)	0.5319	0.5763	0.5476	**0.5948**	0.5634	0.4519	0.3982

Mdb28(F)	0.6502	0.7287	0.7433	**0.7845**	0.7323	0.6020	0.5159
Mdb69(F)	0.6649	0.6718	0.6982	0.7314	**0.7417**	0.5649	0.5302
Average	0.64494	0.67276	0.681	**0.70352**	0.6673	0.5235	0.4801

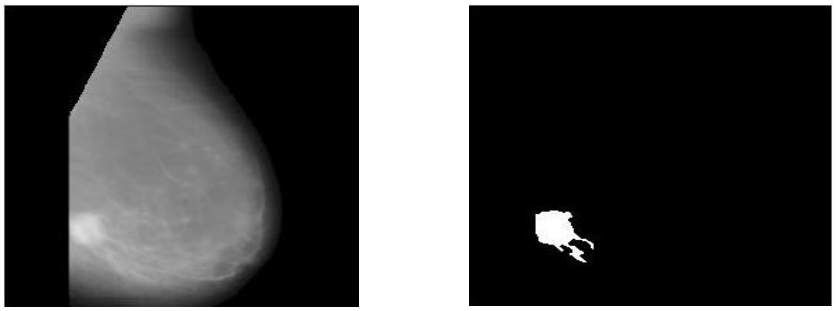
Table 5.1 shows comparison of random index values of various segmentation methods for 10 (5 D and 5 G) sample images. From this table, it is observed that the average value of random index for t=0.13 are higher than all other thresholds signifying that the used threshold is performing better in comparison to others. Table 5.2 shows comparison of random index values of various thresholds for five abnormal Fatty mammograms. From this table, it is observed that the average value of random index for t=0.20 are higher than all other thresholds, signifying that the used threshold is performing better in comparison to others. From both tables, it is clear that if we are increasing or decreasing the threshold values from t=0.13 and t=0.20, RI values are degraded. Therefore, these thresholds are taken for the termination of region growing algorithm.

Step wise procedure for segmentation based on modified region algorithm:

1. Take pre-processed mammograms as input.
2. Find the pixel (seed point) having the maximum intensity.
3. Calculate the *Contrast* of the mammograms using GLCM and Eq.-5.1.
4. Depending on the type of mammogram, the threshold t for the termination of region growing algorithm is determined.
5. If (*Contrast >1.8*) then threshold is taken as t = 0.13

   Else threshold value is set as t = 0.20
6. Start region growing algorithm using seed point and selected threshold t.
7. The new point, from 4-neighboring pixels, is added to the segmented region if the distance (mean of intensity difference) is less than threshold t.

8. Repeat the steps *6 to 7* until the distance between seed point and the neighboring pixel is higher than threshold t.



(a) (b)

Fig. 5.6 (a-b): Segmentation of sample mammogram

Fig. 5.6 (a-b) show the final outcomes of proposed segmentation for same sample mammogram. Fig. 5.6a shows the input mammogram for this phase, and Fig.5.6b shows the result of proposed segmentation.

5.2.2 Feature Extraction

Through multi-resolution it is possible to zoom the mammographic ROIs which improves texture visualization [90]. Getting inspiration from this concept, we propose a wavelet-based CS-LBP feature for efficient discrimination and retrieval of mammograms where discrete wavelet transform (DWT) is used to decompose the ROIs into a number of sub-images at different levels of resolution while preserving the low and high-frequency detail [89]. In this way, the most dominating texture description can be extracted from the ROIs of the mammograms by using the decomposition property of wavelets. As we know that, X-ray mammogram has strong edge distribution in the horizontal, vertical and diagonal directions (LH, HL, and HH) [33]. Therefore, this work has taken these detail coefficients (i.e. D1, D2, and D3) at 2-level decomposition and extracted CS-LBP features from each detail. i.e.

WCS-LBP(D1)=[F1 F2 F3……...F16],

WCS-LBP(D2)=[F17 F18 F19……..F32], and

WCS-LBP(D3)=[F33 F34 F35…….....F48]

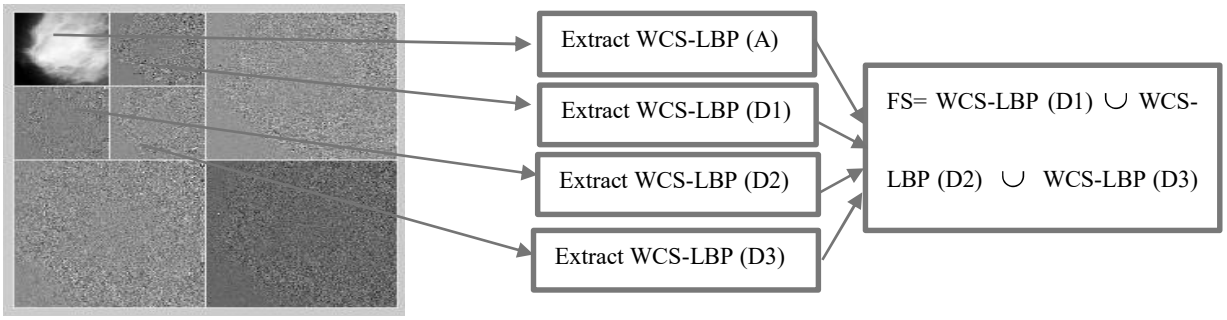


Fig. 5.7: Process of the proposed WCS-LBP feature extraction

Also, for taking the advantage of good energy compaction, we have also considered fine details from LL sub-band. So, extracted approximated CS-LBP feature, which is represented as:

WCS-LBP (A) = [F49 F50 F51…….F64]

So final wavelet based CS-LBP feature set:

FS= WCS-LBP (D1) ∪ WCS-LBP (D2) ∪ WCS-LBP (D3) ∪ WCS-LBP (A)

i.e. FS= [F1 F2 F3……F64]

Fig.5.7 shows the complete procedure for feature extraction. Final feature set (FS) is described as feature vectors of the mammogram and they are stored for the representation of the images. Algorithm 5.1 explains the complete procedure for pre-processing as well feature extraction.

Algorithm 5.1 Feature extraction

Ensure: *K* is the total mammograms (Input), FS is the output of this algorithm, shows final feature descriptor matrix .i.e. FS [1: K, 1: n], function Artifact() suppressed tagged label, scratches and other artifacts, pect() removes the pectoral muscle and filter

the noises, Seg() is the proposed segmentation which find the region of interest for further feature extraction. DWT() performs the discrete wavelet transform on each ROIs. Coef () holds three detail and one approximation coefficient matrix and finally function CS-LBP() extract the texture feature from each coefficient. FD is approximation (LL) and detail coefficient (LH, HL, HH) matrix and FM is feature matrix for each image. P is neighboring connectivity (P=8), l=2 is number of decomposed level of wavelet and r is the number of wavelet coefficient (4, i.e. LL, HL, LH and HH).

Steps

1. Initialize the required values for r, l, and p
2. $n \leftarrow r \times 2^{p/2}$
3. for i←1 to K do
4. $A_i \leftarrow Artifact(I_i)$
5. $P_i \leftarrow pect(A_i)$
6. $ROI_i \leftarrow Seg(P_i)$
7. for j←1 to l do
8. $WaveT_i \leftarrow DWT(ROI_i)$
9. for d←1 to r do
10. $FD_{j,d} = coef(WaveT_i)$
11. $CS\text{-}LBP_{j,d,k} \leftarrow CS\text{-}LBP(FD_{j,d})$
12. Compute CS-LBP and append to FM_{jd}
13. end for
14. end for
15. end for
16. FS[K, n]←concatenate (FMs)

5.2.3 SOM Clustering and Mammogram Retrieval

In this section, This work has used SOM for the clustering of mammogram samples where extracted dataset FS is fed to SOM algorithm and it clusters the mammograms based on the visual similarity.

SOM clustering algorithm for mammogram samples is as follow [149]:

Algorithm 5.2 SOM clustering

Ensure:

Input: feature vectors (FS) of *K* samples, each having length *n* and learning rate $\eta(t)$ which lie in the interval of $0 < \eta(t) < \eta(t-1) < 1$

Network Architecture:

Input layer: N=322 units (length of training vector)

Output layer: C=2 units (number of clusters), where each input unit is fully connected with weights to output unit.

Output: vector Y of length C (Y_1, Y_2…...Y_C).

Steps:

1. Select output layer network topology– Initialize current neighborhood distance, D (0), to a positive value.
2. Initialize weights vector to small random values w_j
3. Let t = 1
4. While computational bounds are not exceeded do (until feature map stop changing)
5. Select an input sample K_l. i.e feature vector of set FS.

6. Compute the square of the Euclidean distance of K_l from weight vectors (w_j) associated with each output node

$$\sum_{K=1}^{n}(I_{l,k}-w_{j,k}(t))^2$$

7. Select output node *j** that has weight vector with minimum value from step ii.
8. Update weights to all nodes within a topological distance given by D(t) from *j**, using the weight update rule:

$$w_j(t+1)=w_j(t)-\eta(t)(I_l-w_j(t))$$

9. Increment t
10. End while

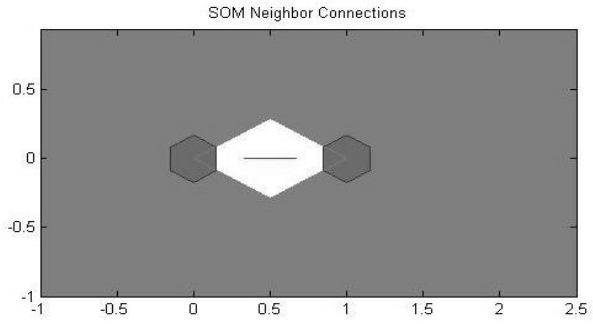


Fig. 5.8: MAP Structure and SOM Neighbour

The algorithm of SOM works as follows: Initially, A 2-D grid of neurons of a specified map size is pre-defined, which is hexagonal structure as shown in Fig.5.8. In this figure, the blue hexagons represent the neurons and the red lines connect two neighboring neurons. The distances between neurons are indicated by red line. For each sequential (subsequent) training iteration, an input data vector is randomly chosen from the feature set and sent through the network. The distance between its weight vector and the input vector is computed by each neuron of the network. Then find the winning neuron, which has the minimum distance between its weight vector and the input vector. Further, winning neuron and its neighbouring neurons weights are updated based on the

SOM learning rule. The number of iterations to train the SOM is usually set to a large value. Thus, the SOM is trained to cluster or classify an unknown input vector that is closest to the SOM's weight vector.

Here, we get an output vector Y of length *C* (Y_1, Y_2…...Y_C) where *C* (<K) is the number of clusters. Each of the *K* vectors in the training data is classified as falling in one of *C* clusters or categories. SOM produces cluster centers which are basically the weights assigned to the neurons by the algorithm. In the On-line phase, the user submits a query image to the CBIR system to search for images having similar content. The system extracts all of its contents (features) and represents it with a feature vector for the image. The similarity between cluster centers (Y_i) and feature vector decides the cluster to which the query image belongs. Once the desired cluster has been decided (based on Euclidean distance between cluster centers and feature vector) the intra-cluster similarity is calculated by the ED from the images in that cluster. The algorithm then indexes these images, and the system ranks the images in ascending order of the Euclidean distance and returns the topmost results that are most similar to the query image.

5.3 Result Analysis and Discussion

5.3.1 Searching Time Analysis

Searching of query image from large database of images is a challenging problem. In traditional (conventional) CBIR, searching methods calculate the similarity between the query image and all images of database. Further, all these images are ranked according to their ascending order of similarity values, and retrieved all the high ranked images. This exhaustive search is more time consuming. In this work, we have reduced the

searching time by narrow down the search space within the closest cluster. Analysis of searching time is given below.

Let T_1=calculation time for similarity measure between query image and all the images of database, and T_2=Sorting time for ranking of images.

So, retrieval time for traditional exhaustive search:

$$T = T_1 + T_2$$

(5.2)

Here, we have taken average case of quick sort for sorting and ranking of images.

So, for *K* number of images in the database, it takes:

$$T_2 = O(K \log K)$$

(5.3)

So $$T = KT_{1s} + O(K1 \quad K)$$

(5.4)

where T_{1s} is the required time to calculate similarity measure between two images. After the proposed clustering based retrieval framework, the retrieval time is:

$$T_C = CT_{1s} + CT_{1s} + O(c \log c)$$

(5.5)

where, C is the number of clusters, and c is the number of images in the closest cluster. As we know that in clustering, *K* numbers of images are divided into C clusters based on their similarities.

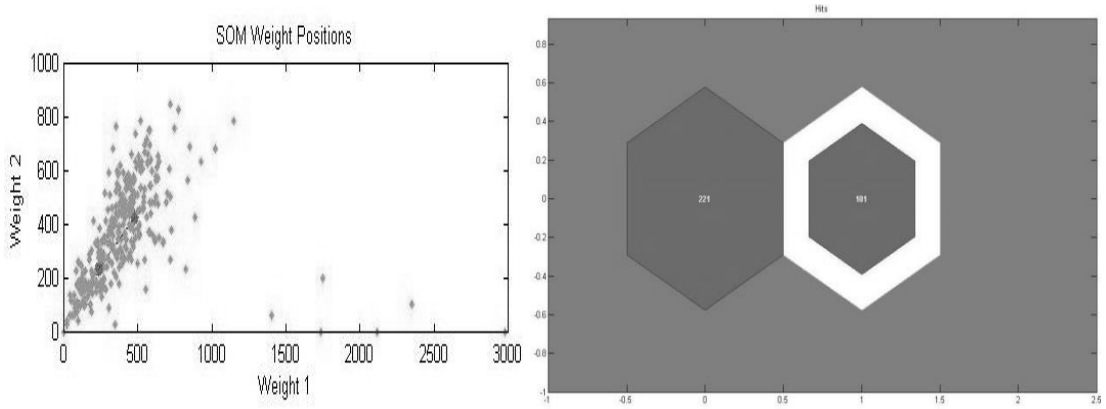
So C<<*K* and c<<*K*

Hence $T_C << T$

So, traditional exhaustive searching requires more searching time as compare to proposed clustering based approach, also may have scalability problem for big

databases. But proposed clustering based framework is easily scalable for large image databases and reflects good response time.

Fig. 5.9a shows, the SOM weight position where we plot the input vectors as green dots and shows how SOM classifies the input space by showing blue-gray dots for each neuron's weight vector and connecting neighboring neurons with red lines. Fig. 4.9b shows, SOM sample hits, a number associated with each hexagonal. Basically, these hexagonal are weighted neurons, which represent how many samples (images) are associated with each neuron. It is best if the data are fairly evenly distributed across the neurons. In this chapter, the used topology is a 2-by-1 grid, so there are 2 neurons, and maximum number of hits associated with a neuron is 221, reflects 221 samples (images) in that cluster. This work is tested on normal and abnormal mammograms, where smallest and largest clusters have 221 and 101 mammograms respectively.



(a) (b)

Fig. 5.9 (a-b): SOM weight position and sample hits

Proposed work, searching time improvement, as compare to exhaustive search can be estimated using *Speedup* metric. Table 5.3 reflects the retrieval *Speedup*, where we got the best *Speedup* of 3.19 times, if the query image belongs to the smallest (have minimum number of images) cluster. This value of *Speedup* says that proposed

framework searching time is 3.19 times faster than traditional CBIR. Further, worst and average searching times are 1.46 times and 2.33 faster than traditional CBIR.

Table 5.3: Retrieval speedup in various cases

Cases	Number of Similarity Comparisons	Speedup
Normal	221	1.46
Abnormal	101	3.19
Average	161	2.33

5.3.2 Retrieval Analysis

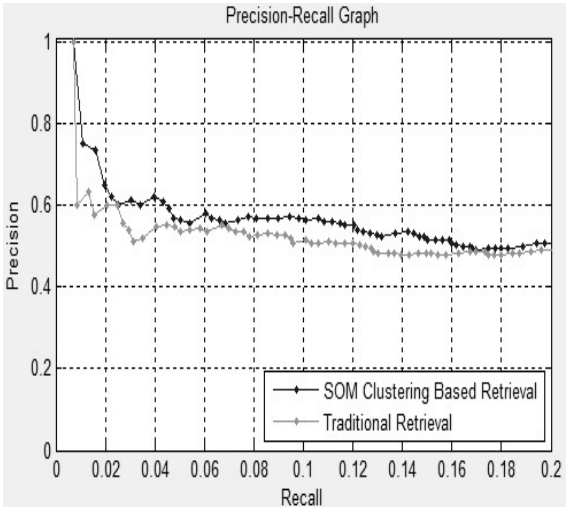
In the existing literature of mammogram image retrieval, generally the performances of retrieval system are analysed for abnormal and normal classes, in which retrieval of abnormal class is quite important. In order to analyse the retrieval performance such as average precision and recall, we have used the following criteria for abnormal query.

Table 5.4: Criteria for measurement of performance evaluation of CBIR

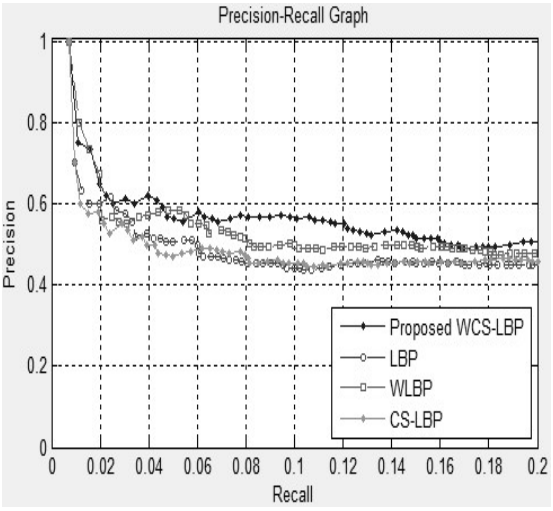
Score	Criteria
1	If retrieve mammogram belongs to class of query mammogram
0.5	The retrieved mammogram belongs to one of the abnormal classes, but not the class of query image.
0	The retrieved image does not belong to any abnormal class

For different comparative discussions for the retrieval, the experiments are carried out to randomly select 60 abnormal mammograms (10 images from each class)

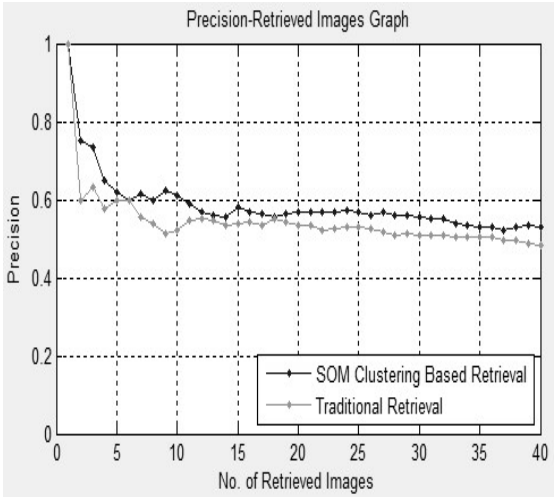
as the query images with the number of retrieved images set as 40, and performance measures are calculated as per the reported criteria in Table 5.4.



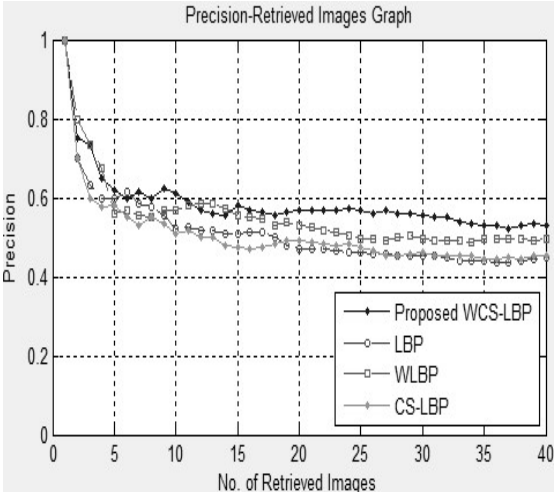
(a)



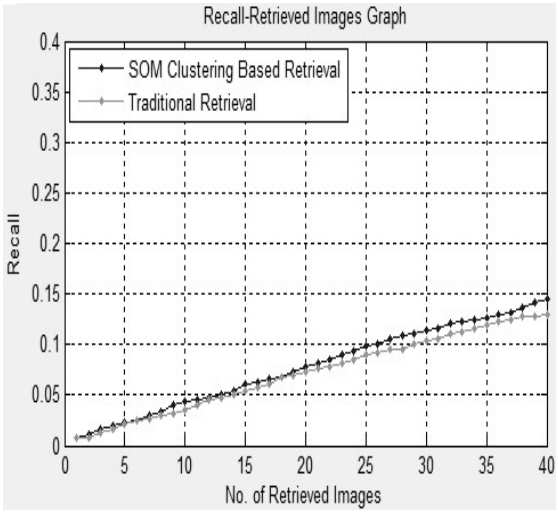
(d)

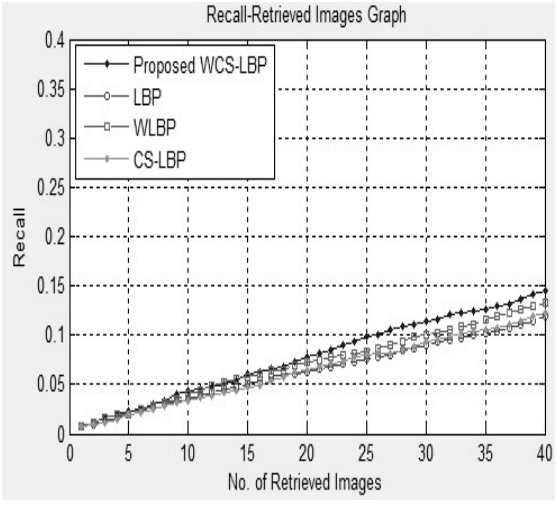


(b)



(e)





(c) (f)

Fig.5.10 (a-f): Different comparative retrieval for abnormal classes: Fig. 5.10(a-c) Traditional versus SOM Clustering and Fig.5.10 (d-f) SOM clustering and different variants feature set.

Fig. 5.11: Comparative analysis of feature dimensions

Fig.5.10 shows the retrieval performance in terms of following parameters, viz. P-R curve, Precision versus the number of retrieved images and recall versus the number of retrieved images. Fig. 5.10 (a-c) show the comparison with traditional (conventional) CBIR where searching is performed on all the images of the database. From these figures, it is observed that proposed clustering based retrieval approach perform significantly encouraging than tradition exhaustive search based retrieval. Also, it reduces the searching space as well as response time as earlier discussed. From the Fig. 5.10 (d-f) it is clearly visible that average precision and recall of proposed wavelet based centre local binary patterns are consistently encouraging than LBP, W-LBP and other variant features. Therefore, as compared to LBP and W-LBP, WCS-LBP is quite efficient in terms of retrieval performance as well feature dimension because the searching calculation of WCS-LBP is 16 times than faster than W-LBP. Feature dimension for three variant features is given in Fig.5.11. Feature dimension for three variant features is shown in Fig.5.11. We have also tested the retrieval performance for the normal mammogram, where we have taken 10 random healthy mammograms as a

query. Fig. 5.12 shows the average precision of abnormal and normal class for 10 number of retrieved mammograms. From the Figure, it is observed that average precision of normal class mammogram is outstanding. It gives maximum 98 % average precision for proposed WCS-LBP features followed by 87 %, 91% and 96% for LBP, CS-LBP, and W-LBP features respectively. Here, we notice out the good retrieval performance for the normal mammogram, because normal mammogram belongs to the single class and quantities of normal mammograms are just double, while abnormal mammograms retrieval performance depends upon the pre-processing-segmentation steps. Further, Abnormal mammograms have six different classes with a limited number of mammograms.

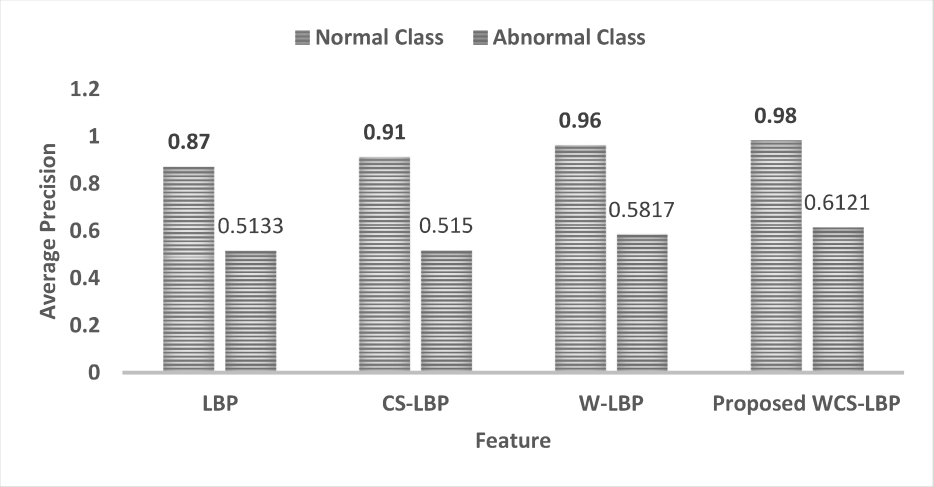


Fig. 5.12: Average precision for normal and abnormal mammograms

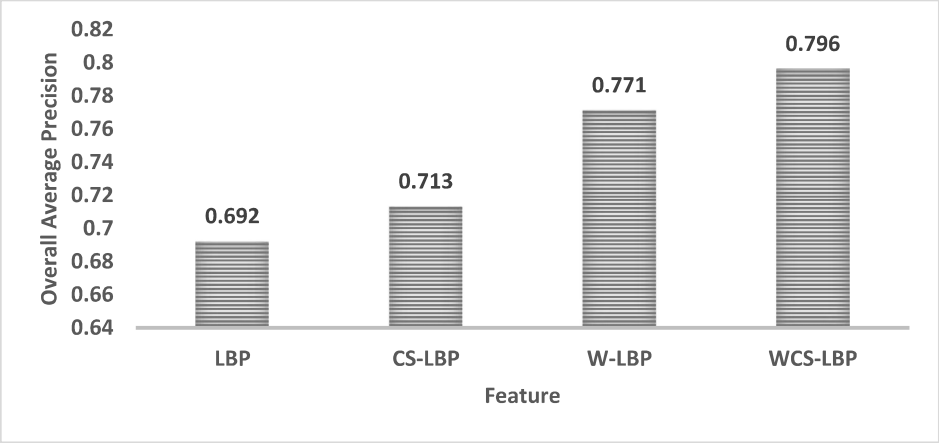


Fig. 5.13: Overall mean of average precision for the proposed framework

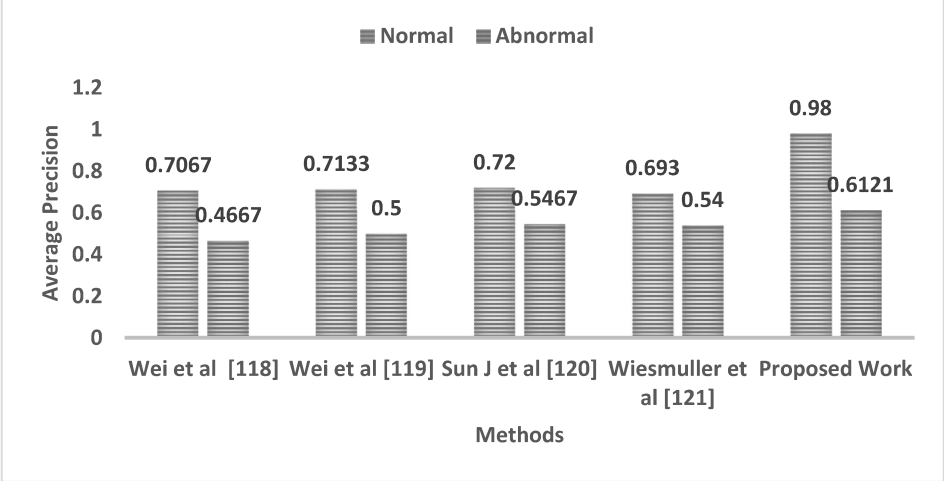


Fig. 5.14: Comparative analysis for normal and abnormal query

From the Fig. 5.13, it is clear that the overall average precision of proposed approach is better by 6.935%, 4.855%, and 1.515 %, with respect to the overall average precision of approaches introduced by LBP, CS-LBP, and W-LBP. In order to compare this work with other state-of-art methods, average precision rate is calculated for 10 images of retrieval. In this finding, this work has taken 20 independent mammograms (10 from each class) as different queries; belongs to the Normal and Abnormal classes. According to the reported results in Fig. 5.14, the obtained average precision for normal

and abnormal classes for proposed work is significantly encouraging than other state-of-art methods.

Fig. 5.15(a-e) show; the snapshot of image retrieval for five different queries. It gives the glimpse of effectiveness for the proposed work. In Fig. 5.15a query of Fatty abnormal mammogram is shown where 9 relevant images are retrieved. In Fig.5.15b retrieval of Ill-defined masses (malignant) is shown where 7 retrieved images belong to the class of abnormal or suspicious categories.

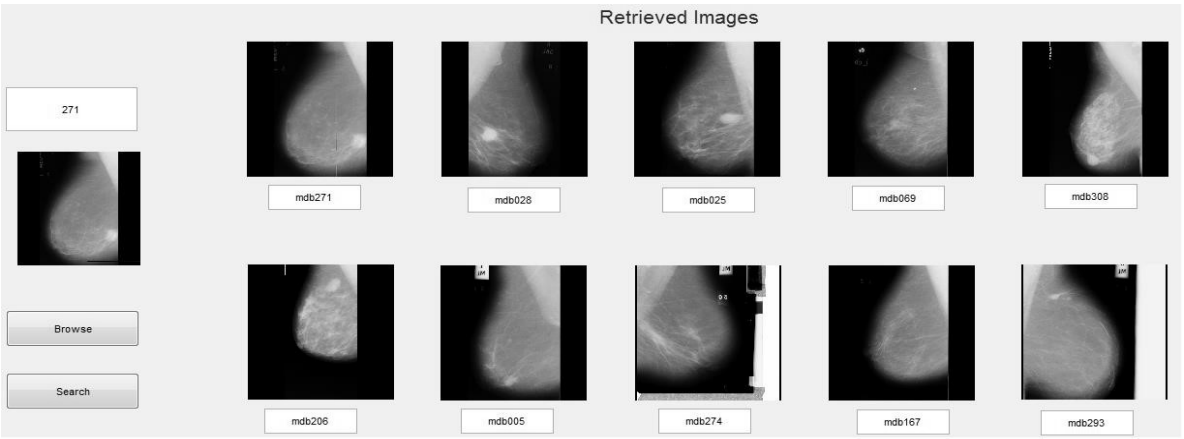


Fig. 5.15a. Image retrieval for Fatty query

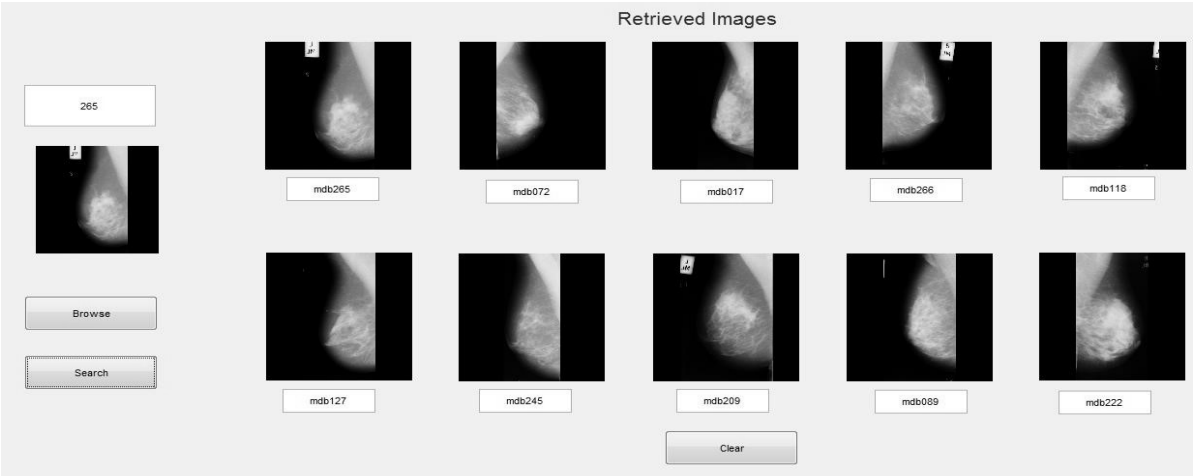


Fig. 5.15b. Image retrieval for Ill-defined masses query

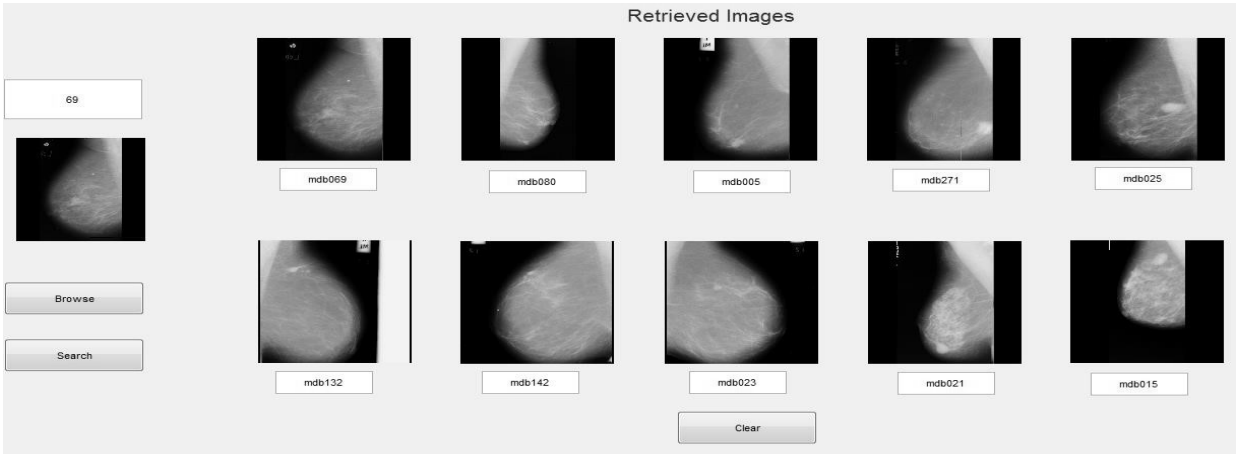


Fig. 5.15c. Image retrieval for Circumscribed masses query

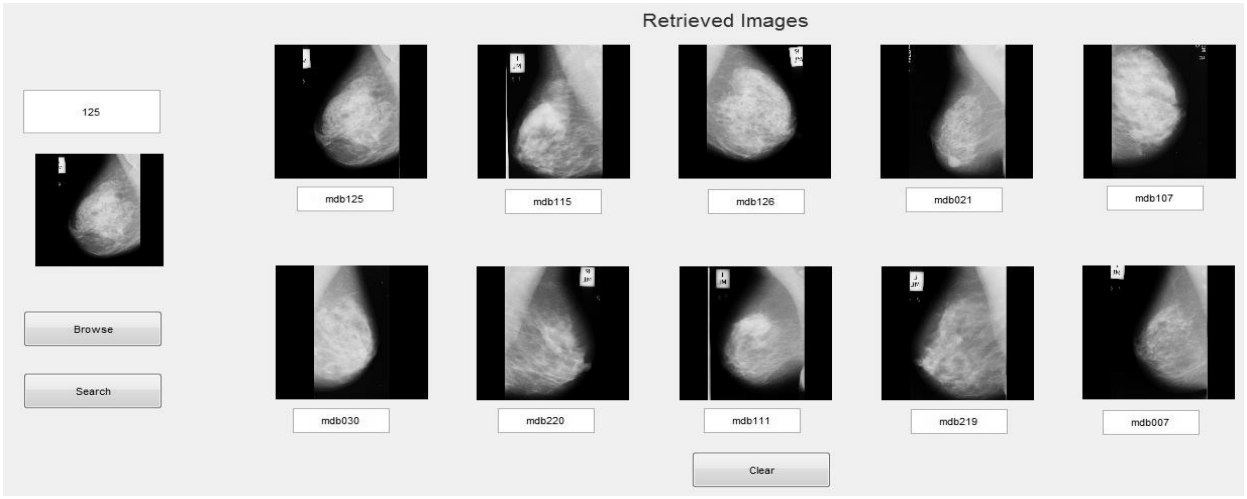


Fig. 5.15d. Image retrieval for Architectural distortion query

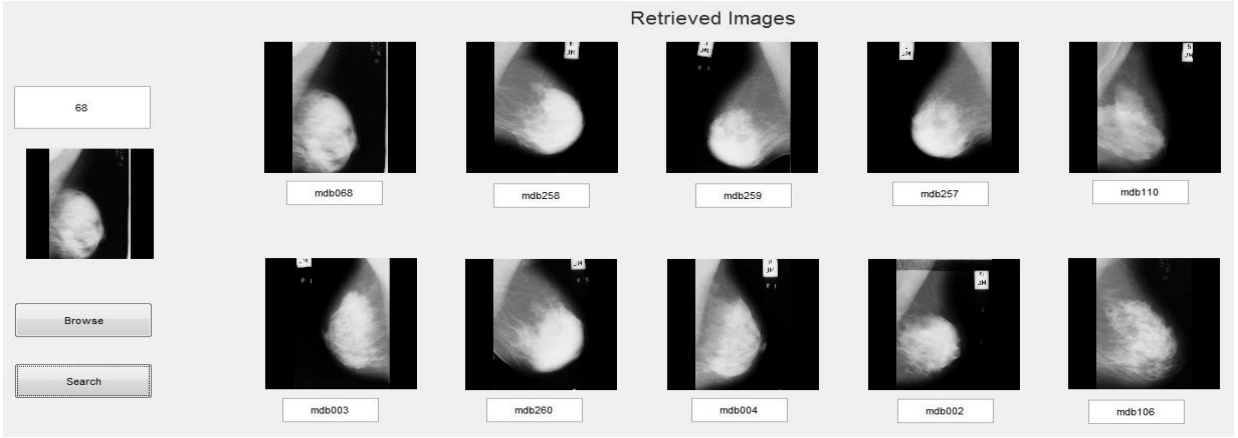


Fig. 5.15e. Image Retrieval for Dense mammogram

Fig.5.15 (a-e) Sample retrieval form different classes: (a): Fatty, (b): Ill-defined masses, (c): Circumscribed Masses, (d): Architectural Distortion, and (e): Dense classes

Fig. 5.15c shows the retrieval for suspicious circumscribed mass (benign). In this retrieval, all the retrieved images are relevant to the query. Fig. 5.15d shows the retrieval for abnormal (malignant) class, which belongs to the architectural distortion, where out of 10, eight images are correctly retrieved. Further, Fig. 5.15e shows the retrieval of normal dense query where 9 mammograms are correctly retrieved. These sample queries and corresponding retrieval, confirmed the effectiveness of this work.

However, the proposed framework outperforms for SOM with 2 clusters. Further, beware the fact that this work searching time is not dependent on image database, it depends on the cluster size. So, adding the new class of images will not affect our retrieval time. Furthermore, this work drastically reduces the exhaustive searching time and reflects a superior response time with good retrieval performances.

5.4. Conclusions

CBIR system fetches relevant mammograms based on visual similarity of the current mammogram, which is very beneficial for early diagnosis of breast cancer. There are several factors, like, image acquisition parameters, exposure time and energy level are responsible for increasing the difficulty during segmentation and influencing the quality of mammogram. This work resolved these issues by using the application of connected component labelling, morphological image processing, adaptive k-means and seeded region growing algorithm. There are four key contributions of the work presented in this chapter. Firstly, fully automated removal of artifacts and pectoral muscles. Secondly, modified region growing for effective segmentation which provided perfect breast contour representation for breast profile region. Further, designed wavelet-based CS-LBP features captured the strong texture characteristics and finally introduced SOM clustering based retrieval framework. This framework retrieved the most relevant

mammograms in less searching time as compared to the traditional exhaustive search method.