Chapter 4: Effective Mammogram Classification-cum-Retrieval Based on Center Symmetric-LBP Features in Wavelet Domain Using Random Forests

In this chapter, firstly an effective method for the classification of mammograms using random forests and wavelet based center-symmetric local binary pattern (WCS-LBP) is presented. For the classification of mammograms, multi-resolution CS-LBP texture characteristics from non-overlapping regions of the mammograms are captured. Further, we examine most relevant features using support vector machine- recursive feature elimination (SVM-RFE). Finally, we feed the selected features to decision trees and construct random forests which are an ensemble of random decision trees. Using wavelet based local CS-LBP features with random forest, we classify the test images into different categories having the maximum posterior probability. The proposed method shows the significantly encouraging classification performance as compared with other variant features and state-of-art methods. The obtained performance measures are 97.3 % accuracy, 97.3 % precision, 97.2 % recall, 97.2 % F-measure and 94.1 % Matthews correlation coefficient (MCC). Further, for the mammogram retrieval, training of the random forests is done with a small set of known ground truth annotation, and after training, classifier works as a manager which categorizes the remaining mammograms into different libraries (i.e. Normal or Abnormal). However, at the query time, same features are extracted and fed to the random forests. This classifier detects the class of query and searching is automatically narrowed down to the corresponding library or cluster. This mammogram classification and retrieval model

filters out maximum irrelevant images and retrieves most relevant mammograms in less searching time.

4.1 Introduction

Mammogram classification is a crucial and challenging problem, because it helps in early diagnosis of breast cancer and supports radiologists in their decision to analyse similar mammograms out of a database by recognizing the classes of current mammograms. CBIR techniques to diagnose the stage of breast cancer, because it supports radiologists in their decision to find out similar historical cases from the pre-stored database, and help radiologist for comparing the current case with past cases [87-88]. However, image analysis is not supported by CBIR system, creates a need to classify the mammograms manually and then annotate it by taking the help from pathologists, doctors or medical experts [91]. This manual classification of overwhelming number of mammograms is repetitive, labour intensive, and requires medical experts. In addition, due to the experimental situations (different image quality and human working subjectivity) classification results can lead to misclassifications and are not reliable [140]. Thus, for resolving the limitations of manual classification, various content (features) based mammogram image classification and retrieval methods have been proposed. Unlike text-based retrieval and manual classification, content-based approaches index images using texture, colour, and shape. Due to the complexity of the pre-processing operation, CBIR for mammogram database faces a considerable computational burden. Therefore, mammogram classification is important because it categories the images into different classes based on its appearance, and treated as a pre-processing step to speed up the accuracy, searching and retrieval

performances of CBIR systems. Also, mammogram classifications can also help for the early diagnosis of breast cancer by detecting cancer in the mammogram.

The success of accurately classifying mammograms depends on what features that are extracted from mammograms and fed into the learning model. Hence, we need to transform processed images into features that are better representing the task of classification. As we know that mammogram visual appearance are much close to texture based appearance [88]. So in this study, we have proposed wavelet based CS-LBP (WCS-LBP) texture feature which dominantly captures all abrupt edges and fine details of the mammograms. CS-LBP is extended version of the LBP based on the concept of center symmetric [139], capable of effective classification for the normal and abnormal mammogram. Further, feature selection is a very relevant step in creating an accurate predictive classifier. It can be used to identify and remove irrelevant and redundant features from data that do not contribute in discriminating among normal and abnormal mammograms or may, in fact, decrease the accuracy of the model [87]. Fewer attributes are also desirable because they reduce the speed and complexity of classifier training and execution. Moreover, the normal and abnormal mammograms contain some similar characteristics. Hence, it is practical to eliminate similar features between normal and abnormal mammograms. So, we examine most relevant features using support vector machine- recursive feature elimination. Further, for successful diagnosis of breast cancer, the generalization capabilities of the classifier are most essential. So, to overcome the over-fitting problem we have used random forests classifier [133].

From the experiments, it has been found that wavelet based CS-LBP performance is more promising than wavelet based LBP, and also it has been shown that the CS-LBP descriptor is faster than LBP. There are huge differences in the sizes of both descriptors. Although, being small feature dimension, the classification

performances of WCS-LBP is much effective than wavelet LBP. The performance of the proposed mammogram classification framework is also evaluated and compared with other types of similar work reported in the literature. Further, efficacy of this classification framework is used for the retrieval of mammograms. This chapter work is organized as follows; Section 4.2 presents the proposed method and models, Section 4.3 sheds some light on experimental analysis and finally, Section 4.4 is concluding section with some further discussions.

4.2 Methods and Models

The work presented in this chapter are divided into two sections. Section 4.2.1 gives the complete framework for mammogram classification using WCS-LBP and random forests. In section 4.2.2, this work uses the same extracted features and classification framework for the ease of mammogram management and retrieval.

4.2.1 Proposed Models for Mammogram Classification

The proposed mammogram classification framework is shown in Fig. 4.1, where mammograms are firstly pre-processed by suppressing the artifacts like labels, pectoral, and noises. Further, mammograms are decomposed up to 2^{nd} level using DWT, which conserve the edges information and energy details of mammogram in detail and approximate coefficients. Furthermore, approximate coefficient and sum of all details coefficient are divided into four non-overlapping blocks and CS-LBP based texture feature are extracted from each block. However, to select the best features SVM-RFE [144] feature selection method is applied, and finally, random forests is used for the classification of mammograms. Descriptive details of this work are given in the sections 4.2.1.1 to 4.2.1.5. Finally, using the efficacy of this mammogram classification system,

this chapter introduces the classification cum retrieval framework which is discussed in the next section 4.2.2.

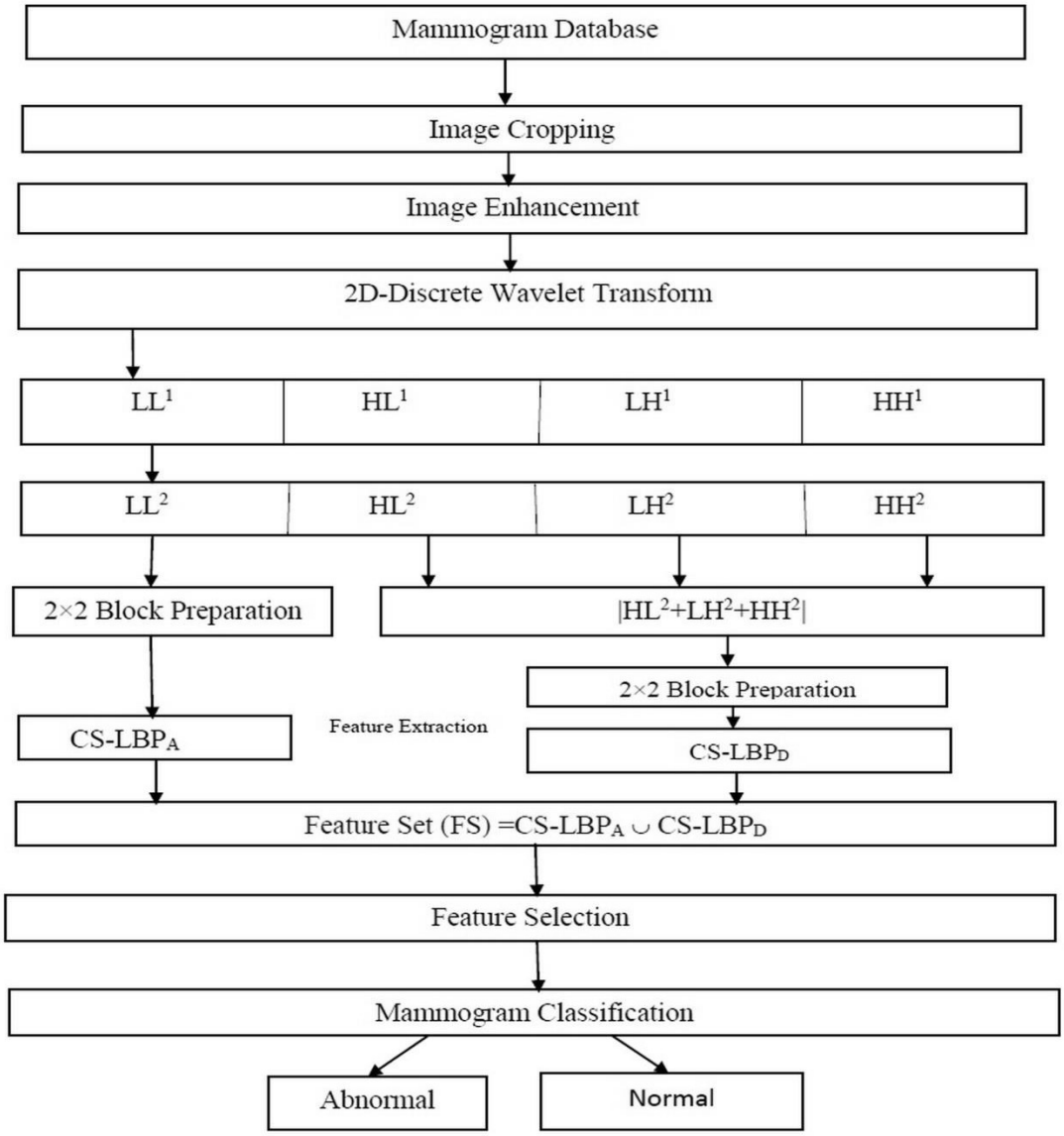

Fig. 4.1: Proposed framework for mammogram classification

4.2.1.1 Image Cropping & Enhancement

The first step in the design and development of an automatic CAD tool for the classification of mammogram, is pre-processing of mammograms which include cropping of ROI, because most of the images comprised of the background with artifacts such as scratches, pectoral muscles, and tagged-labels which may interfere in

proper extraction of features and result in erroneous classification. Hence, a cropping operation is applied to the mammograms to cut off the unwanted portions of the images. For the extraction of ROIs from the abnormal mammogram, we have taken the marked ground truth details. i.e. center of the abnormal area and corresponding radius, and for normal ROIs, we have cropped any random location from normal mammograms.

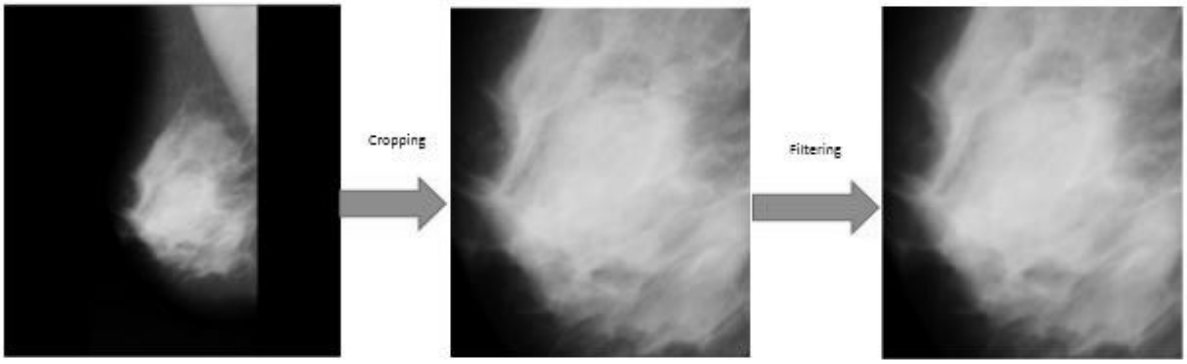

Fig. 4.2: Cropping and filtering of a sample mammogram

Further, Adaptive median filter has been found to smooth the non-repulsive noise and lines from 2-dimension (2D) X-ray images without blurring edges [141]. So this is particularly suitable for enhancing X- ray mammograms images, hence used by proposed work. Fig. 4.2 shows the steps of cropping and filtering, where final ROI is free from background information and noises.

4.2.1.2 Wavelet Transform

As we know that zooming of mammographic ROIs can make texture visualization more clear, and this can be possible through the concept of multi-resolution. So being inspired this concept, this work has used DWT, decomposes the ROIs into a number of sub-images in different resolution levels preserving the low and high-frequency detail. This decomposition property of wavelet leads to extract most dominating texture description from the ROIs of the mammograms. In this work, we have used 2-D Daubechies (Db4)

wavelet family which is applied on each ROI to get approximate and detailed coefficients [39-40, 90].

The generation of 2D-DWT is described as follows:

Wavelets are functions generated from mother wavelet by scaling and translations in frequency (time) domain. If $\psi(t)$ is a mother wavelet, then another wavelet $\psi_{a,b}(t)$ is represented as [40, 90]:

$$\psi_{a,b}{}^{k}(t) = \frac{1}{\sqrt{|a|}}\psi^{k}\left(\frac{t-b}{a}\right) \tag{4.1}$$

where *a* and *b* are two real number (variable) represents the parameters for *scaling and translations* in the time axis, and *k* ϵ *{h, v, d}* represents the orientation parameters in horizontal, vertical and diagonal directions respectively.

So, wavelet transform (*W*) of a function or signal f *(t)* is mathematically represented by

$$W(a,b) = \int_{-\infty}^{+\infty} \psi_{(a,b)}{}^{k}(t) f(t) dt \tag{4.2}$$

For the discrete wavelet transform, it is essential to discretize the dilation (*a*) and translation (*b*) parameters. There are many ways for the discretization of *a* and *b*. The most popular approach of discretizing *a* and *b* is dyadic sampling where the corresponding decomposition for *m, n* natural number is achieved by putting:

$$a = 2^{m} \text{ , and } \quad b = n2^{m} \tag{4.3}$$

Using these values, discrete wavelet family is created as:

$$\psi^{k}{}_{m,n}(t) = 2^{-\frac{m}{2}}\psi^{k}(2^{-m}t - n) \tag{4.4}$$

So, the wavelet coefficients for function *f (t)* is given by

$$W(m,n) = 2^{-\frac{m}{2}} \int f(t)\psi^{k}(2^{-m}t - n)dt \tag{4.5}$$

Also, we can construct discrete scaling families such using scaling function ϕ (x) as.

$$\phi^{k}{}_{m,n}(t) = 2^{-\frac{m}{2}}\phi^{k}(2^{-m}t - n) \tag{4.6}$$

These are the orthonormal basis of sub-spaces and related to resolution 2^m. We can define the wavelet atoms by translating and scaling three mother atoms like ψ^{h}, ψ^{v}, and ψ^{d}. These oriented mother atoms are computed as the tensor product of one dimensional $\psi(x)$ and $\phi(x)$ given by:

$$\begin{aligned}
\phi(x) &= \phi(x1)\phi(x2), \\
\psi^{h}(x) &= \psi(x1)\phi(x2), \\
\psi^{v}(x) &= \phi(x1)\psi(x2), \\
\psi^{d}(x) &= \psi(x1)\psi(x2)
\end{aligned} \tag{4.7}$$

For the transformation of two-dimensional mammogram image represented by a 2D array X [M, N] with M rows and *N* columns. The simple approach for 2D implementation of the DWT is shown in Fig. 4.2, where first we apply the one-dimensional DWT on mammogram along the row-wise to produce an intermediate result and then perform the same along column-wise on this intermediate result to produce the final first level result. This is possible due to the separable properties of two-dimensional scaling and wavelet function, i.e. product of two one-dimensional scaling and wavelet function, as written in Equation (4.7).

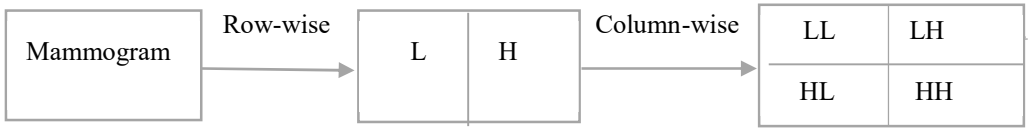

Fig.4.3: First level decomposition

When we apply the one-dimensional transform in each row, it produces two sub-bands in each row. First is the low-frequency sub-bands of all the rows (L), thin version of size *M* x N/2, and other is high-frequency sub-bands of all the rows to produce the sub-band (H) of size *M* xN/2, which contains high-frequency information around

discontinuities (edges in an image). Further, we apply one-dimensional DWT column-wise on these L and H sub-bands, produce four sub bands LL (low-low), LH(low-high), HL(high-low), and HH (high-high) of size M/2 x N/2 respectively. Where LL is a fine version of the original input image, contain the approximated details, and LH, HL, and HH are the high-frequency sub-band containing the detailed information. This decomposition of the image into approximation and detail coefficient via low-pass and high-pass filtering can be performed at any level by recursively applying the same decomposition on consecutive low-pass approximation (LL) coefficients. In this work, we have taken 2-level decomposition of db4 wavelet family, decomposition output for an input mammogram is shown in Fig. 4.4.

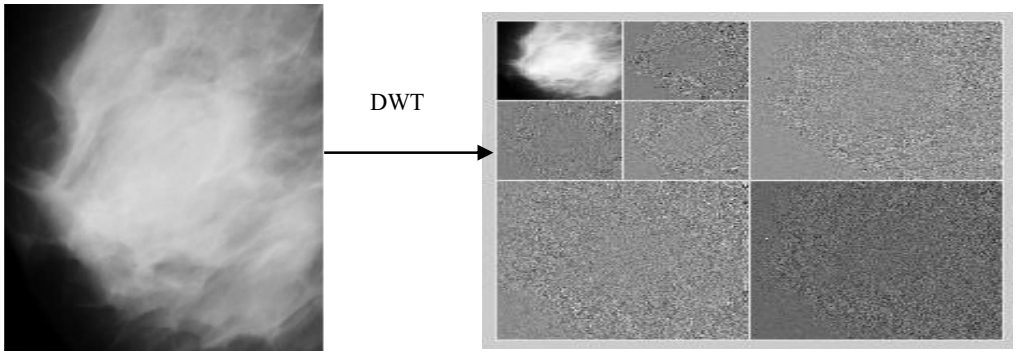

Fig. 4.4: Two- level wavelet decomposition of input mammogram

4.2.1.3 Feature Extraction

In mammogram classification overlapping problem biased the classification performance. For avoiding the problems of high overlapping between different mammogram classes and high intra-class variability within the same class, we partition the transformed ROIs into 2×2 non-overlapping local blocks. This local partition has solved the problem of overlapping [140]. In this work, 2-level low-pass filtered sub-image (LL^2) and the sum of high wavelet energy ($HL^2+LH^2+HH^2$) are divided into 2×2 non-overlapping sub-regions. Further, we extract CS-LBPs features from each block

which hold the spatial information in the local wavelet domain. Generally, medical images are characterizes using texture and shape based parameters. However, as low-dose X-ray images consist of a meaningless dark background region and interesting bright foreground region and have different colourless shapes. So, texture is the most suitable descriptor for the mammograms.

This chapter introduces a center symmetric-local binary patterns (CS-LBP), small dimension descriptor which captures the strong texture characteristics. CS-LBP is modified version of local binary pattern [140], proposed to solve the problem of long histogram in LBP. The procedure of CS-LBP feature extraction is shown in Fig. 4.5, where instead of comparing each neighbour pixel with the center pixel, the center-symmetric (CS) pairs of pixels will be compared. The mathematical model of CS-LBP is expressed below:

$$CS-LBP_{(P,R,T)} = \sum_{i=1}^{\frac{P}{2}} S\left(g_i - g_{i+\frac{P}{2}}\right) 2^i \tag{4.8}$$

$$S(x) = \begin{cases} 0 & x \le T \\ 1 & else \end{cases}$$

where P is neighbour connectivity, g_i and $g_{i+P/2}$ correspond to the gray level of center-symmetric pairs of pixels of equally spaced pixels on a circle of radius R and T threshold. From the equation, it is clear that CS-LBP is directly related to gradient operator because it considers gray-level differences between pairs of opposite pixels in a neighbourhood.

CS-LBP feature takes advantages of both the properties of the gradient-based feature and local binary pattern, hold the characteristics such as tolerance against illumination changes and robustness against monotonic grey-level changes while keeping the low computation cost. As compared to LBP, CS-LBP halves the number of comparisons for the same neighbouring and produces only 16(2^4) different binary

patterns**.** This relation leads to applying CS-LBP in wavelet domain for further mammogram classification.

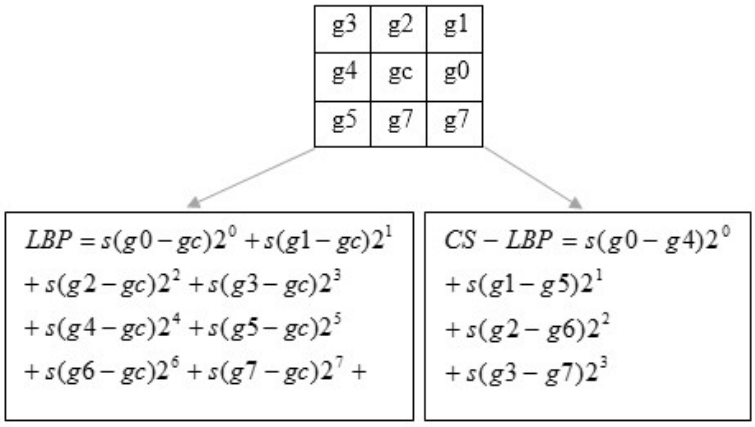

Fig. 4.5: 8-neighbouring connectivity in LBP and CS-LBP features

- **Proposed Local Wavelet-Based CS-LBP**

For the extraction of texture features in wavelet domain, inspired from Beura et al. [90] and Du et al. [142] where, Beura et al. [90] had used gray level co-occurrence matrix (GLCM) based statistical features from 2-level decomposed high pass detailed coefficients (LH, HL, HH). But these detail coefficients can only capture the abrupt illumination changes in the mammogram. However, low-pass approximated coefficient is also important, because it allows good energy compaction of the original mammogram with relatively lower noise. Further, Du et al. [142] had used the LBP in wavelet domain for handwriting image recognition. Using this method, handwriting images are decomposed at S-level then local LBP histograms from all of the multi-scale sub-images are extracted and concatenated. So, the feature dimension of histogram became larger, is depended on the number of decomposed level i.e. for 2 level of wavelet decomposition this generates one low pass filtered (LL) sub-image and six high-pass filtered sub-images**.** As we know that, LBP generates 256 different patterns for the P=8 neighbour. So, the final dimension of the LBP histogram is 1792 [256×7

(i.e. 6 Detailed level+ 1 Approximated level)]. Using CS-LBP we can reduce the feature dimension of histogram up to 112 [i.e. 16×7]. In this study, LL^2 and others relevant abbreviation are used for 2nd level decomposition.

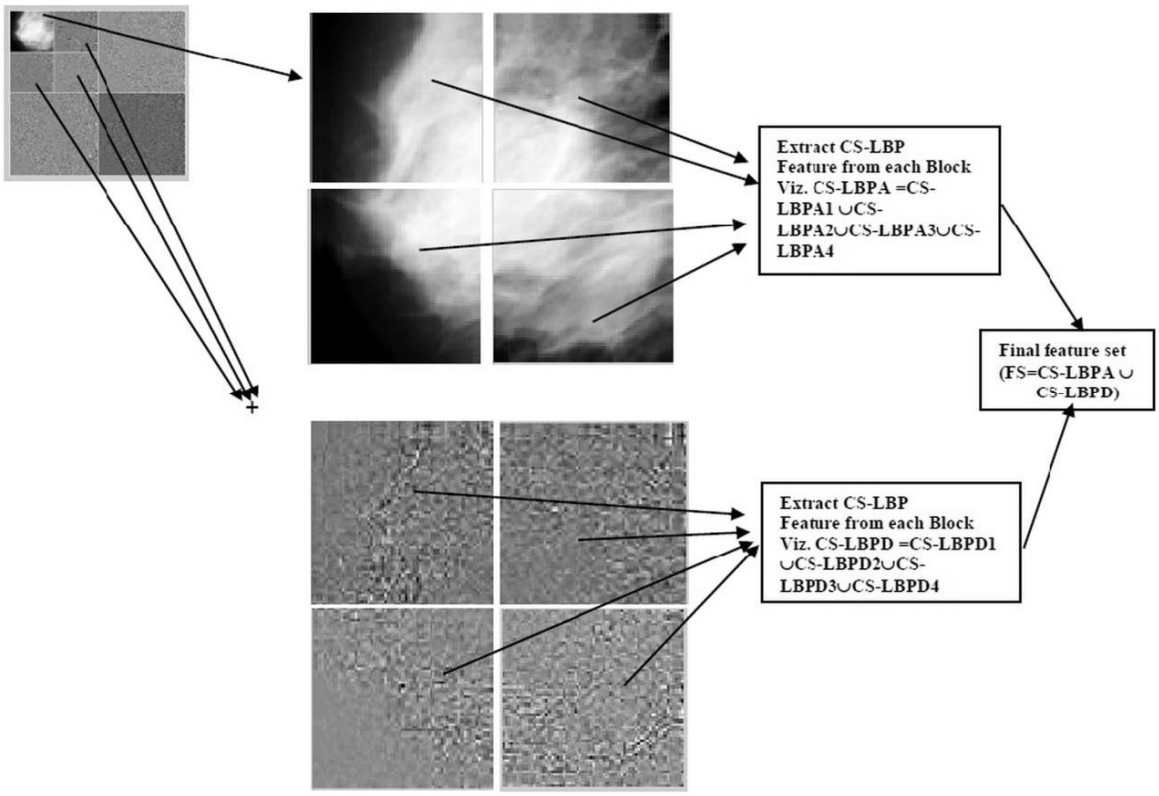

Fig. 4.6: Block preparation of LL^2 sub-bands and $|LH^2+HL^2+HH^2|$ sub-bands

Moreover, wavelet-based LBP needed eight multiplications, eight subtractions, and eight summations for each pixel at each of the seven decompose sub-images. Therefore, if the size of an image is 256×256, the number of required operations are 8388608 [(8×8× 8) × (LL^2: 64×64×1+ LH^2, HL^2, and HH^2: 64×64×3), which reduces drastically by using CS-LBP [(4×4× 4) × (LL^2: 64×64×1+ LH^2, HL^2, and HH^2: 64×64×3).

Taking advantage of previous contribution [90, 142] and improving upon their limitations, we have used low pass filtered and high-pass filtered sub-images and

extracted CS-LBP histogram rather than LBP. Complete feature extraction processes are also shown in Fig. 4.6. Fig. 4.7 shows the LBP and CS-CBP histograms for LL^2(2^{nd} level approximated information) sub-band of the divided mammogram. From the histogram, it is observed that the dimension of features is 16 and 256 for CS-LBP and LBP descriptor.

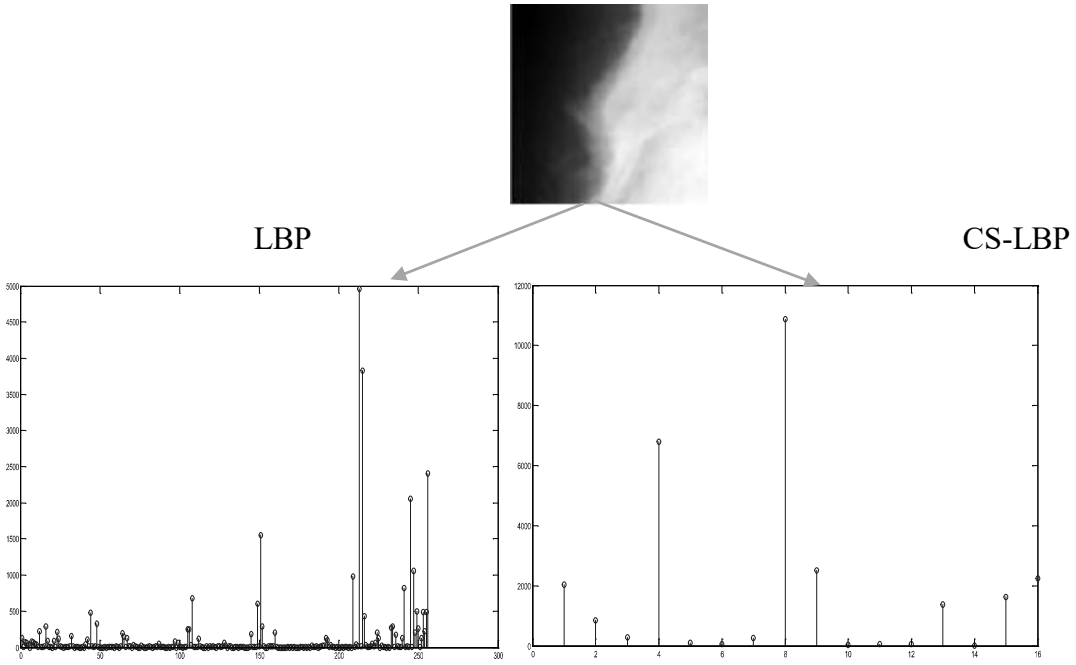

Fig. 4.7: LBP and CS-LBP histogram of a decomposed block

For wavelet decomposition, we use db4 (Daubechies) filter, because it yields better frequency resolution and achieves better spatial resolution than other wavelets [143], and for taking the advantage of good energy compaction, we have considered LL sub-band. The feature set for LL^2 (approximated coefficient):

As mammogram is divided into 4 non-overlapping blocks (i.e. A1, A2, A3, and A4) for each LL^2 level image, so final approximated feature vector is calculated as:

$$\mathrm{CS-LBP}_A = \mathrm{CS-LBP}_{A1} \cup \mathrm{CS-LBP}_{A2} \cup \mathrm{CS-LBP}_{A3} \cup \mathrm{CS-LBP}_{A4}$$

where $\mathrm{CS\text{-}LBP}_{A1}$=[F1 F2 F3……..F16],

$\mathrm{CS\text{-}LBP}_{A2}$=[F17 F18 F19…..F32],

$CS\text{-}LBP_{A3}$=[F33 F34 F35…..F48],

and $CS\text{-}LBP_{A4}$=[F49 F50 F51…..F64]

Since the X-ray image has strong edge distribution in the horizontal, vertical and diagonal directions (LH, HL and HH). So this work has taken the sum of all details information at 2 level decomposition, defined as:

$$D=| \quad HL^2+LH^2+HH^2 \quad | \tag{4.9}$$

Further, divides this details coefficient high energy sub-image into 4 non-overlapping blocks (i.e. D1, D2, D3, and D4). So final detailed feature vector is calculated as:

$$CS-LBP_D = CS-LBP_{D1} \cup CS-LBP_{D2} \cup CS-LBP_{D3} \cup CS-LBP_{D4}$$

where $CS\text{-}LBP_{D1}$=[F65 F66 F67…….F80],

$CS\text{-}LBP_{D2}$=[F81 F82 F83……..F96],

$CS\text{-}LBP_{D3}$= [F97 F98 F99……....F112]

and $CS\text{-}LBP_{D4}$=[F113 F114 F119……..F128]

Final feature set: FS= $CS\text{-}LBP_A \cup CS\text{-}LBP_D$

i.e. FS= [F1 F2 F3……F128]

Feature matrix of *K* number of images=*K*×128

We can write this feature vector into *K*×*n* representation, where *n* is the number of features, and a feature vector can be expressed in a matrix such as:

$FS=[FS_{1,i}\, FS_{2,i}.....FS_{J,i}……FS_{K,i}]$, where i=1 to *n* (i.e. *n*=128)

This feature set (FS) is described as feature vectors of the mammogram and they are stored for the representation of mammograms.

4.2.1.4 Feature Selection

Being motivated by the advantages of feature selection, viz. better classification accuracy; working with fewer features to speed-up the modeling time; cost savings from having fewer features, this work incorporated feature selection technique. In this regard, we have used support vector machine based recursive feature elimination (SVM-RFE) algorithm proposed by Guyon et al. [144], is a wrapper feature selection method, returns a ranking of the features by training a SVM with a linear kernel and ignoring the features which have smallest ranking criterion. It is a weighted-feature elimination method, in which at each step, the coefficients of the weight vector of a linear SVM are used for the ranking of features. The features are eliminated according to a criterion related to their support to the discrimination function, and the SVM is re-trained at each step. The working steps of this algorithm are as follow:

1. First, we train an SVM classifier on the training set, subset of feature set (i.e. subset of FS)
2. Rank the order of features using the weights of the resulting classifier.
3. Remove the features which have the smallest weight;
4. Repeat the process with the training set of the remaining features.

To evaluate the worth of attributes by using an SVM classifier, attributes are ranked by the square of the weight assigned by the SVM. For selecting the most informative features, we have evaluated full training data set (i.e. FS feature set) with SVM feature evaluator, and use attribute ranking, which ranked the features importance in increasing order. Final feature importance ranking for all 128 features is given in increasing order. Among all these high ranked features, we have selected a small sub-set of high weighted features and eliminate the remaining features. In this study, we have

taken small subset of 9 features which details are further discussed in result analysis and discussion section.

4.2.1.5 Mammogram Classification Using Random Forests

Finally, selected features are stored in the database, used by the random forest classifier for mammogram classification. The basic concept of random forest are based on the formation of weak decision trees in parallel, further, we combine the trees to form a single, strong learner by averaging or taking the majority vote. During the classification of images, feature set are randomly divided into the various subset for the training of decision tree individually. Where each node chooses a small subset of features at random, for finding a feature which optimized the split. After training of each decision tree, final output is derived by ensemble approach using majority voting. The working of this classifier is already demonstrated in Algorithm 3.1.

4.2.2 Mammogram Classification-cum-Retrieval

The proposed framework is based on supervised learning where for the training of the random forests only k % ground truth annotation of mammograms are used. Further, this trained classifier works as a manager and categorizes the remaining mammograms into different libraries (i.e. Normal or Abnormal). For the training, we have used the same feature set. However, at the query time same features are extracted and fed to the classifier. This classifier detects the class of query and searching is narrowed down to the corresponding library. This mammogram classification and retrieval model filters out maximum irrelevant images and speeds up the searching time. This process is treated as a pre-filtering step and applied prior computing similarity measure to speed up the searching time. Fig.4.8 shows the working of proposed method where Majority

voting based ensemble method combined the output of all the decision tress and predict the appropriate class based on voting.

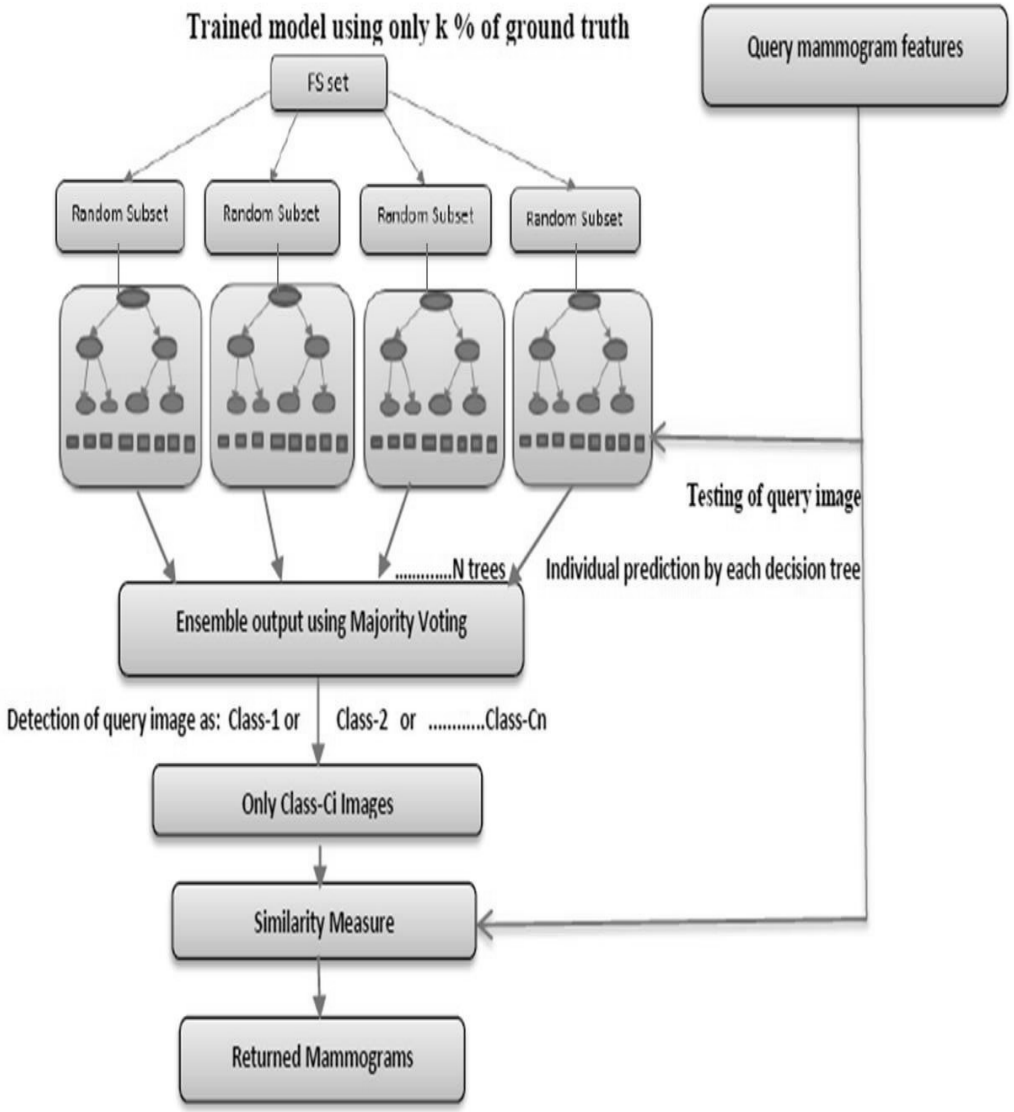

Fig. 4.8: Working diagram of mammogram retrieval

Further, mammograms are retrieved from the cluster or library of the predicted class based on similarity criteria which assigns a rank to each image. The top images obtained from this procedure are then returned. In our case, Euclidean distance between the extracted features of the query image and that of the images of the predicted cluster was used as similarity criterion.

4.3 Result Analysis and Discussion

Result analysis of this work is shown into two sections. Section 4.3.1 discusses about the analysis of mammogram classification, and section 4.3.2 gives the results analysis for mammogram classification cum retrieval framework.

4.3.1 Result Analysis for Mammogram Classification

MIAS database has been used for this study. There are 207 healthy mammograms and 115 abnormal mammograms. Five mammograms of this database consist of more than one abnormal region. Therefore, total of 120 abnormal regions are used for the analysis of this contribution.

In order to evaluate the performance of proposed work, we have used K- fold cross-validation where first we divided the samples into K-folds (10 folds), where K-1 folds are used for training and remaining one fold is used for testing. The performance of LBP and CS-LBP descriptors in local wavelet domain are tested using Random forest classifier, having 100 trees each constructed using 12 random features from LBP and 8 random features from CS-LBP. As the feature dimension of wavelet local binary patterns (W-LBP) is huge, 16 times larger than WCS-LBP, requires extra modelling time to build the model (see Fig. 4.9). During the experimental study, it was found that building model using WCS-LBP features is approximately four times faster than W-LBP. Further, according to the reported results in Fig. 4.10 as the overall mean of average Precision, Recall, F-Measure, MCC and Accuracy of both normal and abnormal classes, the performance measures of WCS-LBP approach is found better in comparison to the W- LBP.

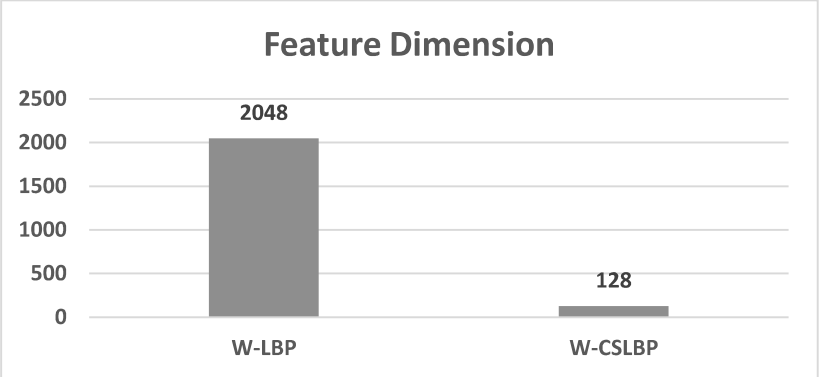

Fig. 4.9: Comparison of feature dimensions

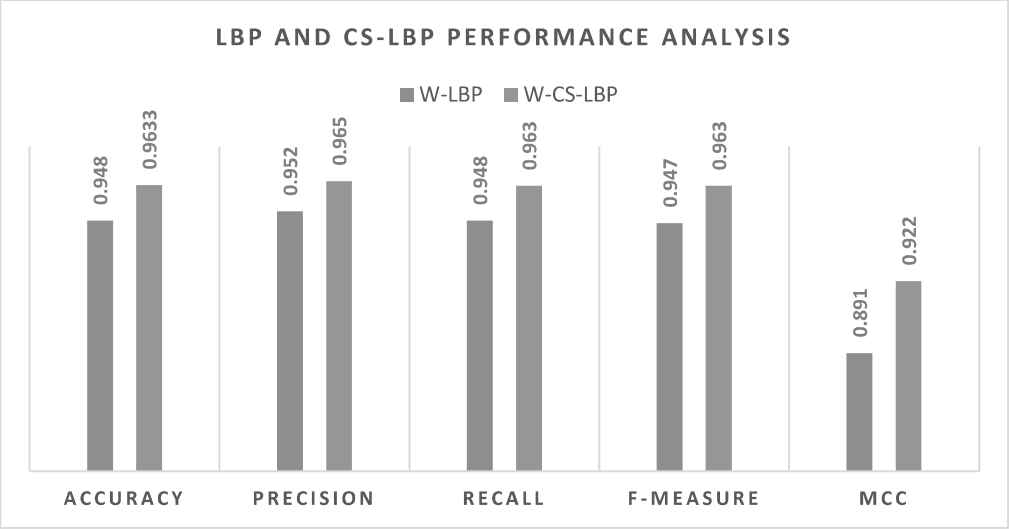

Fig. 4.10: Comparative performance analysis of W-LBP and WCS-LBP features

The overall average Precision, Recall, F-Measure, MCC, and Accuracy of the proposed WCS-LBP feature is better than W-LBP by 7.98%, 5.95%, 3.35%, and 5 % respectively.

Since the classification performance of CS-LBP is more effective than LBP, so in feature selection section we have taken only CS-LBP based features and ranked the importance of all the features in increasing order using SVM-RFE. Further, we have evaluated the classification accuracy of random forest classifier on different number of high ranked features. During this study and from Fig. 4.11, we noticed motivating classification performance for 9 and 19 number of high ranked features. So in this

study, we have taken only 9 high ranked features, well suited for fast modelling and classification as compared to 19 features. Fig. 4.12 shows the statistical distribution of 9 selected features and target class. For target class, the frequency of mammogram classes is shown (207 for normal and 120 for abnormal mammograms).

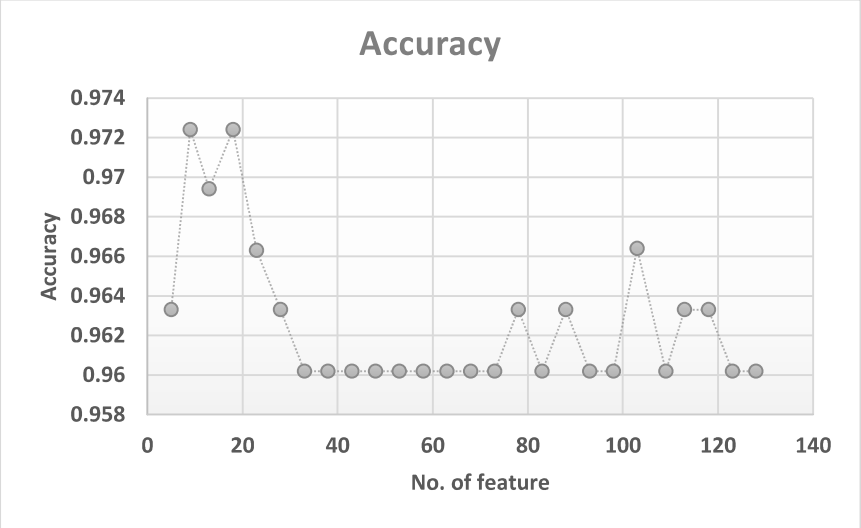

Fig.4.11: Classification accuracy and used number of features

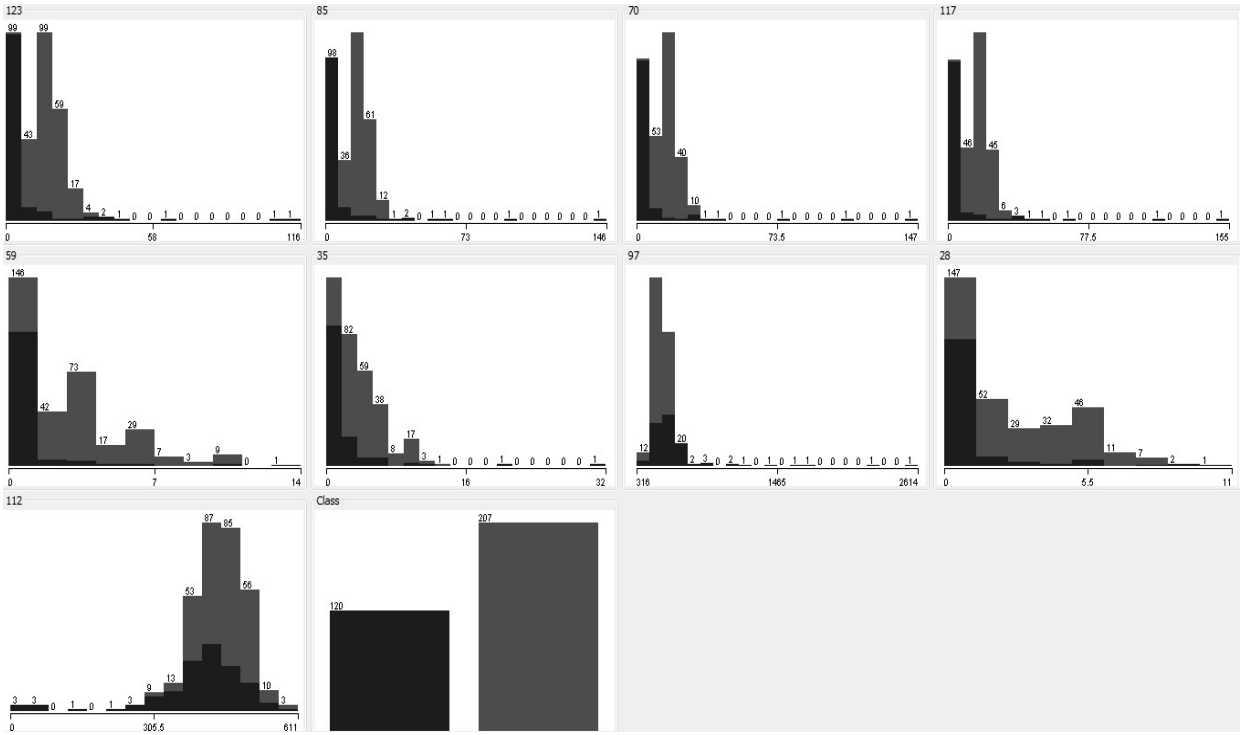

Fig. 4.12: Statistical distribution of selected features through SVM-RFE

For selected features (id's 123, 85, 70,117,59,35,97, 28,112), we have calculated the mean, standard deviation, minimum, and maximum values which are used to represent the quantitative distribution of each sample mammogram. In Fig. 4.12, each graph shows the statistical distribution of each individual feature having the distribution between minimum and maximum for all the samples. In this graph, we have included the distribution of all selected features. Mean and Standard deviation for nine selected features have been calculated using sample distribution of their values. From the figure, it can be observed that the distribution of mammogram instances are only left or right skewed, which shows significant discrimination ability.

Further, for the modelling of random forest, the process starts with number choice of decision trees. Fig. 4.13 shows the performance of random forest for different number of trees, where it reflects maximum 97.25 % accuracy for 30, 50, 60 and 100 number of trees. Random forest of 30 trees, each constructed while considering 4 random features takes 0.03 seconds for building model for classification, while 50, 60 and 100 number of the tree take 0.04, 0.04 and 0.06 seconds, respectively. So, due to fast modelling, we has taken only 30 number of trees for further experimental analysis.

From the confusion matrix Table 4.1, it is clear that the performances of this work are outstanding for both normal and abnormal classes. Here, only one image of the normal class is misclassified as abnormal and eight abnormal images are misclassified as normal. The overall accuracy of this classifier is 97.25 %, means in 327 mammograms; it correctly recognizes 318 images. From Table 4.2, it is clear that the performance of this work is significantly encouraging, we got average precision 97.3 %, recall 97.2 %, MCC= 94.1 %, F-measure= 97.2% and least 4.4% false positive rate.

Table 4.1: Confusion matrix for mammogram classification

Classes	Actual Abnormal	Actual Normal
Predicted Abnormal	112	8
Predicted Normal	1	206

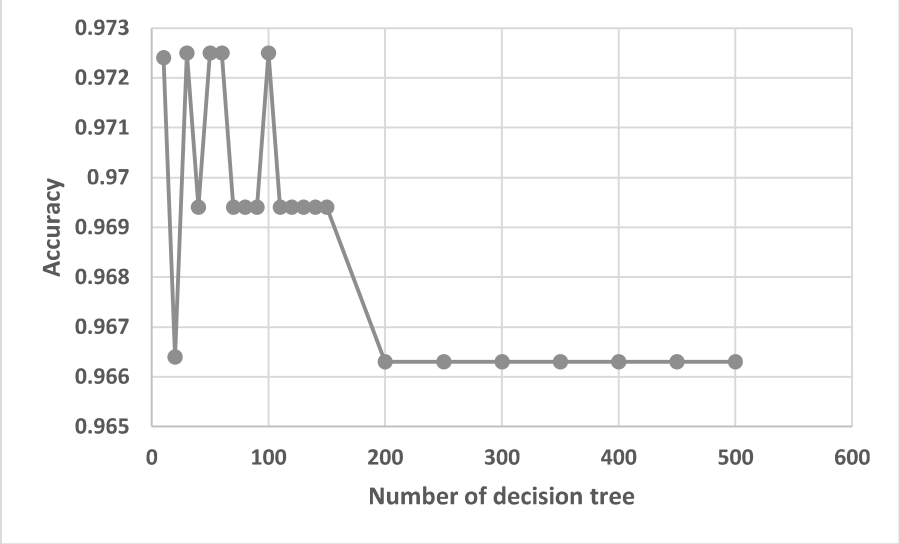

Fig. 4.13: Performance of random forests on different number of trees

Table 4.2 Performance analysis of proposed mammogram classification model

Classes	Precision	Recall	F-measure	MCC	FP Rate
Abnormal	0.991	0.933	0.961	0.941	0.005
Normal	0.963	0.995	0.979	0.941	0.067
Average	**0.973**	**0.972**	**0.972**	**0.941**	**0.044**

Table 4.3: Comparative analysis of the proposed method with other state-of-art methods

State-of-art Work	Accuracy & other performance measures	Used database
Proposed Work	**Accuracy:97.25%,** **Precison:97.3%, Recall:97.2%,** **F-meaure:97.2%** **MCC:94.1%** **ROC Area:97.61%**	MIAS

Srivastava *et al.* [89]	Accuracy: 87%, Sensitivity: 95% Specificity:75%, Accuracy: 87.50%, Sensitivity: 95.83%, Specificity: 62.50%	MIAS
Beura *et al.* [90]	For MIAS: Specificity=97.0% Accuracy 94.2% For DDSM: Specificity=97.9%, Accuracy=97.4%	MIAS, DDSM
Pratiwi M *et al*. [92]	Highest achieved Accuracy, Sensitivity Specificity, 93.98% 94.44% 93.62% respectively	MIAS
Tzikopoulos *et al.* [93]	Accuracy=84.47-85.3%	MIAS
Buciu et al. [94]	Specificity=60.86%, AUC = 0.79, Sensitivity=97.56%,	MIAS
Prathibha et al. [95]	AUC = 0.95	MIAS
Liu et al.[96]	Accuracy=84.2%	MIAS
Subashini et al.[97]	Accuracy= 86.67%	MIAS
Wang et al.[98]	Accuracy=71%	MIAS
Muhimmah et al. [99]	Accuracy=77.57%	MIAS
Oliver et al.[100]	Specificity=91%	MIAS
Miller & Astley [101]	Accuracy=80%	MIAS
Karahaliou et al.[102]	Best achieved classification accuracy 89%, Sensitivity 90.74% and specificity 86.96%).	DDSM
Mutaz et al. [103]	74% Sensitivity=91.6%, Specificity=84.17%	MIAS
Petrosian et al.[104]	76 % specificity, 89%-sensitivity	Local database of 180 mammograms
Kinoshita et al.[105]	Accuracy=81%	Local database
Sameti et al.[106]	Accuracy=72%	Locally collected 58 mammograms
Dhahbi et al.[107]	Accuracy =91.27% for MIAS database	MIAS and DDSM

Mudigonda et al.[108]	Accuracy=83%	MIAS+local images (39-MIAS+15 local) total 54 images
Brijesh et al.[109]	Accuracy=77.8–83.3%	Nijmegen database
Liyang Wei et al.[110]	Accuracy=85%	697 mammograms from 386 cases
Szekeley et al.[111]	Accuracy=88%	Local database
Alolfe et al.[112]	Accuracy=82.5%	Local/MIAS
Jona et al.[113]	Accuracy=94.0%	MIAS
G¨orgel et al. [114]	Accuracy rate=84.8%	MIAS
G¨orgel et al.[115]	Accuracy rate=96.0%	MIAS
Rashed et al. [116]	Accuracy=87.06%	MIAS
Dheeba et al [117]	Sensitivity=94.167%, Precision=92.12% Accuracy=93.67% AUC=0.9685	216 sample mammograms from 54 patients

Table 4.3, presents the comparative analysis of the proposed approach with other CAD tools and classification frameworks. From this table, it is confirmed that the proposed approach is performing significantly encouraging in comparison to most of the under consideration methods.

4.3.2 Result Analysis for Mammogram Classification-cum-Retrieval

In this analysis, we have used k% amount of known ground truth of mammograms from each category for the training of the classifier. The trained classifier then predicted the class of the remaining (100-k) % mammograms with some classification accuracy. Let this classification accuracy be clustering classification accuracy (CCA). For querying, we randomly took 10 sample mammograms of each category from the remaining database as input. In this framework, trained classifier performs the different class-wise CCA, in which searching is performed. For given query image the discussed low-level features (FS) are extracted and given as input to the already trained model. The trained model predicts the category of the query which is nothing but the semantic

concept of the query mammogram. Hence instead of finding similarity between the query image and all the images in the database, it is found between the query image and only the images belonging to the query image category. Therefore, the retrieval results are more accurate and fast. From the Table 4.5, we can see that CCA=94.86% is quite encouraging even for 5 % known mammograms. In this analysis, 5% of MIAS database is used for training. It means ground truth details of only16 mammograms are used for the training of random forests, and remaining 95% mammograms (311) are classified in different clusters based on these training, where this trained classifier managed the 295 mammograms in correct clusters or folders.

Table 4.5: CCA and retrieval performance of the proposed method

Training (k %)	CCA	Average Precision
5	94.86% (295/311)	88.50%
10	93.54% (275/294)	87.50%
15	93.17% (259/278)	85.50%
20	94.27% (247/262)	85.30%
25	93.88 % (230/245)	85%
30	93.01 (213/229)	79.50%
35	93.90 (200/213)	84.50%
40	95.92 (188/196)	91.50%

Further, this table shows the interesting CCA for the increasing number of ground truth, where CCAs are nearly preserved. For k=40%, we have achieved the

95.92 % classification accuracy for remaining 60 % mammograms. To compare our method with the conventional CBIR, we implemented the conventional model on the same dataset taking the same sample query images used to test our method. Identical feature vectors were extracted from all the images in the dataset. These vectors were matched with the feature vector of query image based on Euclidean distance and topmost 10 similar images were retrieved. From the Fig. 4.14, we can see that the proposed approach with only 5% of known ground truth with same set of features gives significantly encouraging retrieval performance (average precision) as compared to conventional CBIR.

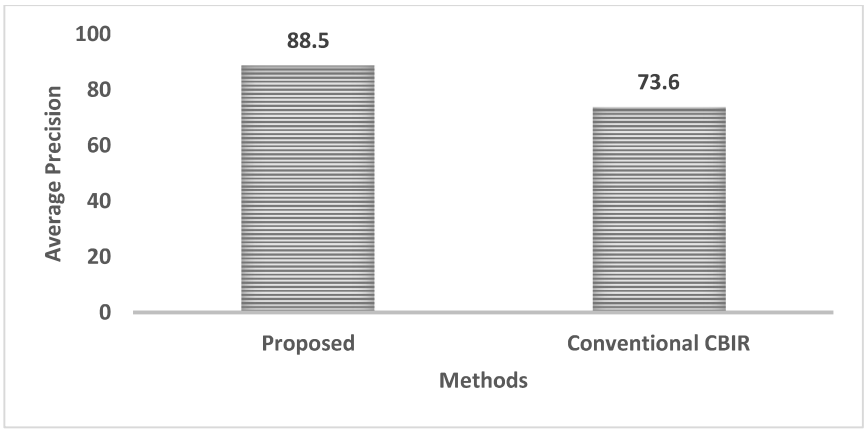

Fig. 4.14: Retrieval accuracy as compare to conventional approach

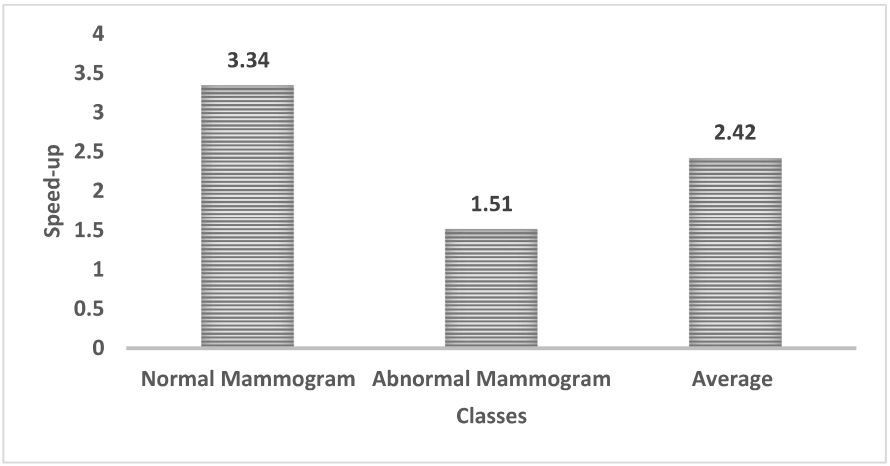

Fig. 4.15: Retrieval speed-up of the proposed classification-cum-retrieval framework

Searching time improvement, as compare to conventional CBIR based exhaustive search methods is estimated using Speedup metric, given in Fig. 4.15. Here, we got the maximum *Speedup* up to 3.34. Means, searching time is 3.34 times faster than conventional CBIR (if the query image belongs to the healthy mammograms). The overall searching time improvement as compared to convention CBIR is 2.42. Therefore, this work retrieval performance is very promising, which gives accurate and fast response for a particular query.

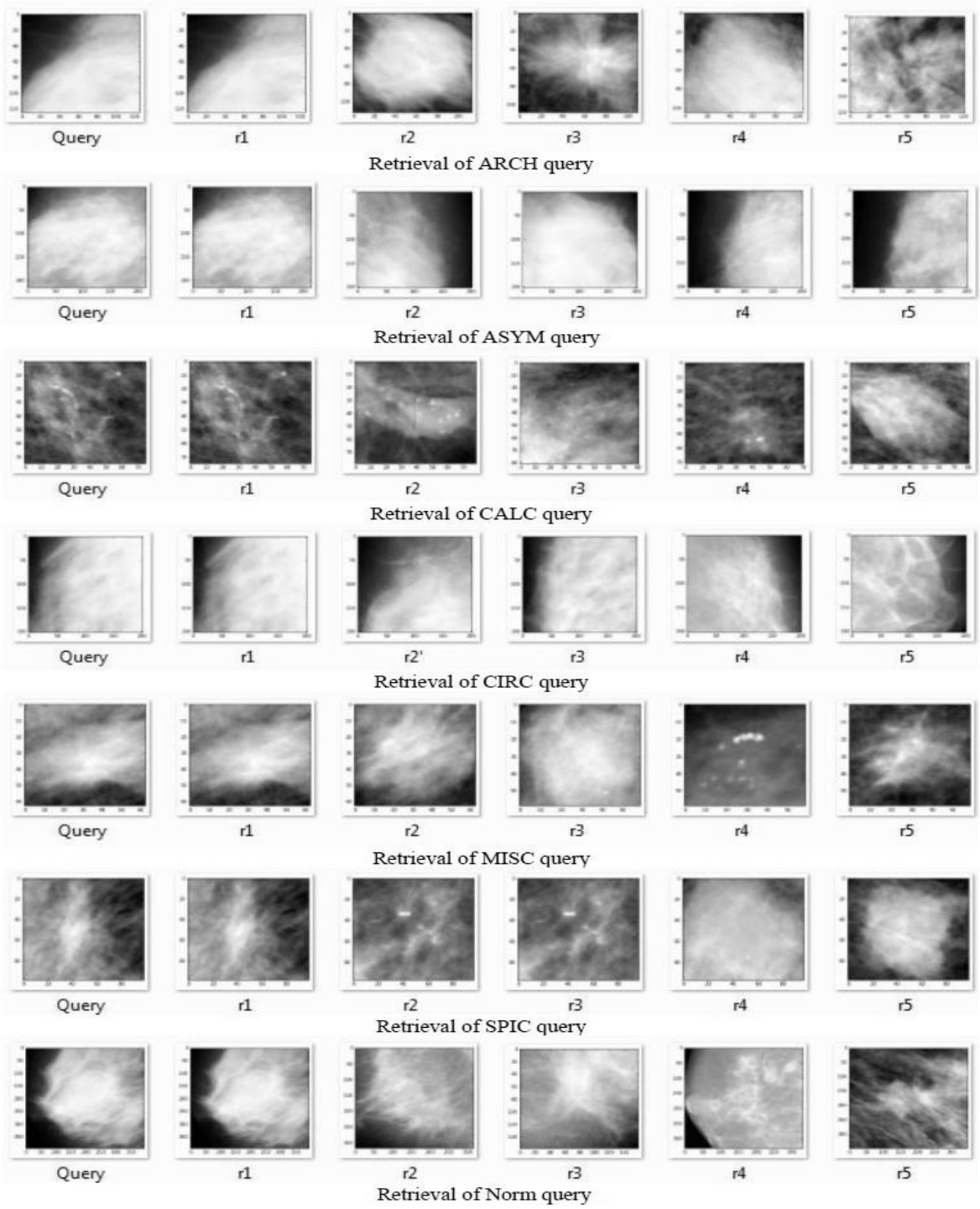

Fig. 4.16: Sample of mammogram retrieval

Fig. 4.16 shows the glimpse of effectiveness for the proposed work, where we have shown the retrieval snapshot for seven different queries (r1 to r5 are ranking in increasing order). In seven queries, six queries belong to the abnormal classes and one

query from healthy mammogram. Abbreviation of these six abnormal queries are already given in section 2.7. From this figure, we can see that maximum of retrieved mammograms from each class are relevant to the query.

4.4 Conclusions

In this chapter, design, analysis and comparative study of a general framework for the mammogram classification and retrieval were presented. For pre-processing of mammograms, labels, and background having no details were cropped and ROI's were extracted, after that adaptive median filter was used to suppress the noises. Further, we introduced random forests with selected local wavelet based CS-LBP features which improved the classification performances. To describe the most dominating wavelet-based texture characteristics, we first extracted CS-LBP features from the non-overlapping local region in wavelet domain and applied SVM-RFE which selects a small subset of most relevant features. Then we applied these selected features to decision trees, construct random forests which are an ensemble of random decision trees. This proposed work was found advantageous in terms of fast feature extraction, small feature dimension, fast modelling and prediction, and significantly encouraging performance as compared to other state-of-art methods.

Further, using same extracted features from small set of know ground truth, we trained the random forest classifier, and this trained model managed the remaining mammogram in different libraries. At the retrieval time, this model predicts the category of the query which is nothing but the semantic class of the query mammogram. Hence instead of finding similarity between the query image and all the images in the database, it is found between the query image and only the images belonging to the query image category. So, the retrieval results were more relevant and fast.