CHAPTER 3
STATE-OF-THE-ART METHODS

3.1 INTRODUCTION

In section 3.2 of this chapter the description as well as the implementation details of the following state-of-the-art texture feature extraction methods is given. GLCM [Wei et al., 2006], [Haralick et al., 1973]; GLAM [Qin and Yang, 2004]; Gabor [Manjunath et al., 1996]; Db4 wavelet [Do et al., 2002]; CDF 9/7 biorthogonal wavelet [Lamard et al., 2007]. In addition to texture features the shape based feature, i.e., the Zernike moments [Li et al., 2009] is also included. In subsection 3.3, the state-of-the-art techniques that are included in feature selection analysis are: the t-test in the case of the GLCM and Gabor texture features (Wei et al., 2005); StARMiner with FD-ASE for ZM features (Felipe et al., 2005), (Li et al., 2009); the GA for CDF 9/7 wavelet based features (Quellec et al., 2010), (DeWouver et al., 1999), (Kay, 1993). The details of the similarity measures have been indentified to find its suitability with different type of features to improve the retrieval performance is given in subsection 3.4. This chapter ends with the concluding remark on the competing methods considered for this research work.

3.2 GRAY LEVEL CO-OCCURRENCE MATRIX (GLCM)

GLCM is one of the most well-known and widely used second order statistical methods to extract texture features [Srinivasan and Shobha, 2008]. The GLCM of an $N \times N$ pixel image $I$ contains the probabilities $P_{d,\theta}(i,j)$ of the changeover from a gray level $i$ to a gray level $j$ in a given direction $\theta$ separated by a pixel distance $d$ that is expressed as:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} C_{d,\theta}(i,j)$$

$$P_{d,\theta}(i,j) = \frac{C_{d,\theta}(i,j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} C_{d,\theta}(i,j)}$$

(3.1)

where $C_{d,\theta}(i,j) = \# \{(m,n),(u,v)\in N \times N : f(m,n)=i, f(u,v)=j, |(m,n)-(u,v)|=d, ((m,n),(u,v)) = \theta\}$, # denotes the number of elements in the set, $f(m,n)$ and $f(u,v)$ corresponds to the gray
levels of the pixel located at \((m, n)\) and \((u, v)\) respectively, and \(N_g\) is the total number of gray levels in the image. \(0^\circ, 45^\circ, 90^\circ\) and \(135^\circ\) are the common choices of \(\theta\) [Lakovidis et al., 2007]. If a texture is coarse then the probability distribution in the GLCM is concentrated on or near its diagonal. On the other hand, for a fine texture the probability distribution in the matrix is away from its diagonal [Kim and Park, 1999]. The fourteen Haralick texture features [Haralick et al., 1973] that are extracted from GLCMs are as follows:

1. **Angular Second Moment (ASM)**

   \[
   ASM = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \left\{ P(i, j) \right\}^2.
   \]

   ASM is also called as energy, measures the textural uniformity (or homogeneity), i.e., pixel pair repetitions. It also detects disorders in textures. Its maximum value is equal to one. The GLCM of less homogeneous image will have large number of small entries that may result in lower energy. A homogeneous scene will contain only a few gray levels, giving a GLCM with only a few but relatively high values of \(P(i, j)\). Thus, the sum of squares will be high and resulting in a higher energy value.

2. **Contrast**

   \[
   Contrast = \sum_{n=1}^{N_g} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P(i, j) \right\}.
   \]

   This statistic measures the spatial frequency of an image and is difference moment of GLCM. It is the difference between the highest and the lowest values of a neighbouring set of pixels. It measures the amount of local variations present in the image. A low contrast image presents GLCM concentration term around the principal diagonal and features low spatial frequencies.
(3) **Correlation**

\[
\text{Correlation} = \frac{\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} (ij) p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y}
\]  

(3.4)

The correlation descriptor measures the linear dependency of the gray level values in the co-occurrence matrix or the correlation presenting along a scan line of an image [Tahir et al., 2004].

(4) **Variance**

\[
\text{Variance} = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} (i - \mu)^2 P(i, j)
\]

(3.5)

This statistic is a measure of heterogeneity and is strongly correlated to first order statistical variable such as standard deviation. Variance increases when the gray level values differ from their mean. Thus, this feature puts relatively high weights on the elements that differ from the average value of \(P(i, j)\).

(5) **Inverse difference moment (IDM)**

\[
\text{IDM} = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \frac{1}{1+(i-j)^2} P(i, j)
\]

(3.6)

This statistic is also called as homogeneity. It measures image homogeneity as it assumes larger values for smaller gray tone differences in pair elements. It is more sensitive to the presence of near diagonal elements in the GLCM. It has maximum value when all elements in the image are same. GLCM contrast and homogeneity are strongly, but inversely, correlated in terms of equivalent distribution in the pixel pairs population. It means homogeneity decreases if contrast increases while energy is kept constant.
(6) **Sum average (SA)**

\[
SA = \sum_{i=2}^{2N_g} iP_{x+y}(i)
\]

where \( P_{x+y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P(i, j), k=2, 3, ..., 2N_g \) \hspace{1cm} (3.7)

(7) **Sum variance (SV)**

\[
SV = \sum_{i=2}^{2N_g} (i-SA)^2 P_{x+y}(i)
\]

Sum average and sum variance are the average and variance of normalized gray tone image in the spatial domain, respectively.

(8) **Sum entropy (SE)**

\[
SE = -\sum_{i=2}^{2N_g} P_{x+y}(i) \log \{ P_{x+y}(i) \} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \{P(i, j)\}^2
\]

\hspace{1cm} (3.9)

The sum entropy is a measure of randomness within an image.

(9) **Entropy (Ent)**

\[
Entropy = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P(i, j) \log (P(i, j))
\]

\hspace{1cm} (3.10)

Entropy is an indication of the complexity within an image. A complex image produces a high entropy value.

(10) **Difference variance (DV)**

\[
DV = \text{Variance of } P_{x-y}
\]
where \( P_{x-y}(k) = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(i, j), k = 1, \ldots, N_g \) \hspace{1cm} (3.11)

The difference variance is an image variation in a normalized co-occurrence matrix.

(11) Difference Entropy (DE)

\[
DE = - \sum_{i=1}^{N_g} P_{x-y}(i) \log \{ P_{x-y}(i) \}
\]

(3.12)

The difference entropy is also an indication of the amount of randomness in an image.

(12) & (13) Information Measure of Correlation I (IMC1), Information Measure of Correlation II (IMC2)

\[
IMC_I = \frac{H_{XY} - H_{XY1}}{\max\{H_X, H_Y\}}
\]

(3.13)

\[
IMC_{II} = (1 - \exp[-2.0(H_{XY2} - H_{XY})])^{1/2}
\]

(3.14)

\[
H_{XY} = - \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(i, j) \log(P(i, j))
\]

(3.15)

\[
H_{XY1} = - \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(i, j) \log\{P_x(i)P_y(j)\}
\]

(3.16)

\[
H_{XY2} = - \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P_x(i)P_y(j) \log\{P_x(i)P_y(j)\}
\]

(3.17)

(14) Maximal Correlation Coefficient (MCC)

\[
MCC = (\text{Second largest eigen value of } Q)^{1/2}
\]
\[ Q(i, j) = \sum_{k=1}^{N_g} \frac{P(i, k)P(j, k)}{P_x(i)P_y(k)} \] (3.18)

A measure of dependence of two random gray levels, defined as the least upper bound of the values of the correlation coefficients between the real random variables. If there is a linear correlation between the variables, then the maximal correlation coefficient coincides with the usual correlation coefficient.

The computational complexity of GLCMs increases for the analysis of large image data in demanding applications like content based image retrieval [Lakovidis et al., 2007]. The number of operations required to obtain GLCM is directly proportional to the number of gray levels in the image. Hence, a GLCM can have larger dimensionality. Therefore, it is often necessary that the number of gray levels in the image is reduced through quantization [Srinivasan and Shobha, 2008]. Another disadvantage of GLCMs is that they cannot capture spatial relationship between three or more pixels in the image [Qin and Yang, 2005]. In this work, the fourteen Haralick texture features are estimated from four GLCMs with orientations 0°, 45°, 90° and 135°, respectively resulting in a feature vector of size 56. Moreover, these feature vectors are generated from GLCMs with pixel distances 1, 3 and 5, which is referred in this work as GLCM-Dist.1, GLCM-Dist.3 and GLCM-Dist.5 features.

### 3.3 GRAY LEVEL AURA MATRIX (GLAM)

The concept of aura matrix was first introduced by Elfadel and Picard (1990) in their work on Markov Random Field texture modeling. It has been shown that the aura matrix is a generalization of GLCM. The understanding of aura and aura measure is necessary in the computation of aura matrix. In this subsection, the description of aura, aura measure, and aura matrix is given first. Then the procedure to compute the GLAM of a given image is described. The “aura set” and “aura measure” are based on the set theoretic concept and useful for texture analysis.
Aura

An image $X$ is modelled as a finite lattice $S$ with a neighbourhood structure $N = \{N_s, s \in S\}$, where every $N_s$ is a subset of the lattice $S$. Consider two subsets $A, B \in S$, then the aura of $A$ with respect to $B$ for the neighbourhood structure $\{N_s, s \in S\}$, denoted by $O_B(A, N)$ is given by

$$
\bigcup_{s \in A} (N_s \cap B)
$$

(3.19)

**Fig. 3.1** Generation of GLNM (a) Sample binary lattice (b) Symmetric neighbourhood structure. (c) The set of shaded elements is the aura set of $A$ with respect to $B$.

A binary $5 \times 5$ lattice is shown in Fig. 3.1(a). Let $A = \{ a | a \in S, a = 1 \}$ and $B = \{ b | b \in S, b = 0 \}$. Let $N = \{n_i, i = 0, 1, 2, 3\}$ as shown in Fig. 3.1(b). Then $O_B(A, N)$ are the elements shaded in Fig. 3.1(c). The aura of a set with respect to itself is called as self-aura. The aura depends on the neighbourhood structure chosen for the lattice. Any choice of the $N_s$'s are valid, but for simplification same neighbourhood structure can be considered and thus the simplified notation for the aura is $O_B(A)$. Moreover, the neighbourhood is site dependent that is the size of the neighbourhood will be different for the boundary and the inner site neighbourhoods [Elfadel and Picard, 1994].

Aura measure

The aura measure is given as,

$$
m(A, B) = \sum |N_s \cap B|
$$

(3.20)
Generally, \( m(A, B) \neq m(B, A) \). The easiest way for measuring the size of a set \( A \) with respect to a set \( B \) is by counting the number of elements \( |O_B(A)| \).

\[
|O_B(A)| \leq \sum |N, \cap B|
\]  
(3.21)

Aura matrix

Aura matrix is a useful tool for pattern processing and represents the aura measure for all the sets of interest on the lattice. The aura matrix is defined as the matrix \( A \) having element \( A = [a_{ij}] \) where \( a_{ij} = m(i, j) \) is the aura measure of the gray levels \( i \) and \( j \).

A neighbourhood structure of four nearest neighbours [Elfadel and Picard, 1994] is used in this work. For a symmetric neighbourhood system aura matrix is rotation invariant. In this work, the elements of normalized GLAM of pixel distance one are used to form the mammogram texture feature.

### 3.4 GABOR FILTER

A set of 24 Gabor filters are obtained by appropriate dilations and rotations of the Gabor function \( g(x,y) \) through the generating function \( g_{mn}(x,y) \) [Manjunath and Ma, 1996].

\[
g(x,y) = \frac{1}{2\pi\sigma_x \sigma_y} \exp \left[ -\frac{1}{2} \left( \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right) + 2\pi j Wx \right]
\]  
(3.22)

\[
G(u,v) = \exp \left\{ -\frac{1}{2} \left( \frac{(u-W)^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2} \right) \right\}; \sigma_u = 1/2\pi\sigma_x, \sigma_v = 1/2\pi\sigma_y
\]  
(3.23)

\[
g_{mn}(x,y) = a^{-m}G(x',y'), a > 1, m,n = \text{integer}
\]

\[
x' = a^{-m}(x\cos \theta + y\sin \theta), \quad y' = a^{-m}(-x\sin \theta + y\cos \theta); \quad \theta = n\pi/K
\]  
(3.24)

\( G(u,v) \) is the Fourier transform of \( g(x,y) \) and \( K \) is the total number of orientations. The filter parameters are computed as follows:
In this work, $K = 6$, $S = 4$, $U_h = 0.4$ and $U_l = 0.05$ are used. Application of these Gabor filters over the mammogram region of interest (ROI) resulted in 24 Gabor filtered outputs. As mentioned in [Manjunath et al, 1996], the mean $\mu_{mn}$ and standard deviation $\sigma_{mn}$ of Gabor filtered outputs are used as feature components to construct a feature vector of size 48, which we refer as Gabor-I. Another type of Gabor feature vector, which is referred as Gabor-II, consists of 144 feature components. To generate this type of feature vector, as mentioned in [Wei et al., 2007] each Gabor filtered output is transformed into a probabilistic matrix from which the contrast, angular second moment, inverse difference moment, entropy, variance and correlation features are computed.

### 3.5 DAUBECHIES (Db4) WAVELET

Wavelet transform is a well-known tool for image analysis. It provides a time-frequency representation of the data as well [Al-Ataby et al., 2012]. Basically, wavelet transform represents a signal $f$ by its linear approximation $\hat{f}$ onto a fixed subspace of dimension $N$ in the orthogonal basis $\{\phi_n\}$. A class of Daubechies wavelets forms an orthonormal basis and can be useful to perform the analysis of images. For an integer $r$, Daubechies wavelet can be defined as [Gonzalez-Garcia et al., 2007],

$$
\phi_{r,j,k}(x) = 2^{j/2} \phi_j(2^j x - k) \quad j, k \in \mathbb{Z}
$$

where, $j$ is a scale, $k$ is a translation and $r$ is a filter. The Daubechies wavelet transform with $r > 2$ (in our case, $r = 4$) presents an energy concentration that preserves the trend of the information when it is considered only as a low pass filter. Daubechies wavelets can be successfully used for discrete wavelet transform based image analysis applying the derived from the mother wavelet high pass and low pass filters in the dyadic sub-band image decomposition.
Due to the normalization of the functional space in the design of the base wavelet, the coefficients in the frequency bands tend to be more dominant and of greater magnitude than the coefficients of the highest frequency bands. In the contrary, the coefficients of the lowest frequency band are grouped in the upper left corner, meanwhile the coefficients of the higher frequency bands are in the other three image corners [Gonzalez-Garcia et al., 2007].

In order to obtain the information contained in the images, one needs to perform sub-level signal decompositions to separate the signal characteristics and to analyze them independently. From this idea so called multilevel filtering approach emerges. By iterating this filtering process until a desired precision level, one gets the well-known multilevel decomposition scheme also known as decomposition tree or wavelet decomposition, depicted in Fig. 3.2. By decomposing the image into frequency sub-bands, one obtains detailed information about. This methodology is known in the literature as multi-resolution analysis. In this figure, H is the low pass filter and G the high pass filter. The filtered signals are then sub-sampled by 2. The coefficients of the orthonormal decomposition pairs in the case of Daubechies 4 wavelet function can be expressed as follows:

\[
G(n) = \begin{bmatrix}
\frac{1-\sqrt{3}}{4\sqrt{2}}, & \frac{3-\sqrt{3}}{4\sqrt{2}}, & \frac{3+\sqrt{3}}{4\sqrt{2}}, & \frac{-1+\sqrt{3}}{4\sqrt{2}}
\end{bmatrix}
\]

(3.27)

\[
H(n) = \begin{bmatrix}
\frac{1+\sqrt{3}}{4\sqrt{2}}, & \frac{3+\sqrt{3}}{4\sqrt{2}}, & \frac{3-\sqrt{3}}{4\sqrt{2}}, & \frac{1-\sqrt{3}}{4\sqrt{2}}
\end{bmatrix}
\]

(3.28)

As per the parameter estimation method mentioned in [Do and Vetterli, 2002], the estimated GGD model parameters, i.e., the texture features, of the Daubechies (Db4) wavelet coefficient distribution in a \( M \times N \) subband \( X = \{x_{i,j}, i = 1, \ldots, M, j = 1, \ldots, N\} \), namely \( \hat{\alpha} \) and \( \hat{\beta} \), are obtained as follows,
Fig. 3.2 Wavelet based image decomposition (a) filtering approach for single level decomposition and (b) subband representation of three level wavelet decomposition.

\[ \alpha = \left( \frac{\beta}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} |x_{i,j}|^\rho \right)^{1/\beta} \]  

(3.29)
where $\beta$ is an approximation of $\hat{\beta}$ determined using the Newton-Raphson iterative procedure [Do and Vetterli, 2002]. In this work, the GGD model parameters are estimated from the details subbands of each level resulted in 18 feature components forming the feature vector for a given mammogram.

### 3.6 COHEN-DAUBECHIES-FEAUVEAU (CDF) 9/7 WAVELET

The lifting scheme implementation [Quellec et al., 2010a] consists of the split, predict, update and scaling steps as shown in Fig. 3.3. The function of each step is as follows: the split separates odd and even samples; predict estimates even samples values using the odd ones; update generates the approximation of odd samples using the prediction error; scaling normalizes the outputs. The lifting scheme for the CDF 9/7 wavelet includes two set of predict ($p_1$ & $p_2$) and update ($u_1$ & $u_2$) filters as shown in Fig. 3.3, whereas, one such set is used for the LeGall 5/3 wavelet. The values used for a given operation in the CDF 9/7 wavelet implementation is as follows: $\alpha \approx -1.58613432$ and $\delta \approx 0.8829110762$ for predicting; $\beta \approx -0.05298011854$ and $\gamma \approx 0.4435068522$ for updating; $\zeta \approx 1.149604398$ for scaling. The prediction errors, updated odd sets and normalized outputs are computed as given in Eq. (3.30), Eq. (3.31) and Eq. (3.32), respectively [Daubechies and Sweldens, 1998].
\[ x_e^1(n) = x_e(n) - \alpha(x_o(n+1) + x_o(n)) \; ; \]
\[ x_e^2(n) = x_e^1(n) - \delta(x_o^1(n+1) + x_o^1(n)). \] (3.30)
\[ x_o^1(n) = x_o(n) + \beta(x_e^1(n) + x_e^1(n-1)) \; ; \]
\[ x_o^2(n) = x_o^2(n) + \gamma(x_e^2(n) + x_e^2(n-1)), \] (3.31)
\[ x_o^3 = \zeta x_o^2 \; ; \quad x_e^3 = \zeta^{-1} x_e^2 \] (3.32)

The above procedure is repeated with \( x = x_o^3 \) for the next level decomposition. In the case of the Le Gall 5/3 wavelet, \( \alpha = 0.5, \beta = 0.25 \) and \( \zeta = 1 \) are the values used in the prediction, updation and scaling steps, respectively. As mentioned in [Quellec et al., 2010a], a maximum of three levels of decomposition is chosen for this work.

Next, as mentioned in [Do and Vetterli, 2002] the GGD model parameters from the details subbands of each level and 32-bin histogram from the third level approximation coefficients are extracted to generate a feature vector of size 50 for a given mammogram.

### 3.7 ZERNIKE MOMENTS

The mammograms are pre-processed and 256 Zernike moments are extracted without the need of previous image segmentation [Felipe et al., 2005]. For a digital image the respective Zernike moments of \( A_{pq} \) of order \( p \) with repetition \( q \) are computed as follows [Li et al., 2009]:

\[
A_{pq} = \frac{p+1}{\pi} \sum f(x_i,y_i)W^*(\rho,\theta), x_i^2+y_i^2 \leq 1
\] (3.33)

where \( i \) runs over all the image pixels, \( p \) is a non-negative integer, \( q \) is an integer subject to the constraint \( p - |q| \) = even, \( |q| \leq p \), \( ** \) denotes complex conjugation.

\[
V_{pq}(\rho,\theta) = R_{pq}(\rho) \exp(jq\theta); \rho = (x^2 + y^2), \theta = \tan^{-1}\left(\frac{y}{x}\right)
\] (3.34)
\[ R_{pq}(\rho) = \sum_{s=0}^{(p+q)/2} \frac{(-1)^s \left[ (p-s)! \right] \rho^{p-2s}}{s! \left( \frac{p+q}{2} - s \right)! \left( \frac{p-q}{2} - s \right)!} \] (3.35)

### 3.8 GENETIC ALGORITHM (GA)

Genetic Algorithm, a global search method for finding optimal solution to a problem is based on Darwin’s fittest principle, which states that the initial population of individuals evolves through natural selection in such a way that the fittest individuals have a higher chance of survival (Goldberg, 1989). Genetic Algorithm method is widely used for selecting features in problem with large complex data (Yang and Honavar, 1998), (Kudo and Sklansky, 2000). Holland’s first proposed genetic algorithm is called the simple genetic algorithm and the working of SGA in detail to find the optimal solutions is given in (Srinivas and Lalit, 1994). SGA is capable of minimizing the set of features statistically extracted in the feature extraction stage. Such minimization reduces

![Fig. 3.3: Genetic algorithm flow chart](image)
the dimensionality of the features (Stolpman and Dooley, 1998). In this work each chromosome, representing a feature set of a particular type is encoded as a binary string. Each element of the chromosome is called gene. Each gene refers to a particular feature represented as 1 or 0 when it is selected or not respectively. The chromosome length is nine and 10 for Db4 and CDF 9/7 wavelet based features, respectively. The population size considered is 20. The control parameters are tournament selection, single point crossover with 0.7 \( P_c \) (probability of crossover) and 0.001 \( P_m \) (probability of mutation) [Sun et al., 2005]. The stopping condition is either the number of generations (e.g 100) or the maximum fitness value, which is equal to the total number of queries. Tournament selection is one of the most popular selection operators in GA [Vladimir, 2003]. It is a method of selecting an individual from a population of individuals in a genetic algorithm. Tournament selection involves running several "tournaments" among a few individuals chosen at random from the population. The winner of each tournament (the one with the best fitness) is selected for crossover. Selection pressure is easily adjusted by changing the tournament size. If the tournament size is larger, weak individuals have a smaller chance to be selected. In the crossover condition, a crossover point is randomly chosen in the range 1 to \( L - 1 \), where \( L \) is the chromosome length. The portions of the two chromosomes beyond the crossover point are exchanged to form two offspring. Mutation involves flipping of a gene in a chromosome based on the mutation rate \( P_m \). SGA is terminated after the stopping condition is reached.

3.9 Student’s t-Test

To select the most discriminant features, a statistical multivariate \( t \)-test is used to assess the significance of the difference between the means of two sample set \( A \) and \( B \), which are independent of each other in the obvious sense, that is, the individual measures in set \( A \) are in no way related to any of the individual measures in set \( B \). The value of the \( t \)-test is obtained as follows [Serdobolskii, 2000]:
\[ D_a = \sum (A_i - \mu_a)^2 \]  
(3.36)

\[ D_b = \sum (B_i - \mu_b)^2 \]  
(3.37)

\[ V = \frac{D_a + D_b}{(n_a - 1) + (n_b - 1)} \]  
(3.38)

\[ \sigma = \sqrt{\frac{V}{n_a} + \frac{V}{n_b}} \]  
(3.39)

\[ t = \frac{\mu_a - \mu_b}{\sigma} \]  
(3.40)

where \( A_i \) and \( B_i \) in equations (3.36) and (3.37) are the \( i^{th} \) element of the set \( A \) and \( B \), while \( \mu_a \) and \( \mu_b \) are the means of the set \( A \) and \( B \), respectively. \( D_a \) and \( D_b \) in the equations (3.36) and (3.37) are the sum of squared deviates of the set \( A \) and \( B \). \( V \) in equation (3.38) is the estimated variance of the source population. \( \sigma \) in equation (3.39) is the standard deviation of the sampling distribution of sample mean differences. \( t \) in equation (3.40) is the value of the \( t \)-test. The degree of freedom (d.f.) is \((n_a - 1) + (n_b - 1)\).

In our case study with 20 normal images (set \( A \)) and 19 abnormal images (set \( B \)), the degree of freedom (d.f.) is 37. According to the Table of Critical Values of \( t \) [Rencher, 1998], the \( t \) value for 37 degrees of freedom (d.f.) is 1.305. When the value \( t \) obtained in this experiment is greater than 1.305, it means that there is a significant mean difference between normal and abnormal mammograms with regard to the given feature.

### 3.10 STATISTICAL ASSOCIATION RULE MINER (StARMiner) AND FRACTAL DIMENSION ATTRIBUTE SIGNIFICANCE ESTIMATOR (FD-ASE)

StARMiner algorithm identifies the most relevant features to discriminate images through mining association rules. The most relevant features, in this case, are the Zernike moments that hold the shape details responsible for the discrimination between image
classes. The algorithm uses statistical measures that describe the behaviour of the attributes in order to find rules of interest.

A statistical association rule is a type of rule that shows an interesting relationship among subsets of data based on the distribution of the quantitative values. Feature vectors describe the images quantitatively. Hence, a suitable approach to find association rules should consider quantitative data. The goal of StARMiner for statistical association rule mining, is to find statistical association rules to select a minimal set of features that preserves the ability of discerning image according to their types. Let \( x_j \) be a category of an image \( f_i \) an image feature (attribute). The rules returned by the StARMiner algorithm have the format \( x_j \rightarrow f_i \). The threshold values used are as follows: \( \Delta \mu_{\text{min}} \) - the minimum allowed difference between the average of the feature \( f_i \) in images from category \( x_j \) and the average of \( f_i \) in the remaining dataset; \( \sigma_{\text{max}} \) - the maximum standard deviation of \( f_i \) values allowed in a category; \( \gamma_{\text{min}} \) - the minimum confidence to reject the hypothesis \( H_0 \). StARMiner mines rules of the form \( x_j \rightarrow f_i \), if the conditions given in equations (3.41) – (3.43) are satisfied.

\[
\mu_{f_i}(V) = \frac{\sum_{k \in V} (f_{i_k})}{|V|} \quad \text{(3.41)}
\]

\[
\sigma_{f_i}(V) = \sqrt{\frac{\sum_{k \in V} (f_{i_k} - \mu_{f_i}(V))^2}{|V|}} \quad \text{(3.42)}
\]

\[
\mu_{f_i}(T_{x_j}) - \mu_{f_i}(T - T_{x_j}) \geq \Delta \mu_{\text{min}} \quad \text{(3.43)}
\]

\[
\sigma_{f_i}(T_{x_j}) \leq \sigma_{\text{max}} \quad \text{(3.44)}
\]

\[
H_0 : \mu_{f_i}(T_{x_j}) = \mu_{f_i}(T - T_{x_j}) \quad \text{(3.45)}
\]
Algorithm

Input: Dataset $T$ of image transactions structured as $\{x, f_1, f_2, f_n\}$ where $x$ represents the image category and $f_i$ an image feature; thresholds $\Delta \mu_{\min}, \sigma_{\max}, \gamma_{\min}$.

Output: The mined rules

1. Scan dataset $T$
2. for each feature $f_i$ do
3. for each category $x_j$ do
4. Calculate $\mu_{f_i}(T_{x_j})$ and $\mu_{f_i}(T - T_{x_j})$
5. end for
6. end for
7. Scan dataset $T$
8. for each feature $f_i$ do
9. for each category $x_j$ do
10. Calculate $\sigma_{f_i}(T_{x_j})$ and $\sigma_{f_i}(T - T_{x_j})$
11. Calculate $Z$ value
12. if $(\mu_{f_i}(T_{x_j}) - \mu_{f_i}(T - T_{x_j})) \geq \Delta \mu_{\min}$ And $\sigma_{f_i}(T_{x_j}) \leq \sigma_{\max}$ And ($Z < Z_1$ or $Z > Z_2$) then
13. Write $x_j \rightarrow f_i, \mu_{f_i}(T_{x_j}), \mu_{f_i}(T - T_{x_j}), \sigma_{f_i}(T_{x_j}), \sigma_{f_i}(T - T_{x_j})$
14. end if
15. end for
16. end for

Attribute selection is a classic goal, which is frequently used to overcome the “dimensionality curse” [Pagel et al., 2000] [Korn et al., 2010]. A carefully chosen subset of attributes improves the performance and efficacy of a variety of algorithms, including indexing, matching, and retrieval. This is particularly true with redundant data, as many
data sets can largely be well-approximated using fewer dimensions. Moreover, attribute selection can also be seen as a way to compress data, since only the attributes which maintain the essential characteristics of the data are kept [Fayyad, 1998].

Fractal dimension attribute significance estimator (FD-ASE) algorithm is used to apply dimensionality reduction of the feature vectors. Considering the contribution of each feature to increment the fractal dimension of the dataset, the algorithm finds dependence relationship between attributes and determines a set of independent ones, discarding the others. The approach of attribute forward inclusion is used to recreate the data set. The fundamental idea is to calculate the partial correlation fractal dimension of data set projections, integrating more and more attributes until its calculated correlation fractal dimension $D$ is achieved. The outline of the FD-ASE algorithm is as follows:

Step 1: Given the data set $A = \{a_1, a_2, ..., a_n\}$, composed by $n$ attributes identified by an index ranging from 1 to $n$, the constant attributes are discarded. To do this, the individual $pD$ of each attribute is computed and every attribute with $pD \approx 0$ is dropped.

Step 2: Calculate the $pD$ of the partial data set consisting of the attribute $a_1$, and progressively calculate the $pD$ of the partial data set resulting from including attributes $a_2$, $a_3$, and so on, until the correlation fractal dimension $D$ of the complete data set is achieved. However, it is possible that, when adding an attribute $a_k$, the fractal dimension does not increase. This means that the attribute $a_k$ is correlated with at least one of the attributes $a_i \mid i < k$, so attributes $a_1$ to $a_k$ define a correlation super group.

It is assumed, without loss of generality, that there is no other attribute $a_j$, $j < k$, such that its inclusion in the data set does not change the resulting $pD$. That is, $a_k$ is the first attribute in the sequence $(a_1, a_2, ..., a_n)$ such that $pD(a_1..a_k) \approx pD(a_1..a_{k-1})$. If the attributes in the sequence $(a_1..a_k)$ constitute a correlation group, there is no attribute in it that can be removed resulting in a different $pD$.

Step 3: The sequence $(a_2..a_k, a_1)$ is formed by having the attribute $a_1$ as the last element in the sequence, maintaining the other attributes in the same order.
Step 4: If \( pD(a_2..a_k) \approx pD(a_2..a_{k-1}) < pD(a_2..a_k, a_1) \), then \( a_1 \) is not an attribute in the same correlation group with \( a_k \), and the attribute \( a_1 \) is removed from the group; otherwise \( a_1 \) and \( a_k \) are in fact in the same correlation group. Steps 3 and 4 must be repeated for every attribute \( a_j \) in \((a_1..a_k)\), computing \( pD(a_1..a_{j-1}, a_{j+1}..a_k) \), \( pD(a_1..a_{j-1}, a_{j+1}..a_k) \) and \( pD(a_1..a_{j-1}, a_{j+1}..a_k, a_j) \). Whenever an attribute is found not pertaining to the target correlation group, it is eliminated from the sequence in the remaining iterations. Therefore, after at most \( 3^k \) calculations of the \( pD( ) \) algorithm, the first correlation group of attributes is discovered.

Step 5: The other attributes \( a_j \) in \( A, j > k \), are tested to determine if they pertain to the same correlation group as \( a_k \). Therefore, every attribute that does not increase the \( pD \) of the correlation group already found is included in this group.

Step 6: The correlation group \( G_1 \) already found is reserved. Being \( g \) the number of attributes in the group \( G_1 \), only the \( b \) attributes that compose the correlation base of the group is kept for the next execution of the algorithm. That is, \( g-b \) attributes are removed from the set of attributes being processed.

Step 7: Repeat steps 2 through 6 until every correlation groups are discovered, which are the most important attributes (dimensions) for a given feature set.

### 3.11 Similarity Measures

Similarity measurement is one of the major tasks of a typical CBIR system, where a distance between the query image and each of the images in the database using their signatures, i.e., features is computed so that required number of top most relevant images can be retrieved. The challenging task is to find good similarity measures between images based on some feature set [Do and Vetterli, 2002]. The simplest way to measure the similarity between two features is to calculate the distance between them using a certain distance metric, which is also one of the most popular methods and widely used in similarity estimation [Yu, Amores, et al., 2006]. At present there are many such measures for determining the dissimilarity between the feature vectors. Euclidean distance is not always the best metric. The fact that the distances in each dimension are squared before
summation, places great emphasis on those features for which the dissimilarity is large [Kokare et al., 2003]. Hence it is important to explore different similarity measures, which overcomes above problems and improve retrieval accuracy. There have been considerable efforts in finding the appropriate measures among such a surplus of choices [Duda et al., 2001]. Therefore, in this work, considerable effort has been taken to find the appropriate similarity measures for various features considered in order to improve the retrieval performance towards mammogram retrieval. This subsection gives the details of three categories of similarity measures, i.e., geometric, statistic and information theoretic, are tested for different feature types are given.

For geometric measures the distance between two feature vectors are based along pair-wise comparisons on dimensions. Statistical measures compare two feature vectors in a distributed manner rather than simple pair wise distance. Information-theoretic measures are various conceptual derivatives from the Shannon’s entropy theory and the feature vectors are the probabilistic distributions [Liu, Song, et al., 2008].

**Euclidean**

The basis of many measures of similarity and dissimilarity is Euclidean distance. It is also referred as Pythagorean metric. It is a true metric, as it satisfies the triangle inequality, and is the most widely used distance measure of all available. The Euclidean distance between feature vectors p and q is given by [Cha, 2007],

$$d_{Euc} = \sqrt{\sum_{i=1}^{d} (P_i - Q_i)^2}$$  \hspace{1cm} (3.46)

where d is the size of feature vector.

**City-block**

It is also known as the Manhattan distance or taxi cab distance, closely related to the Euclidean distance. Euclidean distance corresponds to the length of the shortest path between two points, whereas, the city-block distance is the sum of distances along each
dimension. The city-block distance is a metric, as it satisfies the triangle inequality. The city-block distance between two feature vectors is the sum of the absolute differences of their feature vector values, which is given as [Prasad et al., 2006],

\[ d_{CB} = \sum_{i=1}^{d} |P_i - Q_i| \]  

(3.47)

**Chebyshev**

Chebyshev distance, is also known as maximum metric or L∞ metric [Cyrus, 2000] is a metric defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension [James, Abello et al., 2002] is named after Pafnuty Lvovich Chebyshev [David et al., 2004]. The Chebyshev distance between two feature vectors p and q is given as Eq. 3.48,

\[ d_{Cheb} = \max_i |P_i - Q_i| \]  

(3.48)

**Sorensen**

Sørensen distance [Looman and Campbell, 1960], also known as Bray-Curtis distance [Bray and Curtis, 1957] is defined as Eq. 3.49,

\[ d_{sor} = \frac{\sum_{i=1}^{d} |P_i - Q_i|}{\sum_{i=1}^{d} (P_i + Q_i)} \]  

(3.49)

Its value is always in [0, 1] range. As compared to Euclidean distance, Sorensen distance retains sensitivity in more heterogeneous data sets and gives less weight to outliers [McCune, Bruce et al., 2002].
Gower

Gower distance scales the vector space into the normalized space and then uses the City-block distance which is divided by the feature size value. The Gower distance is defined as Eq. 3.50,

$$d_{gow} = \frac{1}{d} \sum_{i=1}^{d} |P_i - Q_i|$$

(3.50)

Soergel

The Soergel distance between two feature vectors is the sum of the absolute differences of their feature vector values; normalized by the sum of corresponding maximum feature components and is given as Eq. 3.51,

$$d_{sg} = \frac{\sum_{i=1}^{d} |P_i - Q_i|}{\sum_{i=1}^{d} \max(P_i, Q_i)}$$

(3.51)

Canberra

Canberra distance given in equation (3.52) resembles Sorensen distance but normalizes the absolute difference of the individual component. It is mostly used for data scattered around the origin, i.e., zero [Gordon, 1999].

$$d_{can} = \sum_{i=1}^{d} \frac{|P_i - Q_i|}{P_i + Q_i}$$

(3.52)

Cosine

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them and thus often called the angular metric. In the case of information retrieval, the cosine similarity of two feature vector will range from 0 to 1. The cosine distance is insensitive to vector length,
i.e., useful over Euclidean/L_k distances when length has no important meaning to a given comparison. The cosine distance is defined as follows in Eq. 3.53,

\[
s_{\cos} = \frac{\sum_{i=1}^{d} p_i q_i}{\sqrt{\sum_{i=1}^{d} p_i^2} \sqrt{\sum_{i=1}^{d} q_i^2}}
\]  

(3.53)

**Bhattacharya**

Bhattacharyya distance measure is named after A. Bhattacharya, a statistician who worked in the 1930s at the Indian Statistical Institute. The equation (3.54) expressing Bhattacharyya distance has a value between 0 and 1, provides bounds on the Bayes misclassification probability [Bhattacharyya, 1943].

\[
d_B = -\ln \sum_{i=1}^{d} \sqrt{p_i q_i}
\]  

(3.54)

**Squared-chord**

Squared-chord distance between two feature vectors is the sum of the squared differences of the square root the individual feature component, which is defined as Eq. 3.55,

\[
d_{sqc} = \sum_{i=1}^{d} (\sqrt{p_i} - \sqrt{q_i})^2
\]  

(3.55)

**Pearson’s Correlation Coefficient**

There are different forms of the Pearson correlation coefficient formula. For the given feature vectors P and Q, a commonly used form is [Huang, 2008],

\[
d_p = 1 - |\rho|
\]  


where, \( \rho = \frac{\sum_{i=1}^{d} P_i Q_i - \sum_{i=1}^{d} P_i \sum_{i=1}^{d} Q_i}{\sqrt{(\sum_{i=1}^{d} P_i - \sum_{i=1}^{d} Q_i)^2 - (\sum_{i=1}^{d} Q_i - \sum_{i=1}^{d} Q_i)^2}}. \) (3.56)

The Pearson’s correlation coefficient ‘\( \rho \)’ reflects the degree of linear relationship between two patterns. It ranges between -1 to +1, reflecting respectively a perfect negative (positive) linear relationship between the patterns. A zero correlation value implies that there is no linear relationship between the two patterns. The larger the |\( \rho \)| value is, closer the feature vectors \( P \) and \( Q \).

**Mahalanobis**

Mahalanobis distance between two feature vector \( P \) and \( Q \) is computed by [Chen and Chu],

\[
d_M = \sqrt{(P_i - Q_i)^T S^{-1} (P_i - Q_i)}
\]

\[
S = \frac{1}{n} \sum_{i=1}^{d} (P_i - \mu)(P_i - \mu)^T
\]

\[
\mu = \frac{1}{d} \sum_{i=1}^{d} P_i
\] (3.57)

where, \( S \) and \( \mu \) are the sample covariance matrix and mean, respectively. It gives more weight to feature vector with smaller variance and less weight to one with larger variance.

**Chi-square statistic**

Chi-square statistic measure is similar to Euclidean distance, but, it is weighted by the average of the individual feature components. If the feature vectors \( P \) and \( Q \) are closer then the Chi-square value is smaller. It is defined as [Puzicha, et al., 1997],
Fractional distance metrics provide better divergence between the maximum and minimum distances to a given query point than integral distance metrics. This feature makes a proximity query more meaningful and stable. Fractional distance measures have been applied to content-based image retrieval and the experiments also show that retrieval performances of these measures consistently outperform the Manhattan and Euclidean distance metrics when used with a wide range of high-dimensional visual features [Howarth and Rüger, 2005]. It is defined as [Aggarwal et al., 2001],

\[
d_f = \left[ \sum_{i=1}^{d} (P_i - Q_i)^p \right]^{1/p}, \quad p \in (0,1)
\]

**Kullback–Leibler (KL)**

It was originally introduced by Solomon Kullback and Richard Leibler in 1951 as the directed divergence between two distributions. Kullback-Leibler (KL) divergence or relative entropy is a distance measure between two probability distributions \( P(i) \) and \( Q(i) \). The KL divergence is defined as follows [Tourassi, Harrawood, et al., 2007],

\[
D_{KL}(P \parallel Q) = \sum_i \ln \left( \frac{P(i)}{Q(i)} \right) P(i)
\]

In words, it is the expectation of the logarithmic difference between the probabilities P and Q, where the expectation is taken using the probabilities P. The K-L divergence is only defined if P and Q both sum to 1 and if \( Q(i) = 0 \) implies \( P(i) = 0 \) for all \( i \) (absolute continuity). Generally, the higher the KL divergence is, the more dissimilar the two images are.
Symmetric KL (SKL)

KL divergence is not a true distance measure because it is not symmetric (i.e., \(D(q||p) \neq D(p||q)\)). Consequently, different transformations have been utilized in CBIR to provide a symmetric KL divergence measure [Ojala et al., 1996]. Symmetrical KL divergence is equal to zero when the two probability distributions are identical. In this work, two such transformations are explored: (i) average KL divergence and (ii) maximum KL divergence, which are defined as in Eqs.3.61 and 3.62,

\[
D_{AKL} = \frac{D(q \parallel p) + D(p \parallel q)}{2} \quad (3.61)
\]

\[
D_{MKL} = \max \left[ D(q \parallel p), D(p \parallel q) \right] \quad (3.62)
\]

Jensen

The Jensen divergence is an empirical modification of the KL divergence that is symmetric and more robust with respect to noise and histogram binning [Puzicha et al., 1997] and is defined as Eq. 3.63,

\[
D_{JD} = \sum_i \left( q(i) \log \frac{2q(i)}{p(i)+q(i)} + p(i) \log \frac{2p(i)}{p(i)+q(i)} \right) \quad (3.63)
\]

Jensen divergence has values bounded between 0 and 2.

Arithmetic-geometric mean (AGM)

This distance measure is essentially the KL divergence between the arithmetic and geometric mean of the two image distributions \(p(i)\) and \(q(i)\). It is defined as Eq. 3.64,

\[
D_{AGM} = \sum_i \frac{p(i) + q(i)}{2} \log \frac{p(i) + q(i)}{2\sqrt{p(i)q(i)}} \quad (3.64)
\]
**Jeffreys divergence**

The *Jeffrey-divergence* (JD) is a symmetric version of Kullback-Leibler divergence defined by, [Kullback and Leibler, 1951], [Jeffreys, 1946], [Taneja, 2001],

\[
D_{jeff} = \sum_i (p(i) - q(i)) \ln \frac{p(i)}{q(i)}. \tag{3.65}
\]

For \(p(i) \to q(i)\), the Jeffreys divergence behaves like the \(\chi^2\) distance (Michel Marie Deza and Elena Deza, 2009).

**K divergence**

This divergence measure is closely related to I divergence. Its value is \(\geq 0\) and equal to zero if and only if \(p(i) = q(i)\), which is essential for a measure of difference. K divergence is defined as follows [Lin,1991],

\[
D_K = \sum_i p(i) \ln \frac{2p(i)}{p(i) + q(i)}. \tag{3.66}
\]

**Topsoe**

Topsoe divergence is the result of symmetric form of K divergence obtained using the addition method, which is defined as Eq. 3.67,

\[
D_{top} = \sum_{i=1}^d \left( p(i) \ln \left( \frac{2p(i)}{p(i) + q(i)} \right) + q(i) \ln \left( \frac{2q(i)}{p(i) + q(i)} \right) \right). \tag{3.67}
\]

**Jensen-Shannon**

In probability theory and statistics, the Jensen–Shannon divergence is one of the popular methods of measuring the similarity between two probability distributions. It is also known as information radius [Hinrich and Christopher, 1999] or total divergence to
the average [Dagan et al., 1997]. Jensen–Shannon divergence (JSD) is a symmetrized and smoothed version of the Kullback–Leibler divergence defined by Eq. 3.68,

\[
D_{JS} = \frac{1}{2} \left[ \sum_{i=1}^{d} p(i) \ln \left( \frac{2p(i)}{p(i)+q(i)} \right) + \sum_{i=1}^{d} q(i) \ln \left( \frac{2q(i)}{p(i)+q(i)} \right) \right].
\] (3.68)

**Jensen difference**

Jensen-difference measure originates from concavity property of Shannon’s entropy and its concavity property [Burbea and Rao, 1982]. It is also known as information radius [Sibson, 1969]. The Jensen difference between probability vectors is given by Eq. 3.69,

\[
D_{JD} = \sum_{i=1}^{d} \left[ \frac{p(i) \ln p(i) + q(i) \ln q(i)}{2} - \left( \frac{p(i)+q(i)}{2} \right) \ln \left( \frac{p(i)+q(i)}{2} \right) \right].
\] (3.69)

**Modified Kull-back Leibler divergence**

Given the generalized Gaussian density (GGD) model, the probability density function (PDF) of wavelet coefficients in each subband is defined via two parameters \( \alpha \) and \( \beta \). Thus, the closed form for the Kullback-Leibler divergence between two GGDs is obtained as [Do et al., 2002],

\[
D(p(.; \alpha_1, \beta_1) \| p(.; \alpha_2, \beta_2)) = \log \left( \frac{\beta_1 \alpha_2 \Gamma \left( \frac{1}{\beta_2} \right)}{\beta_2 \alpha_1 \Gamma \left( \frac{1}{\beta_1} \right)} \right) + \left( \frac{\alpha_1}{\alpha_2} \right)^{\beta_2} \frac{\Gamma \left( \frac{\beta_2+1}{\beta_1} \right)}{\Gamma \left( \frac{1}{\beta_1} \right)} - \frac{1}{\beta_1}
\] (3.70)

Hence, the similarity measurement between two wavelet subbands can be calculated very effectively using the model parameters. The overall similarity distance
between two images is precisely the sum of KLDs given in equation 3.71 between corresponding pairs of subbands, which is defined by [Do and Vetterli, 2002],

\[ D(I_1, I_2) = \sum_{j=1}^{B} D(p(.; \alpha_1^{(j)}, \beta_1^{(j)})) \| p(.; \alpha_2^{(j)}, \beta_2^{(j)}) ) \] (3.71)

where, \( B \) is the number of analyzed subbands.

3.12 CONCLUSION

Wei et al., (2006) has considered only 11 Haralick texture features, whereas in this research all the 14 features are considered. Two Gabor filter based approaches are considered, in which the work referred by Manjunath et al., (1996) is used in the feature extraction analysis and the one by Wei et al., (2005) is used in the feature selection analysis. In the case of Zernike moments extraction the approach mentioned by Felipe et al., (2005) is adopted in order to avoid prior segmentation process. In this research work the fixed CDF 9/7 wavelet is considered instead of adaptive type mentioned in [Quellec et al., 2010] due to the ambiguity in fixing the range of filter coefficients range. In the case of GA the Roulette-wheel selection is used. The above methods selected as competitors of this research work finds to be applied and compared in most of the work related to mammogram retrieval problem. The performance results of these state-of-the-art techniques are presented and analyzed in the result and discussion chapter.