Mammographic Lesion Classification using Discrete Orthonormal S-Transform

Susnata Roy

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India.

Mammographic Lesion Classification using Discrete Orthonormal S-Transform

Thesis submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Technology

in

Computer Science and Engineering

by

Susnata Roy
(Roll: 110CS0145)

under the guidance of

Prof. Banshidhar Majhi

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India.
May 2014

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Orissa, India.

May 5, 2014

Certificate

This is to certify that the work in the thesis entitled ***Mammographic Lesion Classification using Discrete Orthonormal S-Transform*** by ***Susnata Roy*** is a record of an original research work carried out under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

Banshidhar Majhi
Professor

Acknowledgment

This thesis has been possible due to the help and endeavor of many people.

Foremost, I would like to express my gratitude towards my project advisor, Prof. Bansidhar Majhi, whose mentor-ship has been paramount, not only in carrying out the research for this thesis, but also in developing long-term goals for my career. His guidance has been unique and delightful. He provided his able guidance whenever I needed it. Yet he always inspired me to be an independent thinker, and to choose and work with independence.

I am grateful to Prof. Ratnakar Dash for providing useful insights into my project.

I am also very thankful to Mr. Shraddhananda Beura for listening, offering me advice, and gladly extending his support throughout the process.

I would also like to extend special thanks to my project review panel for their time and attention to detail. The constructive feedback received has been keenly instrumental in improvising my work further.

I would like to thank other researchers in my lab and my friends for their encouragement and understanding.

My parents receive my deepest love for being the strength in me.

Susnata Roy

Abstract

Breast cancer is the leading cause of cancer in women. Early detection of breast cancer through periodic screening improves the chances of recovery. However, the small and subtle signs of the early disease make the task of accurate diagnosis particularly arduous for radiologists. Computer aided diagnosis of the mammographic images is currently very popular as it helps radiologists classify lesions as normal or abnormal, benign or malignant.

This thesis presents an efficient mammographic lesion classification approach for the detection of breast cancer. The approach uses the two dimensional discrete orthonormal S-transform (DOST) method to extract the coefficients from the digital mammograms. A feature selection algorithm based on statistical two-sample t-test method is used for the selection of significant coefficients from the high dimensional DOST coefficients. The selected significant coefficients are used as features for the classification of mammographic lesions as benign or malignant. This scheme utilizes a back propagation neural network as the classifier. The scheme is validated using MIAS database. The result shows an optimal classification accuracy rate of 97.4% and a performance index value of AUC = 0.97 in receiver operating characteristic (ROC) curve. These results are very promising in comparison with existing discrete wavelet transform (DWT).

Contents

Certificate ii

Acknowledgement iii

Abstract iv

List of Figures vii

1 Introduction 1

2 Literature Survey 5

3 Discrete Orthonormal S-Transform 9

3.1 Signals and their types 9

3.2 Wavelet Transform 10

3.3 S-Transform 12

3.4 Discrete S-Transform 12

3.5 Properties of S-Transform 12

3.6 Advantages of S-Transform 14

3.7 DOST 15

4 Proposed Method 17

4.1 Materials and methods 17

4.1.1 Mammogram dataset 18

4.1.2 Image Preprocessing 19

4.1.3 Feature Extraction 20

4.1.4 Feature Selection 22

4.1.5 Feature Classification 24

5 Simulation and Results 26

6 Conclusion and Future Work **35**

Bibliography **36**

List of Figures

1.1 Two types of views of mammogram. (a) CC view of left breast, (b) CC view of right breast, (c) MLO view of left breast, (d) MLO view of right breast. 2

1.2 CAD for lesion classification. 3

3.1 Decomposition of the ROI into subcomponents using 2D-DWT. (a) Wavelet decomposition at two resolution level, (b) Original ROI (mdb015), (d) Subcomponents of ROI (wavelet coefficients). 11

4.1 Block diagram of proposed scheme for classification of mammographic images using two dimensional discrete orthonormal S-transform (DOST) and back propagation neural network (BPNN). 17

4.2 Extracted ROIs from different mammographic images (source: MIAS database). The numbered sub figures indicate the type of lesions; such as: [a, b, c, d] - benign type (mdb015, mdb081, mdb107, mdb219), [e, f, g, h] - malignant type (mdb028, mdb117, mdb115, mdb186). 19

4.3 Partitioning of (a) DWT and (b) DOST for six orders. The squares indicate the sub-images for each order. Both transforms use a dyadic sampling scheme but provide different information about the frequency content of the image. 21

5.1 ROC curves. (a) Classification of benign and malignant mammograms by DOST method, (b) Prediction of malignant mammograms by DOST and DWT method. 28

5.2 Classifier performance in terms of mean squared error(mse). (a) By DOST method, (b) comparison of training error between DOST and DWT method 30
5.3 Regression graph at $\alpha = 7 \times 10^{-4}$. (a) For DOST method, (b) for DWT method. 31
5.4 Comparison of different regression values of both DOST and DWT methods for different values of α. 32
5.5 (a). Comparison of gradients of classification (b). Comparison of Mean square errors of classification, at all the reduced sets of feature used in simulation for DOST method. 34

Chapter 1

Introduction

Breast cancer is a leading cause of death among women. According to the American Cancer Society, approximately 232,340 new cases of invasive breast cancer and 39,620 breast cancer deaths are expected to occur among US women in 2013 [1]. One in eight women in the United States will develop breast cancer in her lifetime.

The situation is no different in India. Breast cancer is set to overtake cervical cancer as the most common cancer in women in India in 2020. The Lancet reported an impending cancer epidemic in India: by 2020, 70 percent of those suffering from cancer worldwide will be located in poor countries, with a fifth located in India [2].

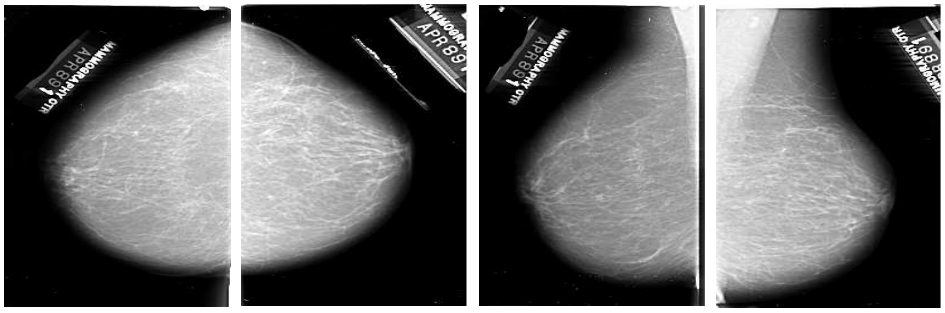
Another study commissioned by GE Healthcare, estimated that by 2030, the incidence of new cases of breast cancer in India will increase from today's figure of 115,000 to around 200,000 per year [3].

An analysis of cancer rates between the years 1982 and 2005, as conducted by The Indian Council of Medical Research, showed that 10 out of every 100,000 women living in Delhi, Mumbai, Chennai, and Bangalore were diagnosed with breast cancer about 10 years ago, compared with 23 women per every 100,000 today [4].

A number of well recognized exogenous and endogenous risk factors have been associated with the genesis of breast cancer. The exogenous factors include high fat diet, lack of physical activity, alcohol abuse, cigarette smoking, socio-economic

status, environmental exposures to pollutants, pesticides, electromagnetic field and ionizing radiations. Among the endogenous factors, the duration of exposure to steroid hormones play a vital role. This is in turn depends on several factors such as late menopause, late pregnancy, and obesity.

Early detection of breast cancer through periodic screening improves the chance of recovery. Mammography is currently the most effective method for a reliable early detection of breast cancer by analyzing the mammograms [5]. The American Cancer Society recommends women aged 40 and above to have a mammogram every year and calls it a gold standard for breast cancer detection [1].Mammograms are the X-ray images of breasts.



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

Figure 1.1: Two types of views of mammogram. (a) CC view of left breast, (b) CC view of right breast, (c) MLO view of left breast, (d) MLO view of right breast.

Interpretation of mammograms is a very important task for radiologists as they suggest patients for biopsy.

1. However, interpretation of mammograms varies among radiologists as it depends on training and experience. This leads to different judgments by different radiologists.

2. Furthermore, differences in image quality, along with the small and subtle signs of the early disease make the task of diagnosis particularly arduous.

3. There always exists possibility of human error due to a number of factors such

as fatigue, distraction and oversight, leading to interobserver and intraobserver variations.

4. Computer aided diagnosis of mammographic images helps improve both the sensitivity and the specificity of the diagnosis.

5. Therefore avoidance of misinterpretation is highly required. It has been observed that $60-90\%$ of the biopsies of suspected cancers by radiologists turned out to be benign [6].

Therefore, computer-aided diagnosis (CAD) is currently a very popular and efficient method which analyzes the digital mammograms and helps radiologists in interpreting mammograms for detection of suspicious lesions and classification.

Regarding this responsibility, one important step is to find out a set of significant features from the mammography images that can distinguish the benign lesions from malignant ones. Different techniques and methods have been studied for the extraction of features and classification of mammograms into benign and malignant classes.

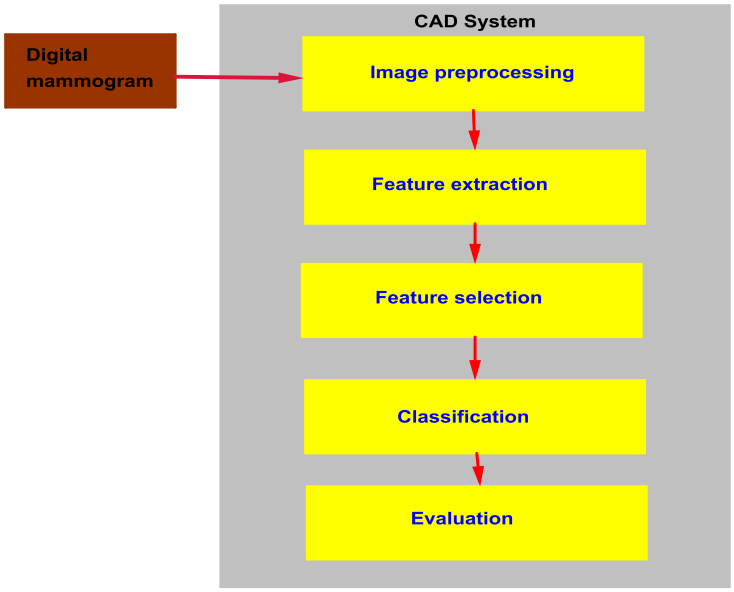


Figure 1.2: CAD for lesion classification.

This thesis is organized as follows. In Chapter 2, the existing work related to computer aided classification of mammographic images has been discussed. In Chapter 3, we introduce the discrete orthonormal S-Transform. We begin with describing the S-Transform in 3.3, its properties and the advantages of S-Transform over other multi-resolution techniques. We then discuss the drawbacks of naive S-Transform and introduce the DOST method in 3.7, which is a faster variant of S-Transform. In Chapter 4, we describe the proposed method for classification of mammograms as benign or malignant. Sub-steps such as preprocessing, feature extraction, subset selection and classification have been elaborated. Results and Simulation are presented in Chapter 5. Finally, Chapter 6 presents the concluding remarks, with the scope for further research work.

Chapter 2

Literature Survey

Liu *et al.* achieved 84.2% accuracy rate by using a set of statistical features obtained using linear phase non-separable linear wavelet transform. Detection is performed from the coarsest resolution to the finest resolution using a binary tree classifier [7].

Pereira *et al.* proposed a method in which they used spatial gray level dependence matrix on wavelet transform of mammograms [8]. They used the texture features to characterize the mammograms as benign or malignant with the help of non parametric K-NN classifier. They found the performance index value of AUC = 0.617 for masses and 0.607 for microcalcifications in ROC curves which were very poor.

M. Fraschini used discrete wavelet transform and neural network to discriminate the benign and malignant masses of regions of interest in mammograms [9]. The performance index value was AUC of 0.91 in the ROC curve.

Talha *et al.* proposed a method to classify the mammographic lesions into benign or malignant for the detection of breast cancer through discrete wavelet transform [10]. They used principal component analysis technique to reduce the wavelet based features and obtained more than 90% classification accuracy. Mammographic Institute Society Analysis dataset has been used for experimentation.

Pratibha *et al.* obtained a classification accuracy of 90.65% for benign and malignant characterization of mammogram samples [11]. They used the combination

of wavelet features and spectral features to analyze the mammograms. Support vector machine was used as the classifier in their approach.

Sertbas *et al.* proposed a method to classify the mammogram masses as benign or malignant by using local seed region growing spherical wavelet transform [12]. They obtained 91.67% classification accuracy rate by using SVM classifier.

Kumar *et al.* proposed a method based on discrete wavelet transform and stochastic neighbor embedding technique [13]. They achieved 90.10% of classification accuracy for the classification of benign and malignant mammograms with the help of support vector machine classifier. In their method the wavelet coefficients of mammograms are reduced by stochastic neighbor embedding technique.

Ganesan *et al.* proposed a one-class classification method to classify the mammographic images as benign or malignant [14]. They used Trace transform, which is a generalization of the Radon transform, to extract the features from the mammograms. Classifiers such as the linear discriminant classifier, quadratic discriminant classifier, nearest mean classifier, support vector machine and Gaussian mixture model have been used. They achieved a maximum accuracy rate of 92.48% by using Gaussian mixture model.

In 1996, Dhawan *et al.* used wavelet transform and gray level image structure features for classification of mammograms. They obtained receiver operating characteristic (ROC) index A_z of 0.81 [15].

In 1998, Chan *et al.* used texture morphology features for classification of mammograms based on gray level co-occurrence matrix (GLCM) and achieved $A_z = 0.89$ of ROC [16].

Manrique *et al.* used a genetic algorithm based on radial basis function neural

network for classification of masses in the year 2006. They achieved 83% classification accuracy rate [17].

Abdalla *et al.* used statistical texture features of mammographic images for the classification purpose in the year 2007. They obtained 82% accuracy with the help of support vector machine (SVM) classifier [18].

Rashed *et al.* have used different types of Daubechies wavelets for feature extraction of mammograms in the year 2007. They achieved 87.06% classification accuracy [19].

Dong *et al.* proposed a method using Gabor filter to classify normal and abnormal in the year 2009. They achieved an average of 80% precision with selected features [20].

Buciu *et al.* used Gabor wavelets to extract directional features from mammographic images in the year 2011 [21]. They used Principal component analysis technique to reduce the feature dimension and achieved 78.26% recognition rate in classification of benign and malignant lesions with the help of SVM.

Mutaz *et al.* used second order statistics and artificial neural network for detection of masses in digital mammograms and obtained an accuracy of 87.92% in the year 2011 [22].

Ramos *et al.* used wavelet transform method for feature extraction of mammograms in the year 2012 [23] The used the genetic algorithm for feature selection and random forest as classifier. They obtained $A_z = 0.90$ as performance index value of ROC.

It has been observed from literature that different feature extraction and selection methods are used for the classification of lesion. Still there is a scope to improve the

classification accuracy. The feature extraction and selection are key steps in lesion classification since these influence the performance of CAD.

So there is a need to develop some new feature extraction and selection method as well as classifiers to increase the accuracy of classification and reduce the complexity.

In this thesis, two dimensional discrete orthonormal S-transform (DOST) method is applied for the extraction of features from mammographic images. From the available set of extracted features, some effective set of features are selected and provided to a back propagation neural network classifier to predict the mammographic lesion as benign or malignant.

Chapter 3

Discrete Orthonormal S-Transform

3.1 Signals and their types

In image processing, a signal is a physical quantity which varies with space and contains information about space.

Signals may broadly be classified into the following two types:

1. Stationary Signals- Stationary signal are the signals which have all the frequency components present at all times of the signal.

2. Non-Stationary Signals- Non-stationary signals are the signals in which all the frequency components are not present at all the times in the signal.

An image is a non-stationary signal. Image consists of edges which divide it into regions. Smooth regions in the image have dominant low frequency components while edges have dominant high frequency components. Since, an image neither consists only of smooth regions, nor only of edges, but a mixture of both, an image is essentially a non-stationary signal.

To analyze a non-stationary signal such as image, we need multi-resolution techniques. Multi-resolution techniques give us time-frequency representation (TFR).

Many multi-resolution techniques exist. Some of them are :

1. Short Time Fourier Transform

2. Wavelet Transform

3. S-Transform

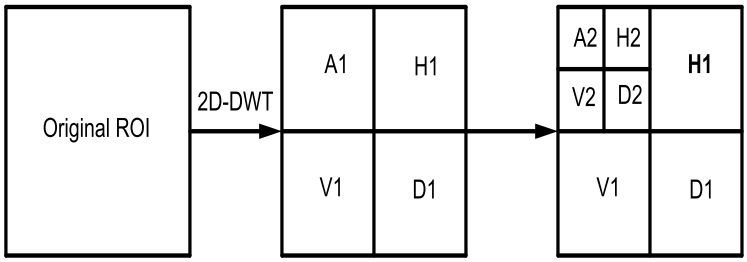
S-Transform has many advantages over Short Time Fourier Transform and Wavelet Transform. A detailed discussion on this is presented in Section 3.6.

3.2 Wavelet Transform

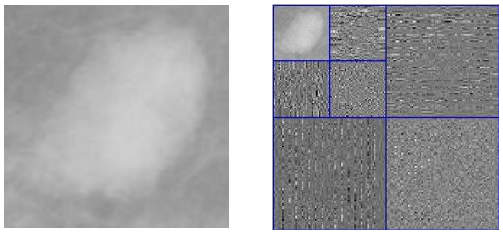
The Continuous Wavelet Transform can be defined as a series of correlations of the time series with a function called a wavelet:

$$W(\tau, d) = \int_{-\infty}^{\infty} h(t) w(t - \tau, d) dt \tag{3.1}$$

Wavelet transform has been widely used in the extraction of features from mammograms. A two dimensional discrete wavelet transform (DWT) is a multi-resolution decomposition method in which an original image $A_{2^{j+1}} f$ at resolution 2^{j+1} is decomposed to three detail images $D^h_{2^j} f$, $D^v_{2^j} f$, $D^d_{2^j} f$ at resolution 2^j in horizontal, vertical, and diagonal directions respectively. It also gives one approximation image $A_{2^j} f$ at coarse resolution. The detail and approximation images are the wavelet coefficient matrices in which each coefficient is considered as a feature of the original image.



(a)



(b) (c)

Figure 3.1: Decomposition of the ROI into subcomponents using 2D-DWT. (a) Wavelet decomposition at two resolution level, (b) Original ROI (mdb015), (d) Subcomponents of ROI (wavelet coefficients).

3.3 S-Transform

The S-Transform for continuous 1-dimensional signal $h(t)$ is given by :

$$S(\tau, f) = \int_{-\infty}^{\infty} h(t) \frac{\mid f \mid}{\sqrt{2\pi}} e^{-\frac{(\tau-t)^2 f^2}{2}} e^{-2i\pi ft} dt \tag{3.2}$$

3.4 Discrete S-Transform

The S-Transform of a discrete 2-dimensional signal $f(x, y)$ is given by:

$$S(x, y, k_x, k_y) = \sum_{\alpha=0}^{M-1} \sum_{\beta=0}^{N-1} F(\alpha + k_x, \beta + k_y) e^{-2\pi^2(\frac{\alpha^2}{k_x^2} + \frac{\beta^2}{k_y^2})} e^{2\pi i(\alpha x + \beta y)} \tag{3.3}$$

Here,

- x corresponds to x-coordinate in space.
- y corresponds to y-coordinate in space.
- k_x corresponds to frequency along x-axis.
- k_y corresponds to frequency along y-axis.
- F is the Fourier transform of original image.

3.5 Properties of S-Transform

1. **Absolutely Referenced Phase Information** : The phase factor $e^{i2\pi ft}$ helps to get absolutely referenced phase information. This phase factor splits the mother wavelet into two parts, Gaussian window and oscillatory exponential kernel $e^{-i2\pi ft}$. The kernel remains stationary while Gaussian window moves. Kernel being stationary, localizes the real and imaginary components of spectrum independently, thus localizing amplitude and phase of spectrum independently.

2. **Relation to Fourier Transform** : The S-Transform is related to Fourier transform in the following way:

$$H(f) = \int_{-\infty}^{\infty} S(\tau, f) d\tau \tag{3.4}$$

Thus, this relationship can be used to calculate Inverse S-Transform.

$$h(t) = \int_{-\infty}^{\infty} \{ \int_{-\infty}^{\infty} S(\tau, f) d\tau \} e^{i2\pi ft} df \tag{3.5}$$

3. **Instantaneous Frequency** : An extension of instantaneous frequency is provided by the S-Transform. S-Transform can be written in polar notation as

$$S(\tau, f) = A(\tau, f) e^{\Phi(\tau, f)} \tag{3.6}$$

where,

$$A(\tau, f) = \sqrt{Real(S(\tau, f)) + Im(S(\tau, f))} \tag{3.7}$$

and

$$\Phi(\tau, f) = tan^{-1} \{ \frac{Im(S(\tau, f))}{Real(S(\tau, f))} \} \tag{3.8}$$

Thus, Instantaneous Frequency (IF) is given by,

$$IF(\tau, f_0) = \frac{1}{2\pi} \frac{d}{d\tau} \{ 2\pi\tau f_0 + \Phi(\tau, f_0) \} \tag{3.9}$$

4. **Linearity** : S-Transform is a linear operation. Thus,

$$ST\{g(t) + h(t)\} = ST\{g(t)\} + ST\{h(t)\} \tag{3.10}$$

Proof of Linearity :

$$ST\{g(t) + h(t)\} = S(\tau, f) = \frac{\mid f \mid}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \{g(t) + h(t)\} e^{-\frac{(t-\tau)^2 f^2}{2}} e^{-i2\pi ft} dt \tag{3.11}$$

which can be rewritten as

$$ST\{g(t) + h(t)\} = S(\tau, f) \quad = \quad \{ \frac{\mid f \mid}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(t) e^{-\frac{(t-\tau)^2 f^2}{2}} e^{-i2\pi ft} dt \}$$

$$+ \{\frac{\mid f \mid}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(t) e^{-\frac{(t-\tau)^2 f^2}{2}} e^{-i2\pi ft} dt\} \quad (3.12)$$
$$= ST\{g(t)\} + ST\{h(t)\}$$

Thus,

$$ST\{g(t) + h(t)\} = ST\{g(t)\} + ST\{h(t)\} \quad (3.13)$$

3.6 Advantages of S-Transform

1. The Short Time Fourier Transform (STFT) has a fixed resolution but S-Transform gives a good time resolution for high frequency components and good frequency resolution for low frequency components, which is best suited for images. S-Transform is equivalent to applying several STFT with different sized windows. Thus, S-Transform is superior to STFT.

2. Wavelet Transform gives phase information local to translated window but S-Transform gives absolutely referenced phase information, which can be used for evaluating phase congruency. It has already been explained in Section 3.5, Property 1.

3. S-Transform can be used for denoising images containing additive noise. For this purpose, we can use the linearity property of S-Transform described in Section 3.5, Property 4.

4. S-Transform is directly related to Fourier Transform but Wavelet Transforms are not related to Fourier Transform. Relationship between S-Transform and Fourier Transform has already been explained in Section 3.5, Property 2. Thus, S-Transform is invertible but not all Wavelet Transforms are invertible.

5. S-Transform also provides superior time resolution compared to wavelet resolution.

3.7 DOST

The S-Transform is more powerful than other multi-resolution techniques like STFT and Wavelet Transform. The phase of the S transform referenced to the time origin provides useful and supplementary information about spectra that is not available from locally referenced phase information in the CWT [24]. The S-transform is advantageous for the analysis of mammographic images as it preserves the phase information using linear frequency scaling.

1. However, the major limitation of S-transform is its high time and space complexity due to its redundant nature, which makes it impractical in many cases.

2. The 2D-ST of an array of size $N \times N$ has a computational complexity of $O\left(N^4 + N^4 \log N\right)$ and storage requirements of $O\left(N^4\right)$.

3. To eliminate this problem of 2D-ST, in this thesis, we use DOST which is also a multi resolution technique for extraction of features from the mammographic images and is based on the S-transform.

4. DOST uses an orthonormal set of basis functions, and therefore, DOST has less computational and storage complexity in comparison to the S-transform, while retaining all the advantageous properties of S-Transform.

5. 2D-DOST provides a spatial frequency representation of an image, with computational and storage complexity as $O\left(N^2 + N^2 \log N\right)$ and $O\left(N^2\right)$ respectively.

With the dyadic sampling scheme in order $0, 1, 2, ..., \log_2 N - 1$, DOST of an $N \times N$ mammogram image $f(x, y)$ is performed by the following steps.

1. Two dimensional Fourier transform (FT) is applied to the image $f(x, y)$ to obtain Fourier samples $F(u, v)$

2. Partition the Fourier sample $F(u, v)$ and multiply it by the square root of the number of points in the partition, and perform an inverse FT. Then the voice image is calculated as
$S(x', y', \upsilon_x, \upsilon_y) = \frac{1}{\sqrt{2^{p_x+p_y-2}}} \sum_{u=-2^{p_x-2}}^{2^{p_x-2}-1} \sum_{v=-2^{p_y-2}}^{2^{p_y-2}-1} F(u+\upsilon_x, v+\upsilon_y)\ e^{2\pi i\left(\frac{ux'}{2^{p_x-1}}+\frac{vy'}{2^{p_y-1}}\right)}$ Here $\upsilon_x = 2^{p_x-1}+2^{p_x-2}$ and $\upsilon_y = 2^{p_y-1}+2^{p_y-2}$ are the horizontal and vertical voice frequencies.

3. Thus, the DOST coefficients of mammogram images are obtained after the transformation.

Chapter 4

Proposed Method

4.1 Materials and methods

The overall block diagram of the proposed method is shown in Figure. 4.1.

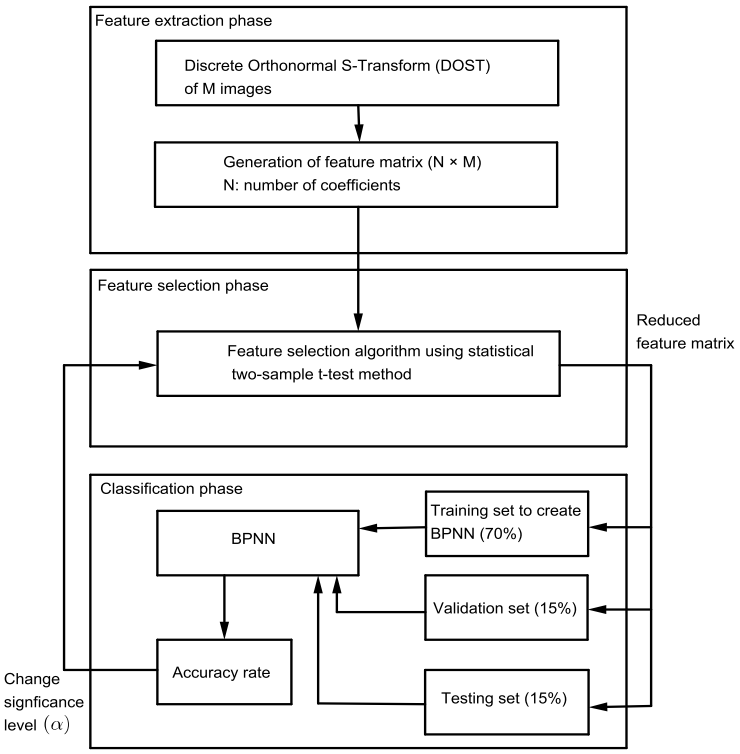


Figure 4.1: Block diagram of proposed scheme for classification of mammographic images using two dimensional discrete orthonormal S-transform (DOST) and back propagation neural network (BPNN).

4.1.1 Mammogram dataset

For the analysis of the schemes, mammographic images are taken from Mammographic Image Analysis Society (MIAS) database [25]. The database contains 322 images, which are under seven categories such as calcification, circumscribed masses, spiculated masses, architectural distortion, asymmetry, other ill-defined masses, and normal. Out of 322 images, 207 images are normal, 115 images are abnormal; and again among abnormal images the number of benign and malignant types are 64 and 51 respectively. Each image has the size of 1024×1024 pixels.

Table 4.1: Distribution of MIAS data set

Type	**Benign**	**Malignant**	**Total**
Microcalcification	12	13	25
Circumscribed masses	19	4	23
Ill-defined masses	7	7	14
Spiculated masses	11	8	19
Architectural distortion	9	10	19
Asymmetry lesion	6	9	15
Normal tissue	-	-	207
Total	64	51	322

4.1.2 Image Preprocessing

All the images of the MIAS database are composed of background, different types of noises, artifacts in the background, and pectoral muscles. All these areas are unwanted regions for feature extraction and subsequent classification. Therefore, it is necessary to extract the region of interest (ROI) which contains the lesion of mammogram. This task is accomplished by manual cropping operation. Figure. 4.2 shows some extracted ROIs containing benign and malignant types of lesions.

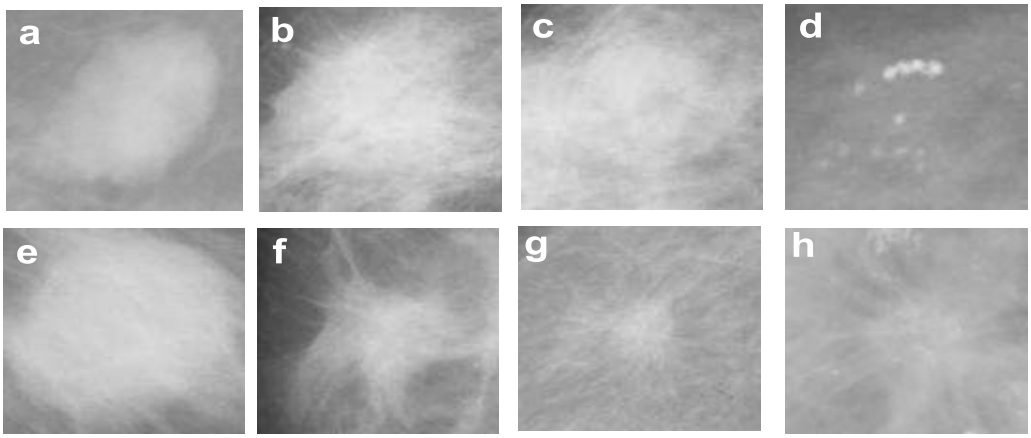
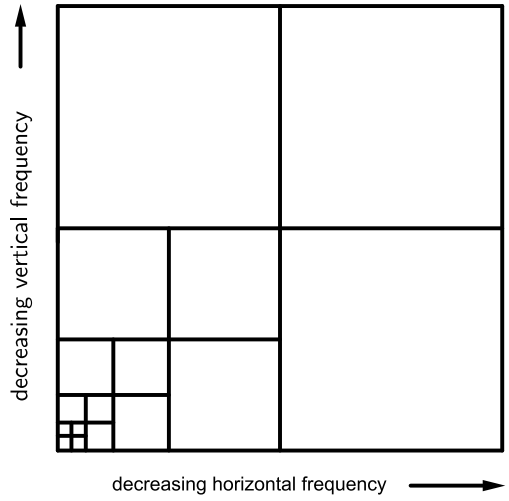


Figure 4.2: Extracted ROIs from different mammographic images (source: MIAS database). The numbered sub figures indicate the type of lesions; such as: [a, b, c, d] - benign type (mdb015, mdb081, mdb107, mdb219), [e, f, g, h] - malignant type (mdb028, mdb117, mdb115, mdb186).

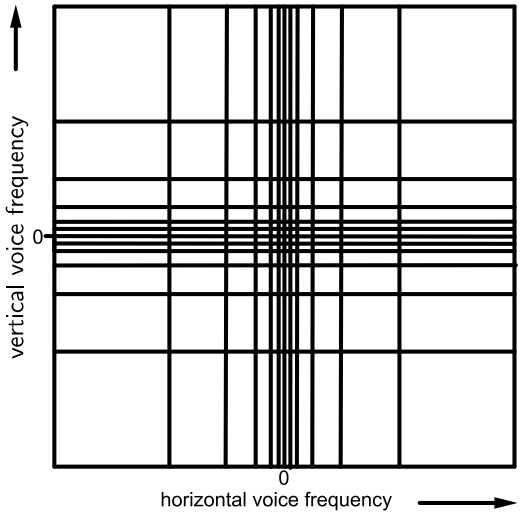
4.1.3 Feature Extraction

In this thesis, DOST has been used to extract features from mammogram images. And a comparison has been drawn out with the results obtained when we used discrete wavelet transform for feature extraction.

The DOST of an image gives the rectangular voice image of $2^{p_x-1} \times 2^{p_y-1}$ points as shown in Figure. 4.3(b). The total number of points in the voice image and original mammogram are same. In DWT, horizontal, vertical and diagonal detail coefficients of an image are obtained for each order. In DOST, voice frequencies (υ_x, υ_y) are obtained that contain a bandwidth of $2^{p_x-1} \times 2^{p_y-1}$ frequencies. In DOST, each $N \times N$ ROI gives the $N \times N$ number of coefficients in which, each coefficient is considered as a feature. A high dimensional feature matrix is constructed by using all these extracted features for all the mammographic ROIs.



(a)



(b)

Figure 4.3: Partitioning of (a) DWT and (b) DOST for six orders. The squares indicate the sub-images for each order. Both transforms use a dyadic sampling scheme but provide different information about the frequency content of the image.

4.1.4 Feature Selection

In feature selection phase, an optimal set of significant features are selected from the extracted feature matrix for the classification. In this thesis, a statistical two-sample t-test method is applied for selection of features. For two classes b (benign) and m (malignant), a two sample t-test is performed and a test decision is returned for the null hypothesis that the data in the vectors b and m come from normal distributions with equal means. The t-test determines whether the data from vectors b and m are related or not. In the proposed feature selection algorithm, a null hypothesis value, $h = 1$ indicates that the null hypothesis is incorrect and rejected. An incorrect null hypothesis implies that data from vectors b and m are significantly different and independent. In the t-test method, the t-value is computed as

$$t = \frac{|\mu_b - \mu_m|}{\sqrt{\frac{(\sigma_b)^2}{N_b} + \frac{(\sigma_m)^2}{N_m}}} \tag{4.1}$$

where, N_b and N_m are the numbers of ROIs in class b and m respectively. Here, μ_b , μ_m are means and σ_b and σ_m are standard deviations. A higher t-value indicates more significant differences between the means of the two vectors. For a certain threshold t-value, the corresponding p-value defines probability of obtaining a t-value higher than the threshold. A significance level, α defines the lower threshold for the p-value. The value of α is in the range 0 and 1. As the α value decreases, the feature reduction increases.

Algorithm 1 illustrates the selection of features.

Algorithm 1: Feature Selection

Require: $feature[1:M, 1:K]$, $target[1:K]$
M: Total number of coefficients obtained from an image
K: Total number of images in dataset
Ensure: $reduced\_feature[1:R, 1:K]$
R: Total number of reduced features
Function **ttest()** computes the null hypothesis value of two vectors at different significance levels

```
1: Create two empty vectors b and m
2: for i ← 1 to M do
3:    Clear contents of vector b and vector m
4:    for j ← 1 to K do
5:       if target[i] = 1 then
6:          Append feature[i, j] to b
7:       else
8:          Append feature[i, j] to m
9:       end if
10:   end for
11:   h[i] = ttest(b,m,α) {α is the significance level}
12:   if h[i] = 1 then
13:      Append feature[i, 1 : K] to reduced_feature
14:   end if
15: end for
```

4.1.5 Feature Classification

The back propagation neural network is used to classify the reduced feature set into different classes. The network is trained by the training set obtained by the feature selection. The neural network tool is used for this purpose. During training, the network is adjusted according to its error. During validation of the network, the validation set is used to measure network generalization. The training is halted when generalization stops improving. The testing set provides an independent measure of network performance during and after training period. For maximum classification accuracy rate and an optimal number of features, the value of α is changed. The process is repeated with the new feature set and stops when optimum classification accuracy rate is obtained with an optimized feature set. The scheme is described in Figure. 4.1.

The performance of the BPNN classifier is evaluated with the help of confusion matrix [26]. A confusion matrix is a table that shows the predicted and actual classification accomplished by the classifier. The confusion matrix for two classes (benign and malignant) and corresponding measures of performance are represented in TABLES 4.2 and 4.3 respectively. Sensitivity and specificity are two important measures for performance evaluation which calculate the percentage of true positive rate and true negative rate respectively. For an ideal performance, both specificity and sensitivity should be high. The evaluation of a classifier performance can also be accomplished by means of receiver operating characteristics (ROC) curves [6]. It is a two dimensional graph which plots sensitivity versus false positive rate (1-specificity). The area under the ROC curve referred by an index AUC is also an important factor for evaluating the classifier performance. AUC with value 1.0 is a perfect performance of the classifier.

Table 4.2: Confusion Matrix for two classes

Actual class	**Predicted class**	
	Positive	Negative
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative)

Table 4.3: Measures of classification performance

Measure	**Definition**
Sensitivity	TP/(TP+FN)
Specificity	TN/(TN+FP)
Accuracy	(TP+TN)/Total number of samples

Chapter 5

Simulation and Results

To validate the proposed feature extraction and lesion classification schemes, experiments are carried out in the MATLAB environment.

In 2D-DOST method, each transformed ROI gives 16384 number of coefficients whereas it is 19648 in case of DWT.

For all 115 abnormal ROIs, the feature matrix contains 1884160 and 2259520 coefficients for DOST and DWT methods respectively.

From the high dimensional feature matrix, most significant features are selected and given to the BPNN classifier for the classification purpose.

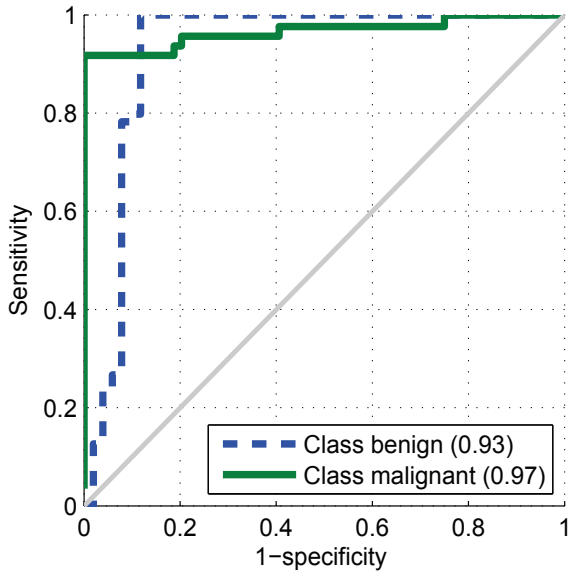
In the classifier 70% of the total set is used for the training. From the rest of the data 15% is used for testing and other 15% is used for validation.

During simulation, the feature sets are selected with different dimensions by changing the value of α. It has been observed that, the classification accuracy is maximum for $\alpha = 7 \times 10^{-4}$.

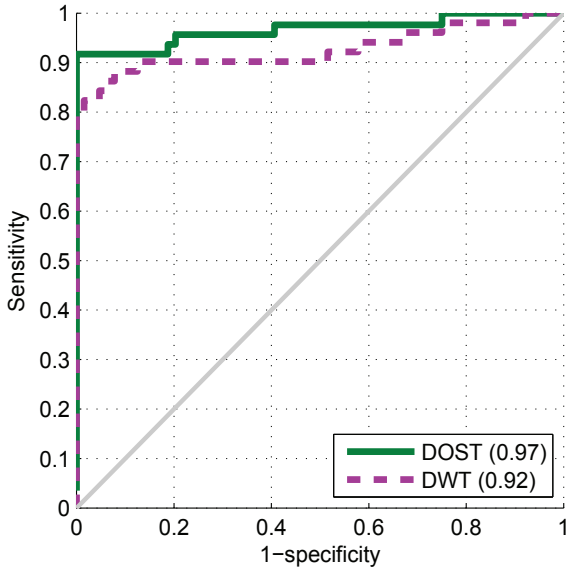
This fixes a network structure 88-10-2 to achieve the maximum classification accuracy. It is found that the maximum classification accuracy rate is 97.4% by using reduced DOST features at $\alpha = 7 \times 10^{-4}$. For the same α, the DWT method gives an

accuracy rate of 90.4.

The ROC curves for benign and malignant classes of lesion using DOST features are presented in Figure. 5.1(a). The values of AUC are 0.93 and 0.97 for the prediction of benign and malignant lesions respectively . For the prediction of malignant lesion in the mammogram, the DOST method is efficient in comparison to the DWT method. As shown in Figure. 5.1(b), the values of AUC are 0.97 and 0.92 for DOST and DWT methods respectively. Different classification performance measures computed during simulations are presented in Table 5.1.



(a)

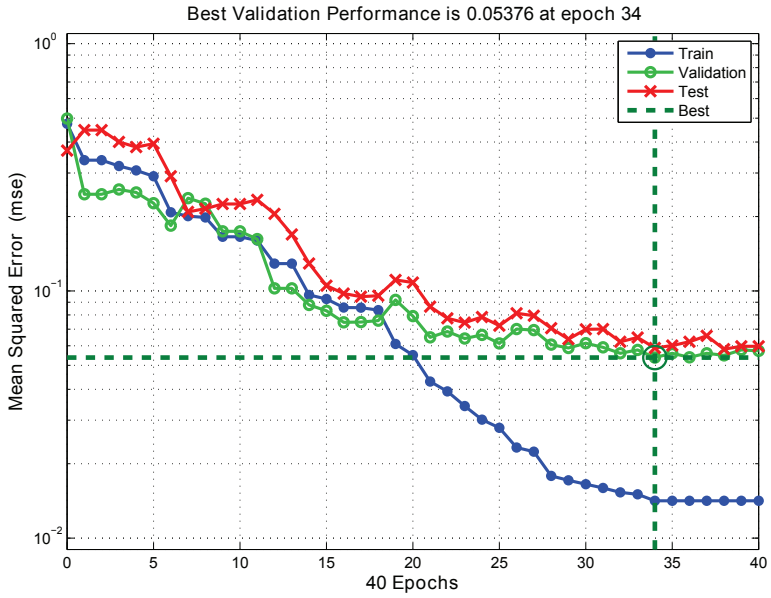


(b)

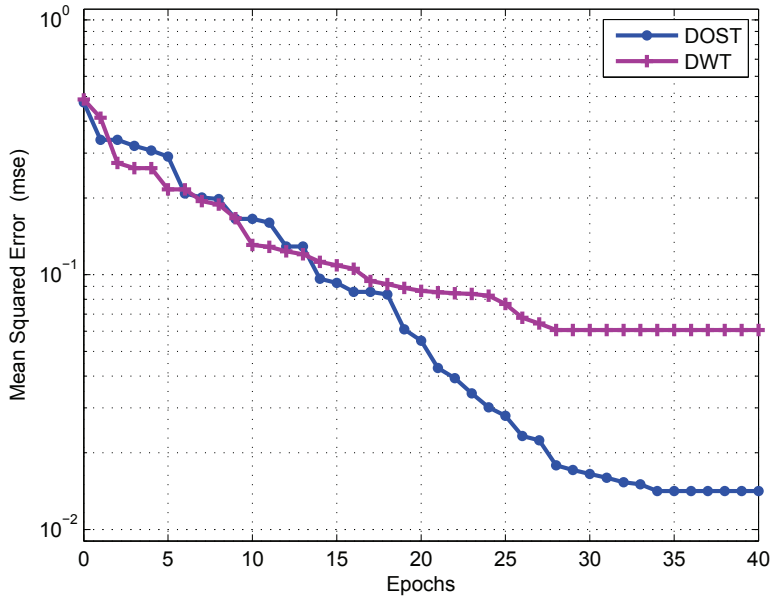
Figure 5.1: ROC curves. (a) Classification of benign and malignant mammograms by DOST method, (b) Prediction of malignant mammograms by DOST and DWT method.

The best validation performance of DOST method is 0.05376 obtained at epoch 34 with less training mean squared error (mse) at $\alpha = 7 \times 10^{-4}$ shown in Figure. 5.2(a). Mean squared error is the average squared difference between outputs and targets. Lower values are better.
Figure. 5.2(b) shows the comparison of training error of both DOST and DWT methods in terms of mse. It is found that the training error of DOST method is promisingly less than DWT method at the same $\alpha = 7 \times 10^{-4}$.

In the simulation, regression values are also studied for the efficiency of classification performance. Regression (R) value measures the correlation between outputs and targets. An R value of 1 means a close relationship. The regression value is 0.9334 at $\alpha = 7 \times 10^{-4}$ in DOST method, which determines less possibility of mis-classification. Figure. 5.3 shows the corresponding regression values of optimal performance of the classifier for both DOST and DWT methods. Decreased regression values are obtained at high dimension of feature sets with higher values of α as shown in Figure. 5.4.

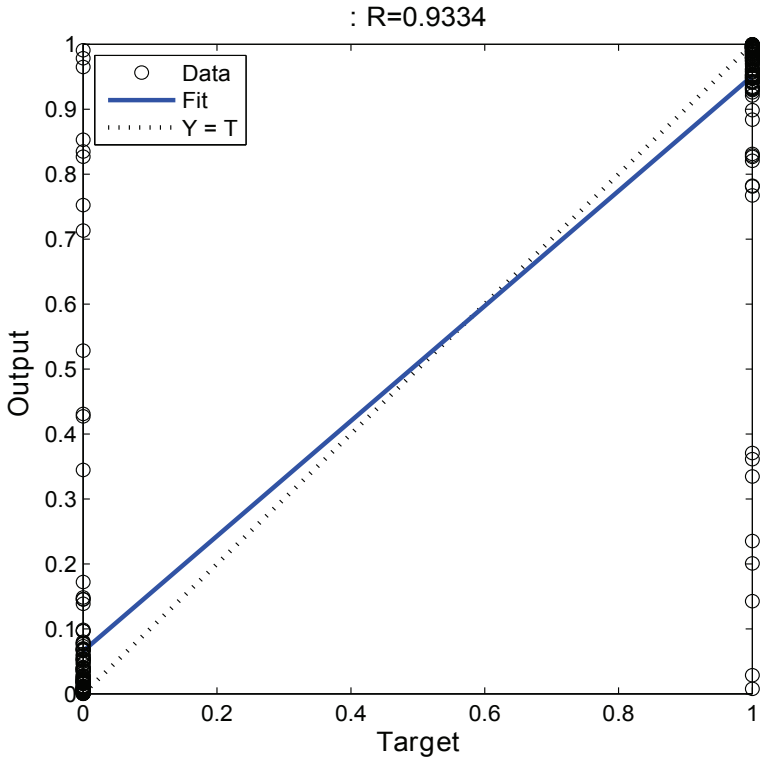


(a)

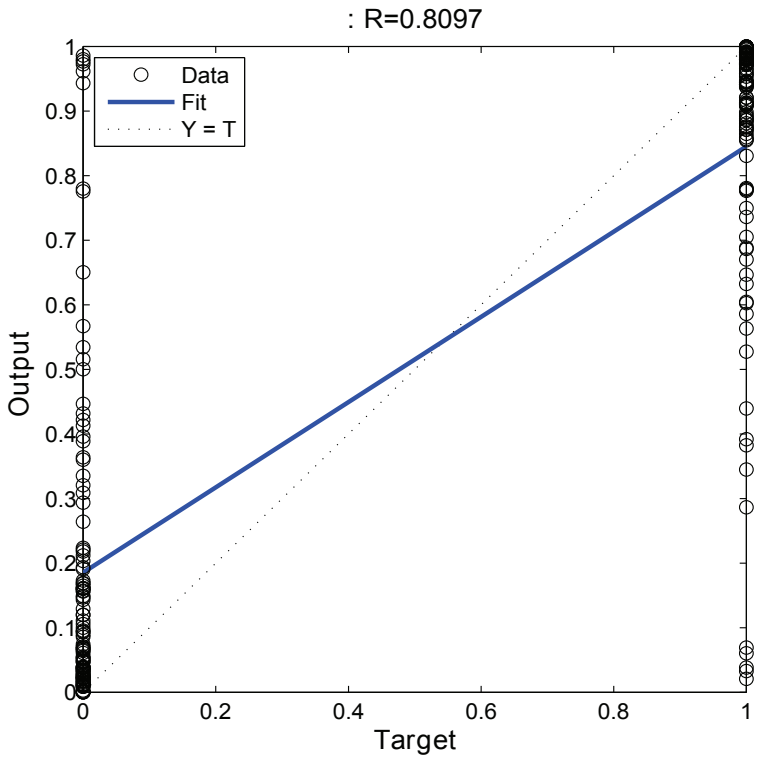


(b)

Figure 5.2: Classifier performance in terms of mean squared error(mse). (a) By DOST method, (b) comparison of training error between DOST and DWT method



(a)



(b)

Figure 5.3: Regression graph at $\alpha = 7 \times 10^{-4}$. (a) For DOST method, (b) for DWT method.

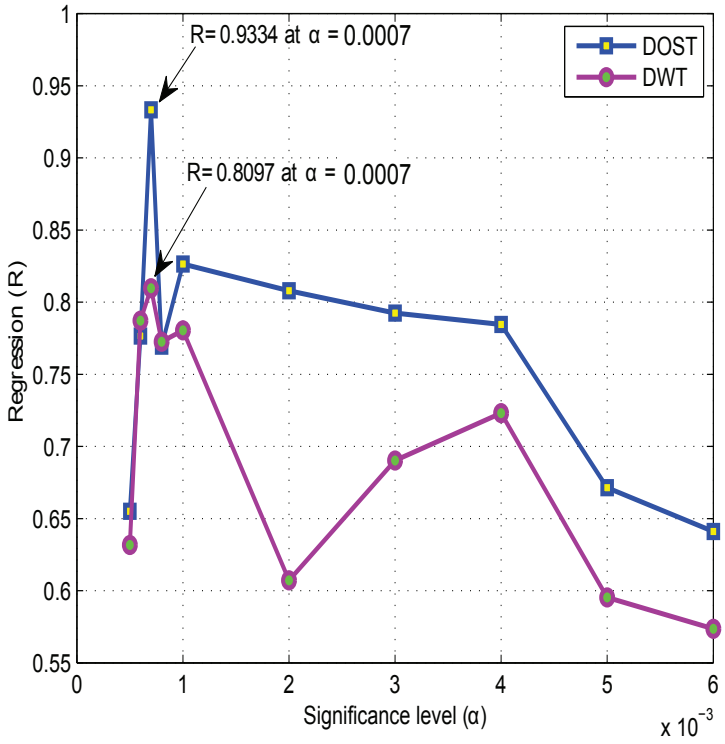
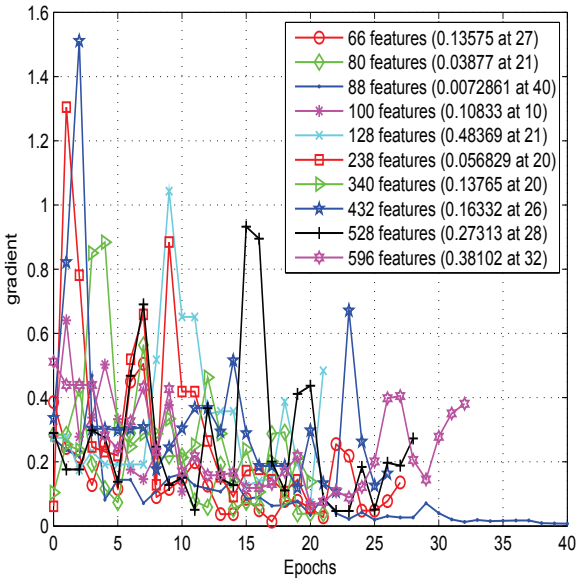


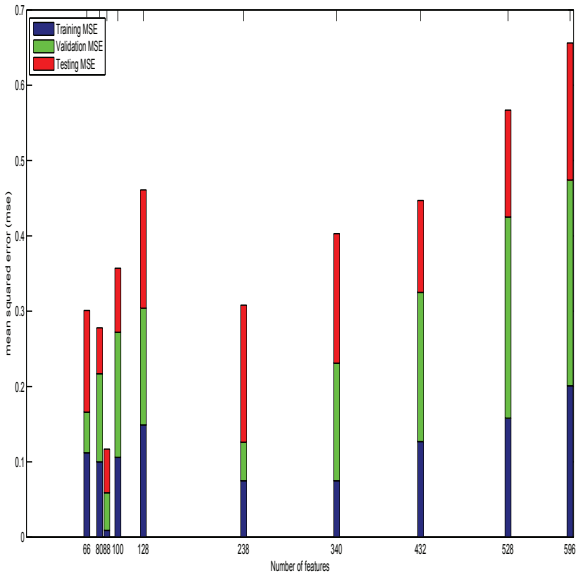
Figure 5.4: Comparison of different regression values of both DOST and DWT methods for different values of α.

Table 5.1: Performance measures for different values of α.

Significance level (α)	Performance measures (%)					
	DOST			DWT		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
6×10^{-3}	76.5	85.9	81.7	78.4	81.3	80.0
5×10^{-3}	70.6	95.3	84.3	82.4	82.8	82.6
4×10^{-3}	84.1	91.5	88.7	84.3	82.8	83.5
3×10^{-3}	86.3	92.2	89.6	70.6	92.2	82.6
2×10^{-3}	88.2	93.8	91.3	66.7	89.1	79.1
1×10^{-3}	88.4	95.2	92.2	74.5	98.4	87.8
8×10^{-4}	74.5	98.4	87.0	72.5	93.8	84.3
7×10^{-4}	94.1	100	97.4	78.4	100	90.4
6×10^{-4}	72.5	100	87.8	82.4	95.3	89.6
5×10^{-4}	78.4	87.5	83.5	60.8	96.9	80.9



(a) Gradient graph



(b) Mean sqaure error histogram

Figure 5.5: (a). Comparison of gradients of classification (b). Comparison of Mean square errors of classification, at all the reduced sets of feature used in simulation for DOST method.

Chapter 6

Conclusion and Future Work

In this thesis, a novel scheme is presented for the classification of mammographic lesions as benign or malignant to support the decision making of radiologists. The scheme utilizes DOST method for extracting the features from the mammographic images. A feature selection algorithm using the two-sample t-test method is applied for selection of significant features from the high dimensional extracted features. Finally a BPNN based classifier is used for classification. Simulation experiments are carried out on MIAS database for the validity of the scheme. The suggested scheme achieves the AUC of 0.97 from the ROC analysis and a classification accuracy rate of 97.4%. The simulation results show that the DOST features are more efficient to distinguish the benign lesions from malignant than its counterparts.

Bibliography

[1] American Cancer Society. Cancer facts & figures 2013. 2013.

[2] Priya Shetty. India faces growing breast cancer epidemic. *The Lancet*, 379(9820):992–993, 2012.

[3] Wilking N. Jonsson B. Prevention, early detection and economic burden of breast cancer. Technical report, GE Healthcare, 2013.

[4] Breast cancer. Technical report, The Indian Council Of Medical Research, 2011.

[5] L Tabar and PB Dean. Mammography and breast cancer: the new era. *International Journal of Gynecology & Obstetrics*, 82(3):319–326, 2003.

[6] HD Cheng, XJ Shi, Rui Min, LM Hu, XP Cai, and HN Du. Approaches for automated detection and classification of masses in mammograms. *Pattern recognition*, 39(4):646–668, 2006.

[7] Sheng Liu, Charles F Babbs, and Edward J Delp. Multiresolution detection of spiculated lesions in digital mammograms. *IEEE Transactions on Image Processing*, 10(6):874–884, 2001.

[8] Roberto R Pereira Jr, Paulo M Azevedo Marques, Marcelo O Honda, Sergio K Kinoshita, Roger Engelmann, Chisako Muramatsu, and Kunio Doi. Usefulness of texture analysis for computerized classification of breast lesions on mammograms. *Journal of Digital Imaging*, 20(3):248–255, 2007.

[9] M Fraschini. Mammographic masses classification: novel and simple signal analysis method. *Electronics letters*, 47(1):14–15, 2011.

[10] M. Talha, G. Bin Sulong, N. Naveed, and M. A. Jaffar. Malignancy and abnormality detection of mammograms using discrete wavelet transformed features and neural network. *Information*, 15(2):707–719, 2012.

[11] BN Prathibha and V Sadasivam. Hybrid transforms domain–based mammogram analysis using c–svm classifier. *International Journal of Medical Engineering and Informatics*, 4(2):146–156, 2012.

[12] Ahmet Sertbas, Osman N Ucan, et al. Mammographical mass detection and classification using local seed region growing-spherical wavelet transform (lsrg-swt) hybrid scheme. *Computers in biology and medicine*, 2013.

[13] S Mohan Kumar and G Balakrishnan. Multi resolution analysis for mass classification in digital mammogram using stochastic neighbor embedding. In *International Conference on Communications and Signal Processing (ICCSP)*, pages 101–105. IEEE, 2013.

[14] Karthikeyan Ganesan, U Rajendra Acharya, Chua Kuang Chua, Choo Min Lim, and K Thomas Abraham. One-class classification of mammograms using trace transform functionals. 2014.

[15] Chitre Y Kaiser-Bonasso C. Dhawan, A.P. Analysis of mammographic microcalcifications using grey-level image structure features. *Transactions on Medical Imaging*, 15(3):246–259, 1996.

[16] Heang-Ping Chan, Berkman Sahiner, Kwok Leung Lam, Nicholas Petrick, Mark A Helvie, Mitchell M Goodsitt, and Dorit D Adler. Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces. *Medical Physics*, 25(10):2007–2019, 1998.

[17] Daniel Manrique, Juan Ríos, and Alfonso Rodríguez-Patón. Evolutionary system for automatically constructing and adapting radial basis function networks. *Neurocomputing*, 69(16):2268–2283, 2006.

[18] M Abdalla Al Mutaz, Safaai Deris, Nazar Zaki, and Doaa M Ghoneim. Breast cancer detection based on statistical textural features classification. In *Innovations in Information Technology, 2007. IIT'07. 4th International Conference on*, pages 728–730. IEEE, 2007.

[19] Essam A Rashed, Ismail A Ismail, and Sherif I Zaki. Multiresolution mammogram analysis in multilevel decomposition. *Pattern Recognition Letters*, 28(2):286–292, 2007.

[20] Aijuan Dong and Baoying Wang. Feature selection and analysis on mammogram classification. In *Communications, Computers and Signal Processing, 2009. PacRim 2009. IEEE Pacific Rim Conference on*, pages 731–735. IEEE, 2009.

[21] Ioan Buciu and Alexandru Gacsadi. Directional features for automatic tumor classification of mammogram images. *Biomedical Signal Processing and Control*, 6(4):370–378, 2011.

[22] Al Mutaz M Abdalla, Safaai Dress, and Nazar Zaki. Detection of masses in digital mammogram using second order statistics and artificial neural network. *International Journal of Computer Science & Information Technology*, 3(3), 2011.

[23] Rodrigo Pereira Ramos, Marcelo Zanchetta do Nascimento, and Danilo Cesar Pereira. Texture extraction: An evaluation of ridgelet, wavelet and co-occurrence based methods applied to mammograms. *Expert Systems with Applications*, 39(12):11036–11047, 2012.

[24] R. G. Stockwell, L. Mansinha, and R. P. Lowe. Localization of he complex spectrum : The s transform. *IEEE Transactions on Signal Processing*, 44:998–1001, 1996.

[25] John Suckling, J Parker, DR Dance, S Astley, I Hutt, C Boggis, I Ricketts, E Stamatakis, N Cerneaz, Siew-Li Kok, et al. The mammographic image analysis society digital mammogram database. 1994.

[26] Ron Kohavi and Foster Provost. Glossary of terms. *Machine Learning*, 30(2-3):271–274, 1998.