Automatic Mammogram Analysis Using Wavelet-Fourier Transforms and Entropy-based Feature Selection

by

© Liuhua Zhang

A Thesis submitted to the

School of Graduate Studies

in partial fulfillment of the requirements for the degree of

Master of Computer Science

Department of Computer Science

Memorial University of Newfoundland

October, 2014

St. John's Newfoundland

ABSTRACT

Breast cancer is the second leading cause of cancer-related death after lung cancer in women. Early detection of breast cancer in X-ray mammography is believed to have effectively reduced the mortality rate since 1989. However, a relatively high false positive rate and a low specificity in mammography technology still exist. A computer-aided automatic mammogram analysis system in this research is proposed to improve the detection performances.

In designing this analysis system, the discrete wavelet transforms (Daubechies 2, Daubechies 4, and Biorthogonal 6.8) and the Fourier cosine transform were first used to parse the mammogram images and extract statistical features. Then, an entropy-based feature selection method was implemented to reduce the number of features. Finally, different pattern recognition methods (including the Back-propagation Network, the Linear Discriminate Analysis, and the Naïve Bayes Classifier) and a voting classification scheme were employed. The performance of each classification strategy was evaluated for sensitivity, specificity, and accuracy and for general performance using the Receiver Operating Curve. The experiment demonstrated that the proposed automatic mammogram analysis system could effectively improve the classification performances, especially using the voting classification scheme based on the selected optimal features.

ACKNOWLEDGEMENTS

My deepest gratitude goes first and foremost to Professor Adrian Fiech, my supervisor, for his constant encouragement and guidance. He introduced me to this study, without his consistent and illuminating instruction, this thesis could not have reached its present form. Second, I would like to express my heartfelt gratitude to Professor Edward Kendall, my co-supervisor. I am indebted to his many hours to read and re-read various drafts of thesis and his helpful comments and suggestions for this study. Without his enlightening instruction, impressive kindness and patience, I could not have completed my thesis. His keen and vigorous academic observation enlightens me not only in this thesis but also in my future study.

Last my thanks would go to my beloved family for their loving considerations and great confidence in me all through these years. I also owe my sincere gratitude to my friends and my fellow classmates who gave me their help and time in listening to me and helping me work out my problems during the difficult course of the thesis.

Table of Contents

List of Tables

List of Figures

List of Symbols, Nomenclature or Abbreviations

CAD – Computer Aided Detection

CADx – Computer Aided Diagnosis

ROI – Region of Interest

DDSM – Digital Database for Screening Mammography

MIAS – Mammographic Images Analysis Society

FFDM – Full-field Digital Mammography

CC – Mamogram Craniocaudal View

MLO – Mamogram Mediolateral Oblique

MRI – Magnetic Resonance Imaging

Db – Daubechies

Bior – Biorthogonal

DWT – Discrete Wavelet Transform

CWT – Continuous Wavelet Transform

FFT – Fast Fourier Transform

WSQ – Wavelet Scalar Quantization

SVM – Support Vector Machine

GOA – Gradient Orientation Analysis

ALOE – Analysis of Local Orientated Edges

SFS – Sequential Forward Selection

SBS – Backward Forward Selection

SFFS – Sequential Forward Floating Selection

PSNR – Peak Signal to Noise Ratio

ROC – Receiver Operator Curve

AUC – Area Under the ROC Curve

LDA – Linear Discriminate Analysis

BP – Back-propagation

NB – Na ïve Bayes

DICOM – Digital Imaging and Communications in Medicine

PGM – Portable graymap

IG – Information Gain

FN – False Negative

FP – False Positive

TN – True Negative

TP – True Positive

HERA – Health Research Ethics Authority

Chapter 1 – Introduction

1.1 Research Rationale

Breast cancer is the most commonly diagnosed form of cancer in women and the second-leading cause of cancer-related death after lung cancer [1]. Statistics from the American Cancer Society indicate that approximately 232,670 (29% of all cancer cases) American women will be diagnosed with breast cancer, and an estimated 40,000 (15% of all cancer cases) women will die of it in 2014 [2]. In other words, 637 American women will be diagnosed with breast cancer, and 109 women will die of it every day. Similar statistics were also found in Canada, where approximately 23,800 (26%) women were diagnosed with breast cancer, and 5,000 (14%) died from it in 2013 [3]. Under this circumstance, detection and diagnosis of breast cancer has already drawn a great deal of attention from the medical world.

Studies show that early detection, diagnosis and therapy is particularly important to prolong lives and treat cancers [4]. If breast cancer is found early, the five-year survival rate of patients in stage 1 could reach 90% with effective treatment. To date, medical imaging technology, which is convenient and noninvasive, is one of the main methods for breast cancer detection. Commonly used medical imaging technologies include X-ray mammography, Computer Tomography (CT), ultrasound and Magnetic

Resonance Imaging (MRI), Positron Emission Tomography (PET), and Single-Photon Emission Computed Tomography (SPECT). Among these technologies, mammography achieves the best results in early detection of asymptomatic breast cancer and is one of the least expensive ones. For this reason, it has become the principal method of breast cancer detection in clinical practice, and one of the most effective ways for general breast cancer survey, though its detection sensitivity is still low. North American countries, the United States and Canada, consider breast cancer general survey and diagnosis as one of the most important parts of their health care systems. As a result, high resolution breast imaging equipment has become widely available [4].

Modern equipment has improved the technical aspects of mammography, but a relatively high false positive rate and a low specificity still exist. This is due to fundamental physical limitations such as unobvious lesions, as well as controllable factors like radiologists' inexperience in reading mammograms. This latter issue has been addressed using double reading, where two radiologists make their own judgments independently based on the same mammogram, and then combine and discuss both opinions. However, this is expensive, and as a result, interest in Computer-Aided Diagnosis (CADx) solutions has emerged [4].

1.2 Background Information

1.2.1 Detection of Masses and Calcifications

Masses are the most common and basic symptoms of breast cancer. In clinically detected breast cancer, 80% - 90% of cases had masses [5]. Having spiculate boundaries is the most important characteristic in identifying malignant breast cancers. Additionally, shapes, sizes, and texture features also affect the diagnosis of breast cancer. Masses in mammography can be recognized as a local, high-contrast area, but the value of contrast is not unique. It changes when imaging conditions, sizes and backgrounds change. The X-ray absorption rates of masses are very close to dense glandular tissue in breast and other dense tissues. In addition, the boundaries of masses are always mixed with background structures, and mass detection has become a difficult task for observers and computer programmers [6]. In breast masses, high density usually reflects malignant tumors, which have irregular spiculate boundaries. In contrast, most benign masses have clear boundaries that are often round or oval [7].

Calcifications (including marcocalcifications and microcalcifications) are important features in breast cancer detection. Tiny glandular clusters of microcalcifications often appear in early stages of breast cancer. Statistics show that 30%-50% of most malignant breast tumors have the symptom of microcalcification [8].

Calcifications in breast cancer mostly refer to calcium phosphate. A few are calcium oxalate calcifications. Calcifications coincide in lumens where ductal

carcinoma causes cellular degeneration. They manifest as piles of sediment or spiculate shapes in mammogram images. Calcifications are also present in ducts and stroma. Calcifications form when necrotic cells release phosphate radical into a calcium rich environment [9].

Automatic calcification detection has been an important research target. Some success has been achieved. However, applying the findings has been challenging for the following reasons: 1) microcalcifications occur in various sizes, shapes and distributions; 2) microcalcifications have low contrast in region of interest (ROI); 3) dense tissue and/or skin thickness make suspicious lesion areas difficult to detect (especially in young women); 4) the dense tissue is easily misunderstood as microcalcification, which results in high false positive rates among most existing algorithms. Therefore, microcalcification detection remains one of the most popular topics in medical image processing research [10].

1.2.2 Mammography

Mammography is a specific kind of imaging technology that uses a low-dose X-ray system to examine breasts [11]. The use of radiography for cancer diagnosis appeared in the late 1920s, but X-ray mammography was developed in the 1960s [12]. Since many pathological conditions, such as breast cancer, are difficult to identify because of the imperceptible physical changes, mammography is aimed at maximizing the visibility of pathology. Two recent advances in mammography include digital

mammography and computer-aided detection (CAD). Digital mammography, also called full-field digital mammography (FFDM), is a mammography system that uses solid-state detectors [13]. Digital mammography provides slightly better detection rates than the older screen-film technology. It could also reduce processing steps and so increase treatment efficiency.

1.2.2.1 Mammography Technology

The principal components of a mammography system consist of the X-ray tube (generates the x-rays), filter (removes unwanted radiation), compression paddle (helps to regularize breast geometry), grid (rejects scattered radiation), and detector. The components are shown in Fig. 1.1. Unlike regular X-ray tubes, mammography equipment uses molybdenum anodes or rhodium anodes. Rhodium provides a more penetrating X-ray, useful for large or dense breasts. The system is designed to maximize spatial and contrast resolution. Modern units use a full field digital matrix detector [11].

Generally, a mammogram image could have two basic views: craniocaudal (CC) view which is taken from above a horizontally-compressed breast and mediolateral-oblique (MLO) view which is taken from the side and at an angle of a diagonally-compressed breast. These views are shown in Fig. 1.2 A and B, respectively.

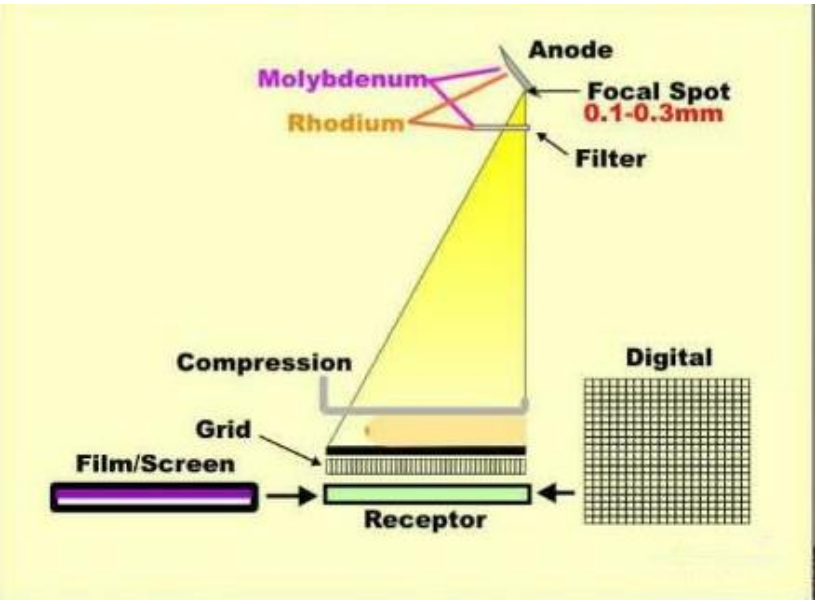


Figure 1.1: The physical structure of the equipment for mammography (retrieved from http://www.sprawls.org/resources/MAMMO/module.htm, Aug., 2012)

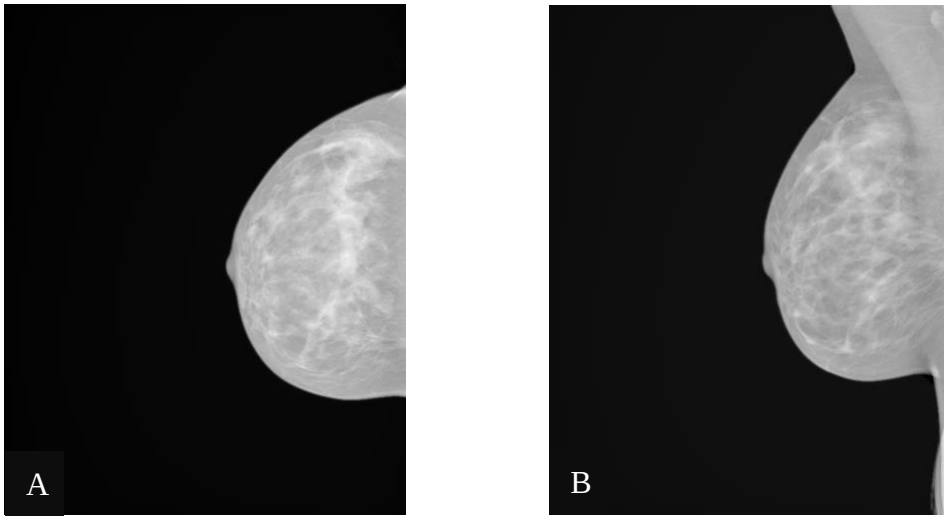


Figure 1.2: Digital mammograms illustrating the conventional views of the breast. A. Craniocaudal view (CC) the compressed breast is viewed from above; B. Mediolateral oblique (MLO) the compressed breast is viewed laterally towards the midline.

1.2.2.2 CAD Technology

In 1967, Dr. Fred Weisberg and others published an article in Radiology stating that breast cancer could be examined by comparing the asymmetry of the medical images of left and right breasts [14]. It was the first time that computer-aided diagnosis was applied to X-ray images. After nearly 40 years of development, CAD has become a piece of technology which has been gradually accepted. Computer-aided detection (CAD) systems combine computer calculation and analysis. They utilize medical imaging processing technology and other possible physiological and biochemical methods. The purpose of CAD software is to assist doctors in detecting disease and improving their diagnostic accuracy. Specifically, a mammogram is passed to the CAD system. The CAD system then searches for any abnormal areas such as density and calcification that may indicate the pathology of breast cancer. These suspicious areas on the images will be marked out by the CAD system, which could be a sign for the radiologist in further analysis.

1.2.3 Terminology of Diagnosis Rates

The performance of a mammography screening system can be measured by two parameters: sensitivity and specificity. Sensitivity (true positive rate) is the proportion of the cases deemed abnormal when breast cancer is present. For example, if 100 women do have breast cancer among 1000 screened patients but only 90 are detected, then the sensitivity is 90/100 or 90%. Sensitivity may depend on several factors, such

as lesion size, breast tissue density, and overall image quality. In cancer screening protocols, sensitivity is deemed more important than specificity, because failure to diagnose breast cancer may result in serious health consequences for a patient. Almost fifty percent of cases in medical malpractice relate to “false-negative mammograms” [15].

Specificity (true negative fraction) is the proportion of cases deemed normal when breast cancer is absent. For example, if 100 cases of breast cancer are diagnosed in a set of 1000 patients, and the screening system finds 720 cases to be normal, the specificity is 720/900 or 80%. Although the consequences of a false positive (diagnosing a normal patient as having breast cancer) are less severe than missing a positive diagnosis of cancer, specificity should also be as high as possible. False positive examinations can result in unnecessary follow-up examinations and procedures, and may lead to significant anxiety and concern for the patient.

1.3 Research Objectives

The primary objective of this research is to design an automatic mammogram analysis system that combines features from the wavelet transform and the Fourier transform to select optimal features, and evaluates performances of different classifiers based on these features. Specific research objectives are

1. Develop a set of pre-processing steps to isolate the tissue in mammogram images and regularize the appearance of the images to make direct comparisons possible.

2. Apply the wavelet transform and the Fourier transform to parse an image and generate a set of scalar features based on the output of the transforms to characterize each image.

3. Employ an entropy-based feature selection method to reduce the number of features extracted from the previous step.

4. Classify mammogram images as normal or cancerous based on three classifiers, and calculate the sensitivity, specificity, and accuracy.

5. Evaluate the performances of the classifiers based on the Receiver Operating Curve, and compare them with a proposed voting classification scheme.

1.4 Scope of Thesis

In this thesis, the scope of the study focused on breast cancer detection using a computer-aided automatic mammogram analysis system. In designing this analysis system, an entropy-based feature selection method was implemented and different pattern recognition methods, including the Back-propagation (BP) Network, the Linear Discriminant Analysis (LDA), and the Naïve Bayes (NB) Classifier, were employed.

In Chapter 2, different data transform methods for mammography including the Discrete Wavelet Transform (DWT) and Discrete Fourier Transform (DFT) are first introduced in their principles, formulations, limitations, and applications. Then, pattern recognition in existing literature is reviewed, and its applications in breast cancer detection are particularly introduced.

In Chapter 3, the mammogram image processing stage, which is the first stage of the proposed mammogram analysis system, is presented. This stage includes two basic steps: mammogram image pre-processing and data transforms.

In Chapter 4, the feature selection and image classification stage in the mammogram analysis system is presented. An entropy-based feature selection algorithm is proposed to reduce the number of features extracted from the transformed mammogram images. Then, three classifiers and a voting classification scheme are used to discriminate normal or cancerous mammograms. Finally, the Receiver Operator Curve (ROC) is analyzed to evaluate the performances of classifiers.

In Chapter 5, the performances of the proposed mammogram analysis system, including sensitivity, specificity, and accuracy, are evaluated and discussed.

The overall conclusions from the research in this thesis are presented in Chapter 6, in which some suggestions for future work are also outlined.

Chapter 2 – Data Transforms and Pattern Recognition

Data transforms and pattern recognition are two essential parts in designing the automatic mammogram analysis system. In this chapter, the Fourier transform and different discrete wavelet transforms are first introduced with their principles, properties, limitations, and applications. Then, the concept of pattern recognition and its general system are presented. Specifically, its applications in breast cancer detection are reviewed.

2.1 Introduction of Data Transforms

Many data processing algorithms, such as compression, filtering, image processing, involve data transformation. Basically, data can be represented by “basis”. In linear algebra, basis refers to a series of linearly independent vectors that define a space. Any data in one space can be represented by a linear combination of these vectors. For example, the essence of Fourier expansion is to express a signal with linear combinations of bases in one space. The nature of the wavelet transform is also related to the transform based on wavelet bases.

Selecting a certain kind of basis or transform is an essential task for different data processing algorithms. For example, in data compression, this basis should be selected

to represent the signal to the greatest extent using fewer vectors. The goodness of fit will determine how much compression can be achieved with acceptable information loss [29].

In the data time-frequency analysis, the Fourier transform is traditionally applied, which is a global transform between the time and frequency domain. Therefore, the Fourier transform cannot express the local properties of signals in the time and frequency domains simultaneously. However, these local properties are the key characteristics of non-stationary signals in some circumstances. In order to analyze and process non-stationary signals, various approaches have been proposed, including Gabor transform, short-time Fourier transform, fractional Fourier transform, line frequency modulation wavelet transform, wavelet transforms, circulation statistics theory and amplitude-frequency modulation signal analysis [30]. The goal is to retain important temporal information in the frequency domain, or partial frequency information in the time domain.

The basic idea of the short-time Fourier transform is that, assuming a non-stationary signal is stationary (or pseudo stationary), represented as a power spectrum in a short interval of a window function $g(x)$, move the window function and make $f(t)g(t-\tau)$ stationary in different limited time width, and then calculate the power spectrum at that different instance [30]. Essentially, the short-time Fourier transform provides only time-resolved and single resolution in signal analysis.

The wavelet transform, as a time-dimensional analysis method, has not only the characteristics of multi-resolution analysis, but also the ability to express local properties in both the time and frequency domains [31]. This transform has a fixed but changeable window size. Consequently, the wavelet transform demonstrates good time resolution and poorer frequency resolution in its high frequency components, and good frequency resolution and poorer time resolution in its low frequency components. It is especially suitable for the detection of the transient abnormal phenomenon in normal signals by showing its composition. Specifically, the continuous wavelet transform, hailed as the "microscope of signal analysis", has a fairly good performance in the fault detection and diagnosis of dynamic systems [32].

2.2 Fourier Transform

The Fourier transform is one of the most important methods in the field of signal processing. It provides a bridge between the frequency domain (Eqn. 2.1) and the time domain (Eqn. 2.2).

$$F(v) = \int e^{-2\pi i(x,v)} f(x) dx \tag{2.1}$$

$$f(x) = \int e^{2\pi i(x,v)} F(v) dv \tag{2.2}$$

The Fourier pair illustrates that data presented in one domain can be represented in the other domain through inverse transformation.

The frequency of an image is the degree of the image's gray level change in the plane space. For example, in an image, the corresponding frequency value of the area with slow gray level change is very low, and vice versa.

2.2.1 Discrete Fourier Transform (DFT)

An infinite number of different frequency sine and cosine curves are required to represent aperiodic signals, which is impossible to implement in the real world. As a result, the Discrete Fourier Transform (DFT) is used for discrete data with limited length in computer programming.

The process of DFT can be illustrated in Fig. 2.1.

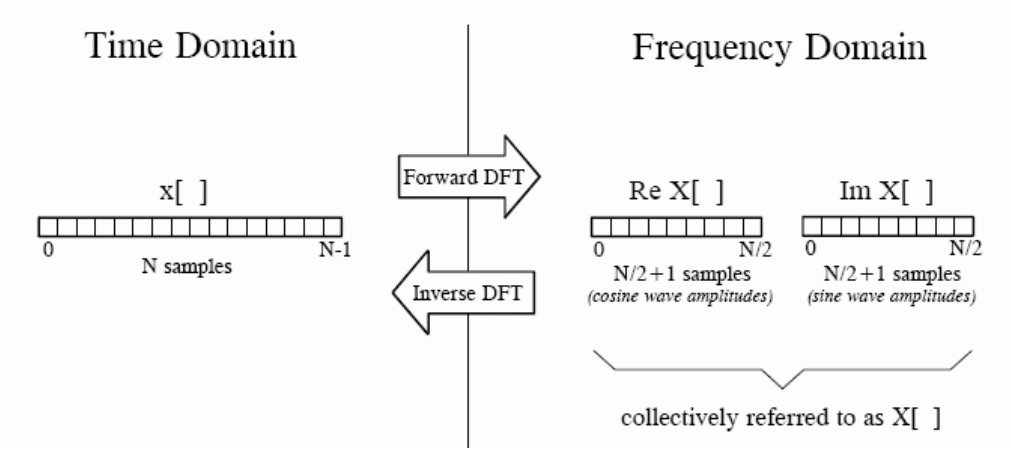


Figure 2.1: Terminology of DFT (retrieved from http://www.dspguide.com/ch8/2.htm, Oct., 2012)

The input signal $x[]$ in the time domain consists of N points, it then produces two signals in the frequency domain: the real part "$Re\,X[]$" and the imaginary part "$Im\,X[\,]$". The values in $Re\,X[]$ and $Im\,X[\,]$ are respectively the amplitudes of cosine and sine wave sets [34].

The sine and cosine wave sets with unity amplitude are called DFT basis functions [34], given by

$$C_k[i] = cos(^{2\pi ki}/_{N}) \tag{2.3}$$

$$S_k[i] = sin(^{2\pi ki}/_{N}) \tag{2.4}$$

where $C_k[]$ is the cosine wave for the amplitude held in $Re\ X[k]$, and $S_k[]$ is the sine wave for the amplitude held in $Im\ X[k]$.

Thus, the original signal $x[]$ can be synthesized as

$$x[i] = \sum_{k=0}^{N/2} Re\bar{X}[k]\cos(^{2k\pi i}/_{N}) + \sum_{k=0}^{N/2} Im\bar{X}[k]\sin(^{2k\pi i}/_{N}) \tag{2.5}$$

In general, the DFT of a discrete signal $g(n)$ is defined as

$$G(K) = \sum_{n=0}^{N-1} g(n)e^{-i\frac{2\pi knT}{N}}, \qquad k = 0, \dots, N-1 \tag{2.6}$$

In a similar way, 2-D FT is a rather straightforward extension of the 1-D transform. Its equation is as follows:

$$G(u,v) = \sum_{x=0}^{N-1}\sum_{y=0}^{N-1} g(x,y)e^{-i\frac{2\pi(ux+vy)}{N}} \tag{2.7}$$

2.2.2 Properties of DFT

There are several properties in the DFT that make it easy to change a signal from one domain to the other domain.

a) Linearity

The Fourier transform is linear, and this property applies to all four members of the Fourier transform family (Fourier transform, Fourier series, discrete Fourier

transform, and discrete time Fourier transform). This means it possesses the properties of homogeneity and additivity [34].

Homogeneity means that a change in amplitude in one domain produces a corresponding change in amplitude in the other domain. For example, in mathematical form (for any constant m), if $x[\,]$ and $X[\,]$ are a Fourier transform pair, then $mx[\,]$ and $mX[\,]$ are also a Fourier transform pair. Additivity means that addition in one domain is equivalent to addition in the other domain.

b) Periodicity and Conjugate Symmetry

The DFT and IDFT are periodic with period *N*. A simple proof is as follows:

$$F(u,v+N) = \frac{1}{N}\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y)e^{-i\frac{2\pi ux}{N}}e^{-i\frac{2\pi(v+N)y}{N}}$$

$$= \frac{1}{N}\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y)e^{-i\frac{2\pi ux}{N}}e^{-i\frac{2\pi vy}{N}} = F(u,v) \qquad (2.8)$$

So, $F(u,v) = F(u+N,v) = F(u,v+N) = F(u+N,v+N)$ (2.9)

c) Convolution

If $x(n)$ has the Fourier transform $X(k)$, and $Y\,(k)$ is the FT of $y(n)$, then

$$X(k)Y\,(k) \;=\; DFT\,\{\{x(n)\} \circledast \{y(n)\}\} \qquad (2.10)$$

Here, $\circledast$ denotes circular convolution [30].

d) Symmetry

If $F(u,v)$ is real, then

$$F(u,v) = F^*(-u,-v) \rightarrow |F(u,v)| = |F(-u,-v)| \qquad (2.11)$$

$$f(x,y)\ real\ and\ even \leftrightarrow F(u,v)\ real\ and\ even \qquad (2.12)$$

$$f(x,y)\ real\ and\ odd \leftrightarrow F(u,v)\ imaginary\ and\ odd \quad (2.13)$$

2.2.3 Applications

Fourier analysis is a useful tool for extracting data from many time domain signals or determining the resolution level in spatial domain images. As mentioned above, frequency encoded data can be transformed to the spatial domain. The best known example of this is MRI data, which is collected in a frequency encoded time domain and then transformed to the frequency encoded spatial domain to provide the MRI image. However, as shown above, the Fourier transform has a serious disadvantage: temporal information loss in time-frequency transformation [33]. Consequently, the Fourier transform may not be suitable for analyzing signals containing unstable or transient characteristics.

Although the Fourier transform can associate the features of a signal's frequency domain with its time domain, and observe respectively from the frequency and time domains, it cannot provide information simultaneously on both. This is because the time domain waveform of a signal is a composite of the frequency domain information [30]. In other words, analyzing a Fourier spectrum provides no information on when a certain frequency is produced. Thus, there is a dichotomy in the information available from Fourier-based signal analysis (namely, the exclusivity of the frequency and time domains).

In practical signal processing, especially for non-stable signals, the frequency domain characteristics of a signal are important [31]. For example, the vibration signal from the cylinder cover of a diesel engine, produced by strike or shock, is a transient signal. This signal is hard to be shown only in either the frequency domain or time domain. Therefore, a new way is required to describe joint time-frequency characteristics of a signal by combining the frequency domain with the time domain. This so-called time-frequency analysis method is also known as the time-frequency localization method.

2.3 Discrete Wavelet Transform

In practical applications, as for the Fourier transform, discretization must be applied to continuous wavelet. The discretization of continuous wavelet $\Psi_{a,b}(t)$ and continuous wavelet transform $W_f(a,b)$ is based on the scaling parameter a and translation parameter b. A continuous wavelet can be defined as [38]

$$\Psi_{a,b}(t) = |a|^{-1/2}\Psi\left(\frac{t-b}{a}\right) \tag{2.14}$$

where $b \in R, a \in R^+, a \neq 0$, ($a$ is a positive value in discretization, R is for the field of real number). Its compatibility condition is

$$C_\Psi = \int_0^\infty \frac{|\Psi(\bar{\omega})|}{|\bar{\omega}|} d\bar{\omega} < \infty \tag{2.15}$$

Assuming $a = a_0^j, b = ka_0^j b_0,\ j \in Z,\ a_0 > 1$, (where Z represents integers) the corresponding discrete wavelet function $\Psi_{j,k}(t)$ can be given by

$$\Psi_{j,k}(t) = a_0^{-\frac{j}{2}}\Psi\left(\frac{t-ka_0^j b_0}{a_0^j}\right) = a_0^{-\frac{j}{2}}\Psi(a_0^{-j}t - kb_0) \quad (2.16)$$

The coefficient of the discrete wavelet can be presented as

$$C_{j,k}\int_{-\infty}^{\infty} f(t)\Psi_{j,k}(t)dt =< f \quad (2.17)$$

Its reconstruction equation is

$$f(t) = C\sum_{-\infty}^{\infty}\sum_{-\infty}^{\infty} C_{j,k}\Psi_{j,k}(t) \quad (2.18)$$

In this equation, *C* is a constant which has nothing to do with the signal. However, the choice of a_0 and b_0 is important because of the requirement of the precision of the reconstructed signals. Based on that, a_0 and b_0 should be as small as possible, since the further the grid points are from each other, the lower is the reconstruction accuracy that can be achieved [39].

In practical calculation, it is impossible to calculate a, b values of the continuous wavelet transform (CWT) for all scaling parameters and translation parameters, and the actual observation signals are discrete. As a result, the discrete wavelet transform (DWT) is usually used. When the DWT is applied, the ½ in the coefficient effectively reduces the resolution of the scale map. The most effective method of computation is the fast wavelet algorithm (also named as pyramid algorithm), which was developed by S. Mallat in 1988 [40]. For any signal, the first step of the discrete wavelet transform is to divide a signal into the low frequency part (called the approximate part) and the discrete part (called the details). The approximate part represents the main characteristics of the signal. The second step is to apply the similar operation to the

low frequency part. But at this time, the scaling factor has been changed. This operation is repeated until the desired scale is reached. In addition to continuous wavelet and discrete wavelet, there are wavelet packets and multi-dimensional wavelets in practical applications [41].

2.3.1 2-D Discrete Wavelet Transform

In 2D wavelets, there is one scaling function and three wavelets:

The scaling function $\varphi^{2D} = \varphi(x)\varphi(y)$ (2.19)

The three wavelets $\Psi_1^{2D} = \varphi(x)\Psi(y)$ (2.20)

$\Psi_2^{2D} = \Psi(x)\varphi(y)$ (2.21)

$\Psi_3^{2D} = \Psi(x)\Psi(y)$ (2.22)

where ϕ and ψ indicate the scaling function and 1-D wavelet, respectively. The discrete wavelet transforms of image $f(x,y)$ of size *M* and *N* is

$$W_\varphi(j_0,m,n) = \frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\varphi_{j_0,m,n}(x,y) \quad (2.23)$$

$$W_\varphi^i(j,m,n) = \frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\Psi_{j,\mathrm{m,n}}^i(x,y) \quad (2.24)$$

The image can be represented by the sum of orthogonal signals corresponding to different resolution scales. The detailed coefficients include the horizontal, vertical and diagonal details of the image. Fig. 2.3 illustrates the general form of the 2D wavelet transform. The decompositions first run along the x-axis, and then run along the y-axis. In the figure, $h_\Psi(-n)$ is an average filter. It outputs the average of its current input and its previous input. $h_\varphi(-n)$ is a moving difference filter. It outputs

half the difference between its current input and its previous input. $2\downarrow$ represents a down-sampling operator. It outputs at half the rate of the input.

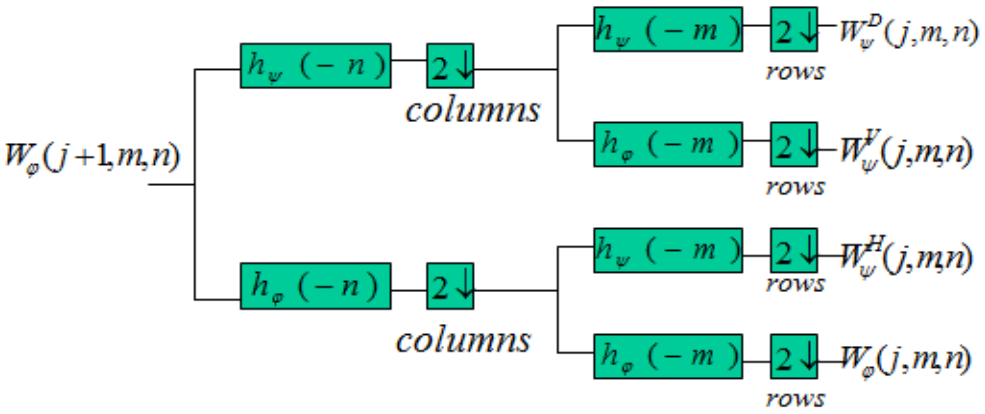


Figure 2.3: Fast 2D wavelet transform flow chart (retrieved from the CS6756 Digital Image Processing course note in Memorial University, winter 2012, Professor Siwei Lu)

Thus, an image can be divided into four bands: LL (left-top), HL (right-top), LH (left-bottom) and HH (right-bottom). An example is shown in Fig. 2.4. The sub-band $W_\varphi(j,m,n)$(LL) contains the smooth information and the background intensity of the image, and the sub-bands $W_\Psi^D(j,\mathrm{m},\mathrm{n}), W_\Psi^V(j,\mathrm{m},\mathrm{n})$ and $W_\Psi^H(j,\mathrm{m},\mathrm{n})$ contain the detailed information of the image. The sub-band $W_\varphi(j,m,n)$ (LL) is obtained by low pass filtering along the rows and then low pass filtering along the corresponding columns. It represents the approximated version of the original image at half resolution. $W_\Psi^H(j,\mathrm{m},\mathrm{n})$ (HL), representing the horizontal high frequencies (vertical edges), is the low pass filtering result along the rows. In contrast, $W_\Psi^V(j,\mathrm{m},\mathrm{n})$ (LH), representing the vertical high frequencies (horizontal edges), is the high pass filtering result along the columns. $W_\Psi^D(j,\mathrm{m},\mathrm{n})$ (HH), representing the high frequencies in

diagonal direction (corners and diagonal edges), is the filtering result by the high pass filter along both columns and rows [42].

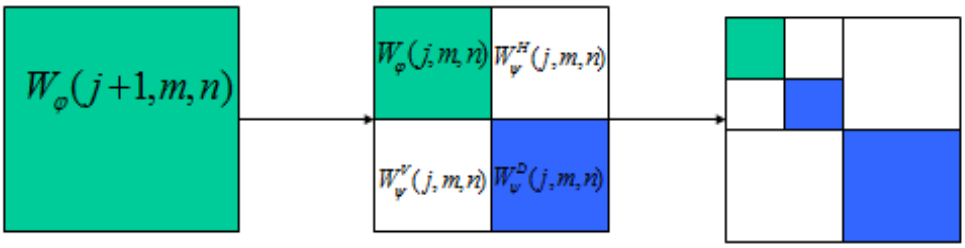
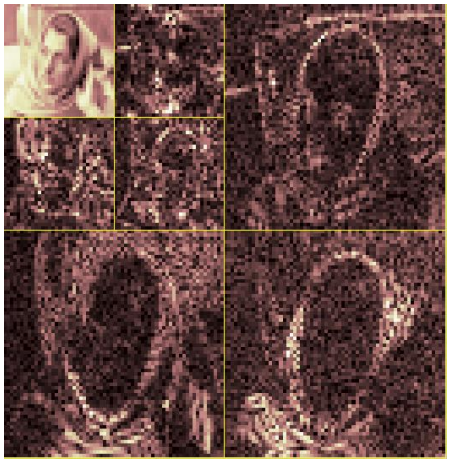


Figure 2.4: One and two level wavelet decomposition process (retrieved from the CS6756 Digital Image Processing course note in Memorial University, winter 2012, Professor Siwei Lu)

(a)

(b)



(c)

(d)

Figure 2.5: An image decomposition example: (a) Original image; (b) decomposition at level 1; (c) decomposition at level 2; (d) synthesized image (Images are taken from the Image Processing Toolbox in Matlab (R2010b))

The image of a woman in Fig. 2.5 (a) is decomposed with the Symlets wavelet (sym4 in Matlab Image Processing Toolbox). The first level decomposition of the original image is shown in Fig. 2.5 (b), and the second level decomposition result is shown in Fig. 2.5 (c). After the inverse DWT, the synthesized image is shown in Fig. 2.5 (d).

2.3.2 Applications

Wavelets can be applied in different application fields, including numerical analysis, image compression, image de-noising, image enhancement, image fusion, feature detection, edge detection and so on.

Wavelet analysis is a powerful and computationally efficient tool for numerical analysis. For example, it has been used in the solution of partial differential equations (ODEs) and integral equations (PDEs) [47].

Image compression algorithms based on the discrete cosine wavelet transform are basically decomposing signals in the frequency domain. In this way, it is easier to obtain important coefficients and achieve the best compression, as the correlation between signals can be removed. Take medical images for example; there is a need of local high resolution. Apparently, simple frequency domain analysis cannot meet that requirement. With the characteristic of time-frequency in the wavelet analysis,

coefficients can be dealt with in both domains. Different compression can be precisely provided in any interested part. Due to the advantages of the decomposition of detailed information, the Wavelet Scalar Quantization (WSQ) method is used to compress the FBI fingerprint database [43]. The new JPEG 2000 (Joint Photographic Experts Group) standard is also based on the wavelet transform. The compression procedure of the JPEG 2000 standard can be divided into three parts, the pre-processing, the core processing, and the bit-stream formation part [44]. The pre-processing part independently compresses an image into rectangular blocks. The core processing, mainly based on the discrete wavelet transform, is to decompose the tile components into different levels. Images are transformed to low-pass and high-pass samples, which represent a low-resolution version and a down-sampled residual version of the original set. After all coefficients are quantized, entropy coding is performed [45].

De-noising is critical in image processing, as noise is usually unpredictable, and exists in every step of image acquisition, processing, and outputting. In wavelet-based image de-noising, one wavelet and the decomposition level N are first chosen. Then one chooses a threshold for each of the N layers, and conducts quantization process to the high frequency coefficients in each layer. Finally, reconstruction is done using the low frequency coefficients in layer N and the modified high frequency coefficients from layer 1 to N [39]. With the Matlab Image Processing Toolbox,

de-noising functions such as ddencmp and wdencmp can effectively implement wavelet-based image de-noising.

In image processing, image enhancement can be conducted by setting a mask or modifying Fourier coefficients in the time domain or the frequency domain. However, these two methods either lose information or involve complex long calculation. Multi-scale analysis in wavelet, which is more flexible, provides a solution using as little amount of calculation as possible and choosing any decomposition levels to achieve satisfactory results [45].

The wavelet transform enables the possibility to distinguish between signal parts with different frequencies; therefore, it can be applied to feature detection. Schneiders proposed a real-time implementation of the DWT to detect features, and it was found that the detection speed of the wavelet filter was faster than a simple threshold-based detection [46].

2.4 Pattern Recognition

Pattern recognition first appeared in the 1920s. With the presence of computers in the 1940s, and the development of artificial intelligence in the 1950s, pattern recognition quickly developed into a subject in the 1960s [48]. Its theory and method in many science and technology research areas have attracted wide attention.

Therefore, it has promoted the development of the artificial intelligence system, which enlarged the possibility of computer applications.

In 1929, Tauschek invented a machine reader, which could recognize the numbers 0 to 9 [49]. Fisher proposed the statistical distribution theory in the 1930s, which laid the foundation of statistical pattern recognition. Two decades later, Noam Chomsky presented formal language theory, and Fu Jingsun presented sentence structure pattern recognition. The theory of fuzzy sets was raised by Zadeh ten years after that. Subsequently, fuzzy pattern recognition methods have been widely developed and applied. Hopfield presented neural network models reviving the artificial neural network, which became widely applied in pattern recognition in the 1980s. Small sample theory and support vector machines gained significant prominence in the 1990s [49].

2.4.1 The Concept of Pattern Recognition

Pattern recognition is a kind of technique dealing with artificial intelligence information. It has been widely applied in areas such as words, fingerprints recognition, and remote sensing. In industry, pattern recognition optimization techniques have produced enormous economic benefits in chemical and light industry, metallurgy, and others.

Broadly speaking, objects are themselves patterns [50]. Narrowly speaking, a pattern is the distribution of time and space information based on observations.

Patterns may be grouped or organized by measurable likeness into pattern classes [49]. The goal of pattern recognition is to identify the class to which a particular pattern belongs.

People can accomplish the job for small numbers of patterns; however, it could be extremely difficult for hundreds of millions of objects. Consequently, people allocated the task to computers. Generally speaking, pattern recognition is the analysis, description, classification and recognition of different sorts of things or phenomena using computers.

2.4.2 Pattern Recognition System

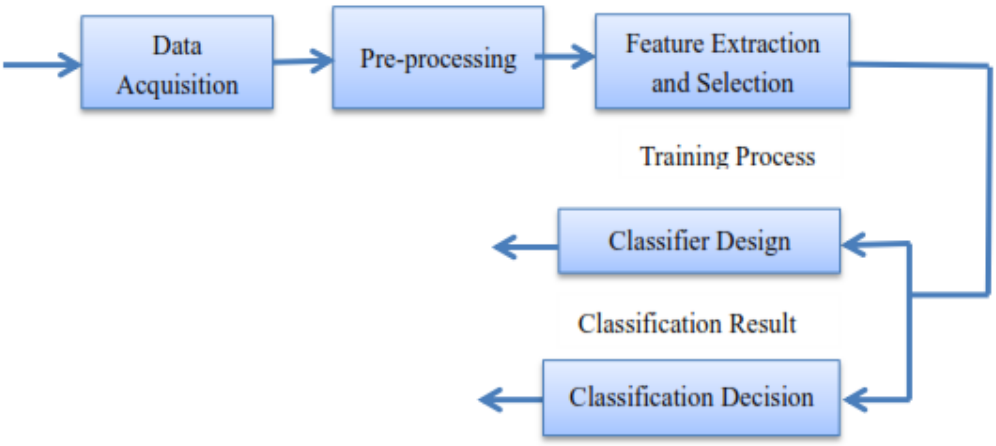


Figure 2.6: The composition of a pattern recognition system

As shown in Fig. 2.6, a pattern recognition system is mainly composed of four stages: data acquisition, pre-processing, feature extraction and selection, and classifier design and classification decision. The goal is to assign a pattern to a certain group or category. The data acquisition stage transfers all kinds of information about the study of objects to digital number or symbol sets that can be accepted by machines. The

pre-processing stage is to remove noise, strengthen valuable information, and recover information degraded during acquisition or transfer. It usually includes binarization processing, edge extraction, image segmentation, digital filtering, de-noising processing, and normalization.

The next stage is to extract relevant data features. For example, in fingerprint recognition, features such as texture, intersection, and shapes can be extracted. The space that contains all original data is called the measurement space, and the related data is eventually classified in the object class feature space. At this point, the high dimensional measurement space has been transformed to lower dimensional feature space. Analysis of the feature space will identify the most relevant features. This process is called feature selection. A group of stable and typical features is the core of a recognition algorithm. Although two recognition algorithms use the same classification strategy, they belong to different algorithms when using different features.

Feature extraction and selection are of vital importance in the recognition process. If a pattern is properly chosen, it will show large variances to different patterns, and we can easily design a classifier with high performance. Therefore, feature selection would directly influence the design of a classifier and the result of the classification. Although feature extraction and selection hold a very important place in pattern recognition, there are no general methods so far.

Determining the optimum number of features to use is not straightforward. Commonly, when a classifier with one set of features cannot satisfy the demands, we would naturally think of adding new features. However, adding features will increase the difficulty of feature extraction and the complexity of classification calculation. In practical applications, it is found that the performance of a classifier will be the same or even worse when the number of features reaches a certain limit. This problem is mainly due to the limited sample size of data. In this case, to satisfy the classification result, samples for learning must be increased at the same time as adding features [48].

A variety of feature selection methods have been developed, which can be divided into three categories: filter methods, wrapper methods, and hybrid methods [24]. In filter methods, the Sequential Forward Selection (SFS), proposed by Whitney in 1971, is one of the most commonly used approximate optimal methods [25]. This method starts from an empty feature set and iteratively adds a new feature from the remaining features. In contrast, the Backward Forward Selection (SBS) removes one feature each time from the full feature set. The Wrapper approaches apply specific machine learning algorithms such as the decision tree or support vector machine (SVM), and utilize the corresponding classification performance to guide the feature selection [26]. The Hybrid method is a combination of the advantages of the Filter and Wrapper methods. An experiment by Jain [27] with different feature selection methods showed

that the sequential forward floating selection (SFFS) algorithm, proposed by Pudil [28], outperformed the other algorithms.

The last stage of pattern recognition is classifier design and classification decision. The output of this part could be a certain pattern that an object belongs to, or the most similar pattern number in a pattern database. The design of a classifier is usually based on the pattern set, which has been classified or described. This pattern set is called the training set, and this result learning strategy is called supervised learning. There is also unsupervised learning, which needs no prior knowledge, but is based on the statistical law or similarity learning to classify each object's category. Among various pattern recognition methods, the most commonly used are pattern matching, statistical pattern recognition, syntactic pattern recognition, fuzzy pattern recognition and neural network pattern recognition.

2.4.3 Applications

Pattern recognition can be applied on different subjects, such as speech recognition, speech translation, face recognition, fingerprint recognition, handwriting character recognition. After decades of research and development, pattern recognition technologies have been widely used in a variety of fields, including artificial intelligence, computer engineering, machine learning, neural biology, medicine, archaeology, geological prospecting, space science, remote sensing, and industrial fault detection [39]. Furthermore, the fast development of pattern recognition

technologies can also greatly enhance the development of military science and technology [48].

In medical applications, such as cancer detection, X-ray image analysis, blood tests, chromosome analysis, electrocardiogram and electroencephalogram diagnosis, pattern recognition is also critically important [48]. Existing research put great effort into the detection of microcalcifications in breast cancer detection. In 2010, Balakumaran first employed dyadic wavelet transform to enhance mammogram quality, and detected 95% of microcalcifications in his experiment by fuzzy shell clustering [16]. Chan et al. proposed a different computer-aided diagnostic method to detect microcalcifications on digitized mammograms, which improved the classification accuracy [17]. This method was aimed at improving the signal-to-noise ratio (SNR) by linear spatial filters. In Jinchang Ren and Zheng Wang's recent work, they proposed an improved SVM approach designed for effective classification of benign and malignant microcalcifications in mammograms. The experiment results showed nearly 20% improvement in terms of the area under the ROC curve (Az) [18].

In 2007, Kage et al. [21] compared the performances of some state-of-the-art methods for mass detection in mammograms. Their experiments were based on two databases that are free to the public: Mammographic Image Analysis Society's digital mammogram database (MIAS) [19] and the Digital Database for Screening Mammography (DDSM) [20]. The results showed that the Gradient Orientation

Analysis (GOA) developed by Brake and Karssemeijer achieved the best results for both databases. The Analysis of Local Orientated Edges (ALOE) method presented by Kegelmeyer et al. [22] achieved the second best results. The standard deviation of folded gradient orientations method, named Liu method [23], achieved the worst results. After adding the novel gradient direction analysis to the Liu Method, the performance was significantly increased.

Chapter 3 – Mammogram Image Processing

The mammogram image processing is the first stage of the proposed mammogram analysis system. This stage includes two basic steps: mammogram image pre-processing and data transforms. In the mammogram image pre-processing step, the original digital mammogram images are de-noised and normalized. In the data transforms step, the normalized images are decomposed by the Fourier transform and three wavelet transforms with different bases (Daubechies db2, Daubechies db4, and Biorthogonal bior6.8) separately. Then, four statistical features, including the mean, standard deviation, skewness and kurtosis of the image intensities, are extracted.

3.1 Mammogram Image Pre-processing

For the automatic mammogram analysis system, the original images are different in size and directions. Furthermore, artefacts and noise may also exist in some mammograms, which would generate wrong or poor analysis result. Thus, several mammogram pre-processing steps were implemented to regularize the appearance of the images, and remove unnecessary artefacts and noise. Based on the studies [62-63], the steps taken in this work include orientation matching, background thresholding, and intensity matching.

3.1.1 Orientation Matching

In this study only the MLO mammogram presentation was used. In these, the right and left breasts point to the opposite sides in the mammogram image. Therefore, it is better to flip one of the breasts to the same direction as the other one. This step ensures that all images pointed to the same direction, preventing changes in the wavelet transform coefficients due only to the directionality change between right and left images. The sharp edge between the tissue and the dark background is a major feature in all images that affects this change. As shown in Fig. 3.1, the intensity of right breast images falls from left to right across this edge, while it rises in left breast images. This would change the sign of the calculated wavelet coefficient.

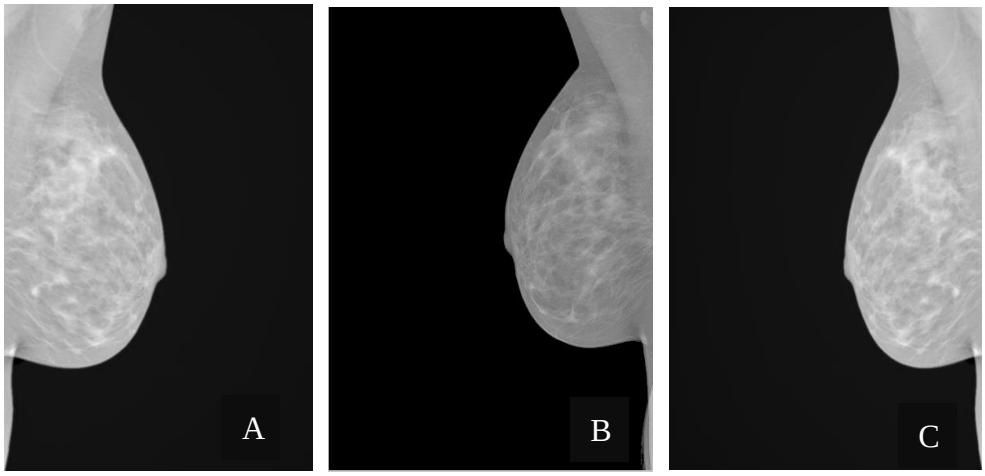


Figure 3.1: An example of MLO view mammogram: A. Right side; B. Left side; C. the image after orientation matching of A.

Fig. 3.1 shows the result of orientation matching of an example of Medial Lateral Oblique (MLO) view mammogram. Fig. 3.1 A and B respectively show the right and left breast images of a patient with tiny microcalcifications in her breast tissue. Fig. 3.1 C shows the reflected image of orientation matching of the right breast.

3.1.2 Background Thresholding

Signal outside the tissue is non-informative and was removed from consideration by binary masking. Thresholding is the simplest method to create binary images, and it normally sets all pixels below a set intensity level to zero [41]. A satisfactory threshold can remove all irrelevant information in the background pixels, and leave foreground objects unaltered. A most commonly used method to choose the threshold is Otsu's Method [54], which assumes that the image to be thresholded contains two classes of pixels or bi-modal histogram (e.g. foreground and background). The method then calculates the optimum threshold separating those two classes so that their combined spread (intra-class variance) is minimal [54]. It also assumes that the foreground and background intensities are normally distributed, and it chooses the threshold level which minimizes the segmentation error between the two regions.

The attenuation of x-rays passing through the tissue affects the intensity in the images, and is influenced by the thickness and density of the tissue. Therefore, tissue pixels which fall below the conservative threshold are predominantly from the edges of the tissue region where the breast tissue is thin and uncompressed. While a few pixel layers may be removed by this method, it was deemed acceptable as any pathology that exists this close to the surface of a patient's skin should be readily detectable by conventional examination without the aid of mammography.

In this work, the binary thresholding, which sets all pixels below a threshold, was set to an intensity of zero and all pixels above the threshold to an intensity of one (see Fig. 3.2). The output image of the process is the pixel-by-pixel product of the binary mask image and the original image. In this way, all background pixels of the output image are set to zero of intensity, while all foreground pixels are unaffected.

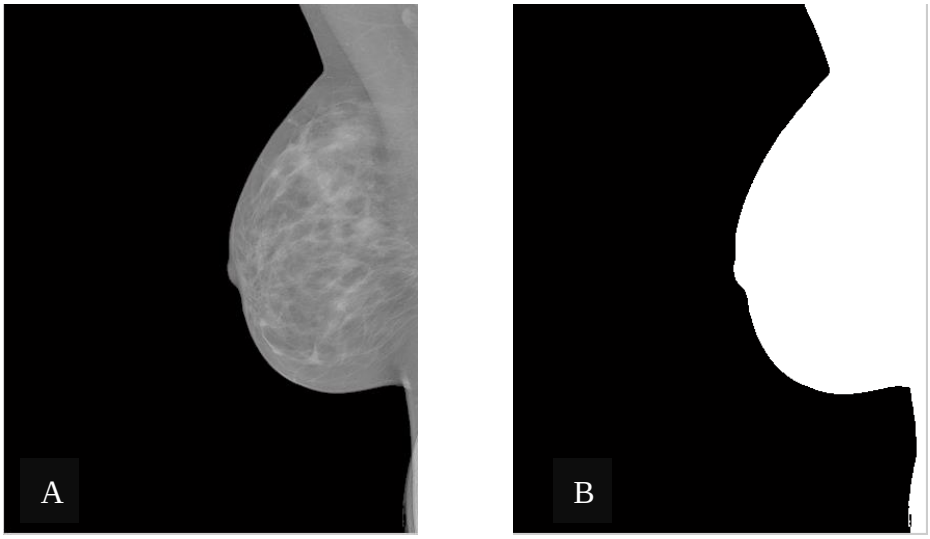


Figure 3.2: A. Mammogram image before background thresholding; B. The thresholded binary image used to mask the original image.

3.1.3 Intensity Matching

Intensity matching is the last pre-processing step applied to the images before they are ready for data tranforms. In this step, all mammograms are linearly scaled to an intensity of 0.0 to 1.0. This intensity matching process can be defined by

$$img\_out = \frac{img\_in}{\max(img\_in)}, \tag{3.1}$$

where *img_in* is the input image following the background thresholding step, and *img_out* is the intensity-matched image whose pixel intensities range from zero to one. This step ensures the uniformity across all different mammogram images, because their pixel intensities ranges could differ with machines settings. It can be seen in Fig. 3.3 that there is tiny difference before and after the intensity matching procedure. The broader spread in intensities would increase the variations in different tissue types and densities. (the maximum relative intensity prior to normalization was 0.92).

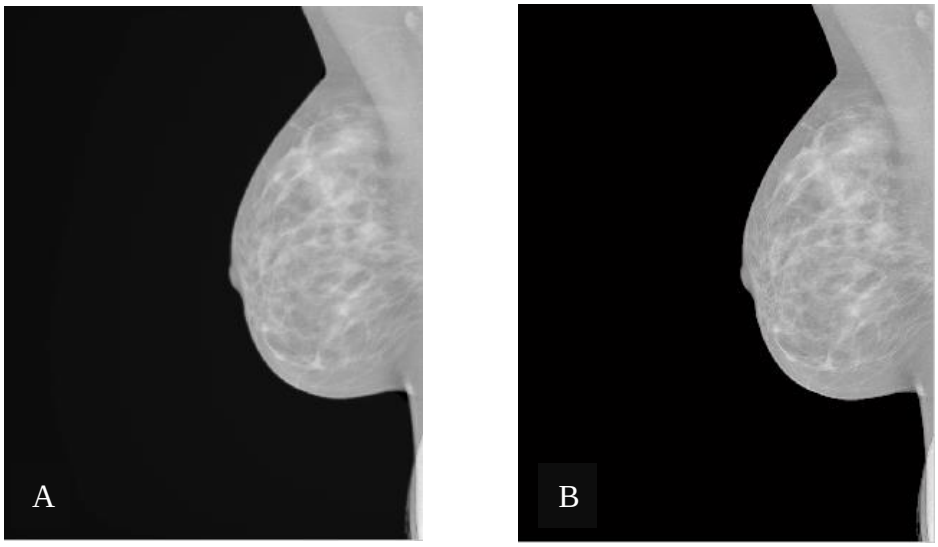


Figure 3.3: Mammogram image before A and after B intensity matching

3.2 Data Transforms

Once the images are pre-processed to minimize the differences between images that were not related to differences in the physical composition of the breast tissue, the wavelet and Fourier transforms were performed on the images.

The images were all sampled to 1024×1024 pixels, which would allow maximum 10 levels of decomposition, since dyadic sampling reduces the dimensions by a factor

of two in each direction after each pass. In this work, only eight levels of decomposition were used. Because the final two levels would consist of four-pixel and one-pixel images, respectively, which are basically useless for mammogram analysis, compared to the size of the entire breast. As a result, these levels are omitted from the wavelet analysis to speed calculation.

3.2.1 Choice of Transform Methods

1. Daubechies Wavelets: db N

This discrete orthogonal wavelet was developed from two-scale equation coefficient $\{h_k\}$ by Ingrid Daubechies, which makes discrete wavelet analysis practicable [33]. The names of the Daubechies family wavelets are written as "db N", where N is the order, and db is the "surname" of the wavelet. Except when N=1 (Haar), db N is asymmetric and has no explicit expressions. But there are explicit expressions for the square modulus of the transfer function of $\{h_k\}$. Assuming that $P(y) = \sum_{k=0}^{N-1} C_k^{N-1+k} y^k$, C_k^{N-1+k} is the coefficient of binomials, then,

$$|m_0(\omega)|^2 = \left(cos^2 \frac{\omega}{2}\right)^N P(sin^2 \frac{\omega}{2}) \tag{3.2}$$

in which, $m_0(\omega) = \frac{1}{\sqrt{2}} \sum_{k=0}^{2N-1} h_k e^{-ik\omega}$.

The Daubechies wavelets are chosen for their sensitivity to various types of intensity gradients. Fig. 3.4 shows the wavelet and scaling functions of two Daubechies wavelets used in this work: Daubechies 2 and Daubechies 4.

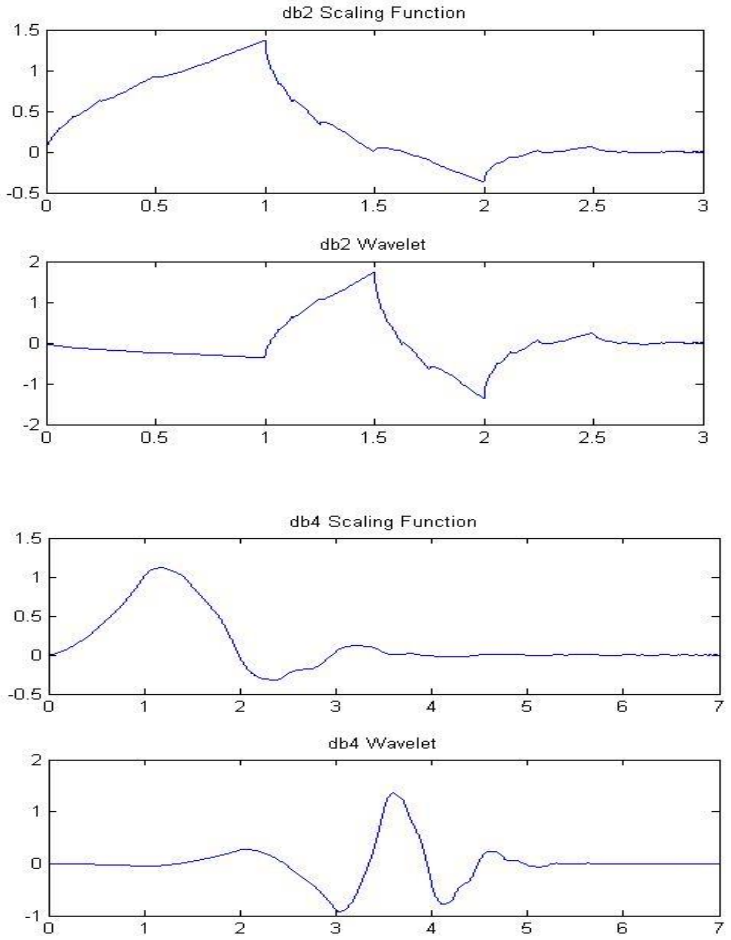


Figure 3.4: Wavelet functions (high pass filters) and scaling functions (low pass filters) for Daubechies 2 and Daubechies 4 [55].

2. Biorthogonal Wavelet Pairs: bior*Nr.Nd*

The main characteristic of the Biorthogonal function is it can feature linear phase, and it is mainly used in the reconstruction of signals and images. The Biorthogonal wavelet family uses a pair of associated scaling filters (instead of the same single one) for reconstruction and decomposition [33]. The Biorthogonal function is denoted as bior*Nr.Nd* form in Table 3.1[36], in which, r denotes reconstruction, d denotes decomposition.

Table 3.1: bior*Nr*.*Nd* form

Nr	*Nd*
1	1,3,5
2	2,4,6,8
3	1,3,5,7,9
4	4
5	5
6	8

The Biorthogonal wavelets are chosen for their ability to provide exact reconstruction. Fig. 3.5 shows the decomposition (analysis) and reconstruction (synthesis) filters for the Biorthogonal bior6.8 wavelet. The wavelets and their associated scaling functions are shown in the discrete form, since this is the form used to decompose the mammogram images [55].

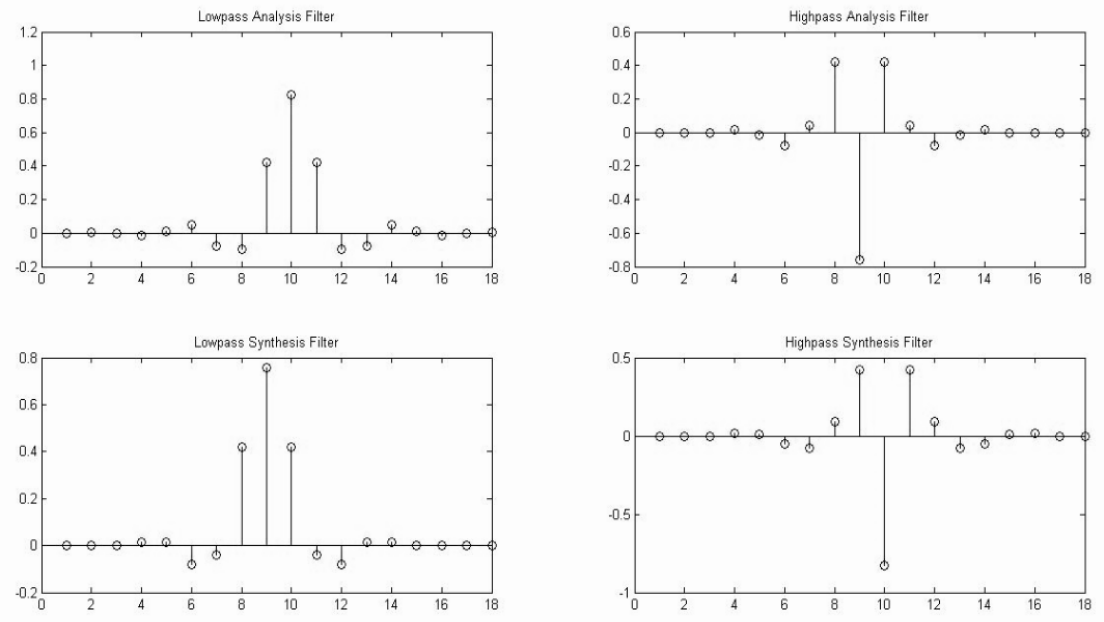


Figure 3.5: Decomposition (analysis) and reconstruction (synthesis) filters for the Bior6.8 wavelet

3. Fourier Transform

The time or space based data is usually transformed to frequency-based data after the Fourier transform, as shown in Fig. 3.6. The Discrete Fourier Transform (DFT) of a vector x of length n is another vector y of length n according to the following equation:

$$y_{p+1} = \sum_{j=0}^{n-1} \omega^{jp} x_{j+1} \tag{3.3}$$

where ω is a complex nth root of unity:

$$\omega = e^{-2\pi i/n} \tag{3.4}$$

Here, i is the imaginary unit, and p and j are indices that run from 0 to $n-1$.

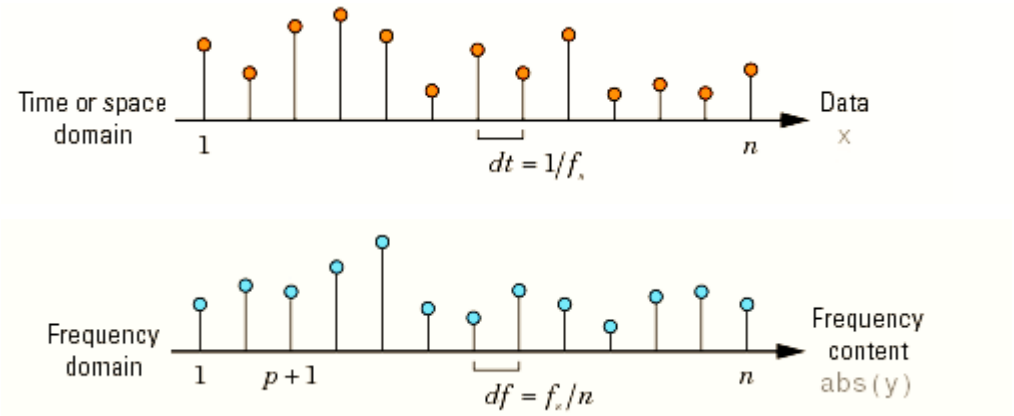


Figure 3.6: Fourier transform between the time/space and frequency domain [55]

4. Comparison

Fig. 3.7 shows the original mammogram and its four detail views obtained at the first decomposition level when the Db4 wavelet basis is used. It is shown that the

wavelet maps have a lower resolution than the original image. Each view is sensitive to different features in the image. For example, the horizontal detail detects vertical changes in intensity, the vertical detail detects horizontal changes in intensity, the diagonal detail responds when the intensity is varying in both directions, and the approximation image is a low resolution version of the original image used as an input to the next coarser level of the decomposition.

Fig. 3.8 shows the Fourier transform view of the original mammogram. Compared with the wavelet maps, it can be seen that the wavelet transform provides multi-resolution decomposition, which means the wavelet maps at different levels reflect the image features of different sizes. Furthermore, spatial information is partially conserved. The wavelet maps in Fig. 3.7 show the spatial distribution of information at particular size scales; in contrast, the Fourier transform would lose the spatial information and simply produce a map of the relative contributions of different frequencies over the entire image. This spatial information is useful for finding localized structures, such as microcalcifications and masses. These structures remain localized after the wavelet transform is applied, and their can then be distinguished from a more homogeneous background.

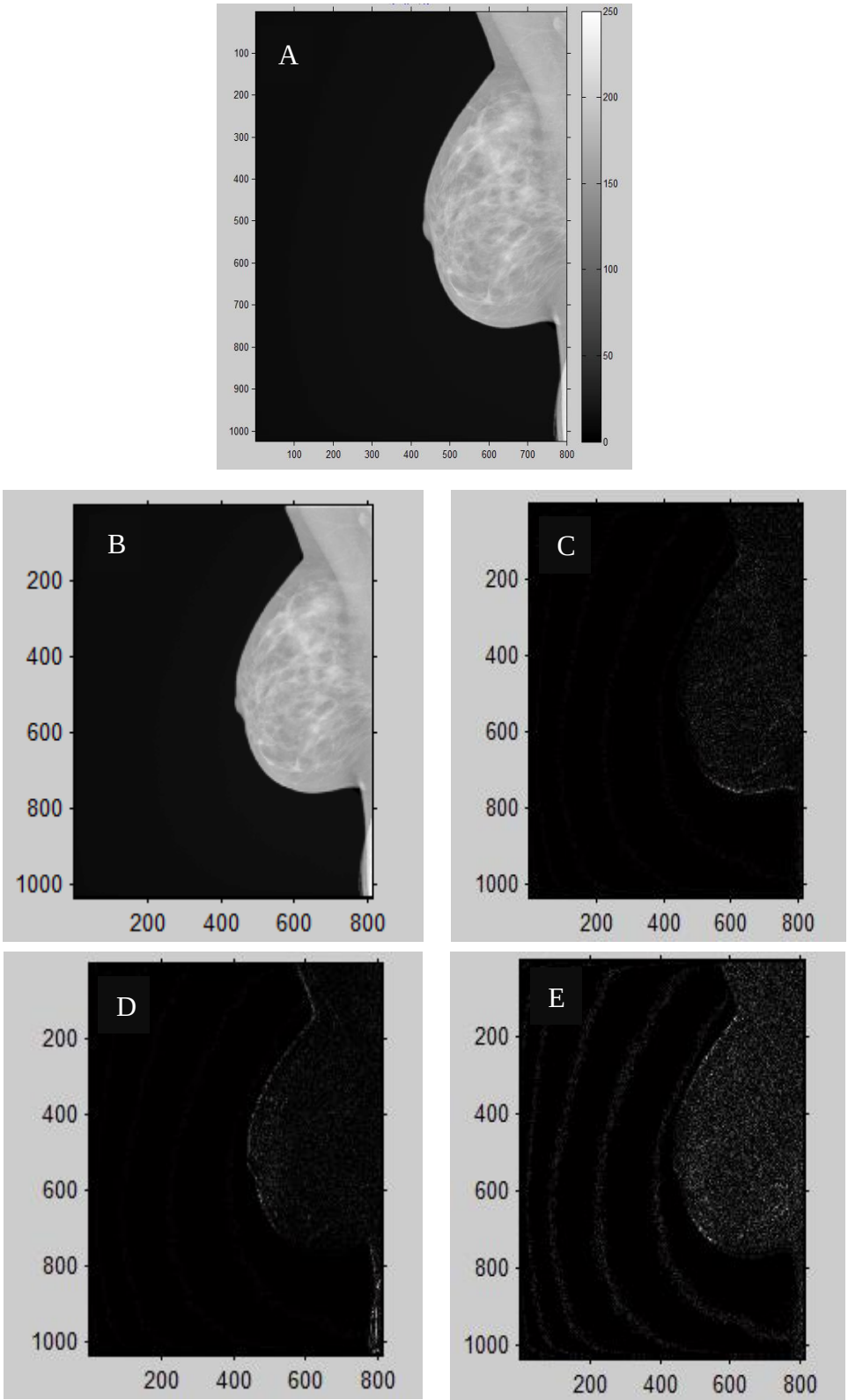


Figure 3.7: First level db4 wavelet decomposition: A. Original mammography image; B. Approximation view; C. Horizontal detail view; D. Vertical detail view; E. Diagonal view.

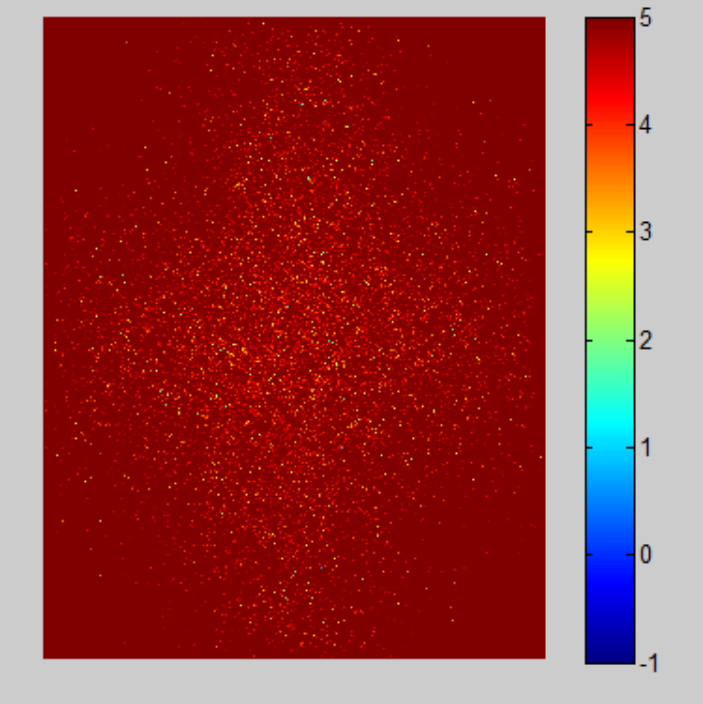


Figure 3.8: The Fourier transform view of the mammogram in Fig. 3.7 A.

3.2.2 Choice of Measurement

In this experiment, four statistical features were extracted: mean intensity, standard deviation, skewness and kurtosis of the pixel intensities. Then, the mammogram analysis system uses some of these features to classify mammogram images as being normal or cancerous.

1. Mean

The mean μ in this paper is obtained by calculating the average pixel value of the tissue region in the mammogram image. The equation is given by

$$\mu = \frac{1}{N}\sum_{i,j} I(i,j) \tag{4.3}$$

where $I(i,j)$ is the pixel value at point (i,j) of the mammogram image. N is the number of pixels in the tissue region of the image. The mean feature measures the average value of each detail views at different decomposition levels.

Microcalcifications are usually tiny and bright. Compared with normal samples, microcalcifications have a slightly higher intensity in the high resolution maps. While masses are usually different in sizes and shapes, they could range from millimetres to several centimetres in width. Therefore, masses cannot be extracted from the background tissue through single scale or wavelet basis. However, masses are located in one region of tissue, and they are usually brighter than normal tissue. As a result, a slightly larger mean intensity can be measured through a wavelet basis, especially when different scales are used to detect masses.

2. Standard Deviation

The standard deviation σ, the estimate of the mean square deviation of grey pixel values, describes the dispersion of a local region. It is defined as

$$\sigma = \sqrt{\frac{1}{N}\sum_{i,j}[I(i,j) - \mu]^2} \tag{4.4}$$

It measures the variability in the brightness of the image over the tissue region. The value of the standard deviation would increase in the high spatial resolution levels of the wavelet map images that contain microcalcifications or masses, because they are brighter than normal parts of mammogram images.

3. Skewness

The third statistic feature measured from each wavelet map image is the skewness of the pixel intensities, which measures the degree of asymmetry. The skewness of a distribution of values is defined as the third central moment of the distribution, normalized by the cube of the standard deviation. It is given by

$$S = \frac{1}{N}\sum_{i,j}[\frac{I(i,j)-\mu}{\sigma}]^3 \quad (4.5)$$

When a distribution has a larger right tail, then it shows a positive skewness. Even there is no significantly difference in the mean value or standard deviation, the skewness still changes because it is sensitive to the addition of a small number of unusually small or large values on a distribution.

4. Kurtosis

The fourth statistic measured from the wavelet maps is the kurtosis of the pixel intensities. The kurtosis of a distribution of values is defined as the fourth central moment of the distribution, normalized by the fourth power of the standard deviation of the distribution. The kurtosis *K* is given by

$$K = \frac{1}{N}\sum_{i,j}[\frac{I(i,j)-\mu}{\sigma}]^4 \quad (4.6)$$

Kurtosis measures the narrowness of the central peak of a distribution compared with the size of the distribution's tails. A distribution with a narrow peak and tails that drop off slowly has a large kurtosis compared with a distribution with a relatively wide peak but suppressed tails. The kurtosis and standard deviation of a distribution may be similar, but kurtosis is more sensitive to points distant from the mean than the

standard deviation. Because of this, kurtosis is sensitive to the presence of microcalcifications and masses. It will rise when the number of unusual bright pixels increases in a wavelet map.

Chapter 4 – Feature Selection and Image Classification

The second and third stage of the mammogram analysis system, the feature selection and image classification stages, are introduced in this chapter. First, an entropy-based feature selection algorithm is proposed to reduce the number of features, which were extracted from the transformed mammogram images. Then, three classifiers (the Linear Discriminate Analysis, the Back-propagation Network, and the Naive Bayes classifier) and a voting classification scheme are proposed and discussed in detail. The classifiers would be used based on the features after the feature selection. Finally, the classification accuracy, sensitivity, specificity, and Receiver Operator Curve (ROC), are presented for the evaluation of the classifiers.

4.1 Feature Selection

Since a large number of potential classification features are generated from each mammogram image, a selection process is needed to choose those features that are most effective at differentiating between normal and cancerous images. Specifically, there are four parameters measured from each wavelet map, with four wavelet maps per level and eight levels of decomposition. Thus, 16 features could be generated form each level of decomposition. To eliminate some of these, it was noted in N. Terki, etc.

[59] that peak signal to noise ratio (PSNR) improved when the level of decomposition increases, and the image quality was better from third level of decomposition. Therefore, level 3 to level 8 of decomposition of the proposed three wavelet transform methods were applied in this work. In this case, 96 features would be generated from each of three wavelet transforms based on the 6 levels of wavelet decomposition.

Then, the generated 96 features from the wavelet transform were combined with the 6 features extracted from the Fourier transform. In other words, 3 different feature sets were created, and each of the feature sets contains features from one wavelet transform and the Fourier transform.

For a feature to be useful in classification, it should be closely and uniquely associated with a certain class [56]. Ideally, the feature will correlate with the desired class independent of the presence of other classes. If these conditions are met, the feature reduction (selection) problem can be addressed by measuring the correlation with that class then establishing a pass threshold. The pass threshold eliminates features that correlate poorly. There are two common approaches used to measure the correlation between two random variables, in this case between feature and class [57]. The first is linear correlation, where the variation in a feature value is compared to the variation in a class value. This is clearly not applicable here where the class variable has two values, normal or suspicious. The second approach and the one adopted for

this research is Information Gain, a concept based on the reduction of entropy in the dataset.

A target range for number of features was determined from the work of, Lei and Huan [58], they proposed a fast correlation based filter approach and conducted an efficient way of analyzing feature redundancy. Their new feature selection algorithm was implemented and evaluated through extensive experiments comparing with other related feature selection algorithms based on ten different kinds of feature types. The number of features ranged from 57 to 650, and the sample size of feature types ranged from 32 to 9338. At the end of the experiment, they recorded the running time of the proposed system and the number of features selected for each algorithm. The results showed that the average selected number of features was 15 for the five compared feature selection algorithms, and the selected features could lead classification accuracy to around 89%. In this research, we chose a threshold of information gain which could lead to around 15 features left.

4.1.1 Principle

Entropy is a measure of the uncertainty of a random variable [58]. The entropy of a variable *X* is defined as

$$H(X) = -\sum_i P(x_i) log_2(P(x_i)) \quad (4.1)$$

and the entropy of *X* after observing values of another variable *Y* is defined as

$$H(X \mid Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j) log_2(P(x_i \mid y_j)) \quad (4.2)$$

where $P(x_i)$ is the prior probabilities for all values of *X*, and $P(x_i|y_j)$ is the posterior probabilities of *X* given the values of *Y*. The amount by which the entropy of *X* decreases reflects additional information about *X* provided by *Y*, and is called information gain [46], given by

$$IG(X \mid \mathrm{Y}) = H(X) - H(X \mid \mathrm{Y}) \tag{4.3}$$

If we have $IG(X \mid \mathrm{Y}) > IG(Z \mid \mathrm{Y})$, it means a feature *Y* is regarded more correlated to feature *X* than to feature *Z*.

4.1.2 Algorithm

The entropy with the feature selection algorithm was implemented by the following steps:

1. Order features based on decreasing entropy values (using Equation 4.1), and build a link list for all features;

2. Calculate the entropy of each feature in the link list related to the classification results using Equation 4.2;

3. Calculate the information gain of each feature using Equation 4.3 based on its two entropies obtained from step 1 and 2;

4. Compare each feature's information gain with the next feature, and move the larger one ahead till the end of the link list;

5. Select the features with the information gain larger than the threshold set in the program.

4.2 Image Classification

4.2.1 Linear Discriminate Analysis

Linear Discriminate Analysis (LDA), a widely used algorithm for pattern recognition, was introduced by Belhumeur in 1996 [48]. The main idea is to reduce the dimensionality of the dataset in a manner that preserves class discrimination. Data is projected onto a vector so as to maximize class-mean separation and minimize intra-class variability. Conceptually, this means that the data are now arranged linearly along a vector. Ideally, the classes of interest are completely separated and the feature or features can be used to classify newly introduced data [48]. After projection, pattern samples in the new subspace have the biggest between-class distance and the minimum within-class distance, which guarantee the best separability in the space.

Suppose there are *N* samples, $\{x_1^{(i)}, x_2^{(i)},\ldots, x_d^{(i)}\}$, in *d* dimensions, which belong to class C. The objective function of LDA is as follows:

$$W_{best} = argmax_w \frac{W^T S_b W}{W^T S_W W} \tag{4.4}$$

$$S_b = \frac{1}{n}\sum_{i=1}^{c}(m^i - m)(m^i - m)^T \tag{4.5}$$

$$S_w = \frac{1}{n}\sum_{i=1}^{c}(\sum_{j=1}^{n_i}(x_j^i - m^i)(x_j^i - m^i)^T) \tag{4.6}$$

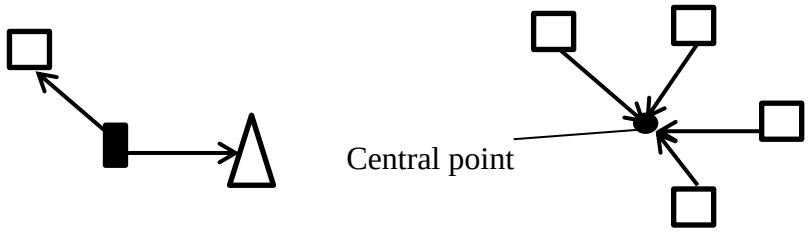
where m is the total sample mean vector, n_i is the number of samples in class C_i, m^i is the average vector associated to C_i class, x_j^i is the $j-th$ sample vector in the $C_i - th$ class. S_b and S_w are named between-class scatter matrix and within-class matrix, respectively.

As mentioned above, the objective of LDA is to make the data points of different classes as far apart from each other as possible. In addition, it also aims at making the data points from the same class as close as possible. These two purposes can be decomposed into the following functions:

$$max \sum_{i=1}^{c} n_i W^T (m^i - m)(m^i - m)^T w \tag{4.7}$$

$$min \sum_{i=1}^{c} (\sum_{j=1}^{n_i} W^T (x_j^i - m^i)(x_j^i - m^i)^T w \tag{4.8}$$

This procedure can also be illustrated in Fig. 4.1(a) and (b).



(a) (b)

Figure 4.1 Data points with the same shape belong to the same class. (a): Diagram of between-class procedure; (b): Diagram of within-class procedure.

4.2.1.1 Algorithm

1) Constructing a matrix of feature vectors. All feature samples were read in as a matrix $X = \{x_1, x_2, \dots, x_n\}$. Each feature data x_i was regarded as a node i, and in the same way, another feature data x_j was regarded as a node j. Node i and j were connected with a line if x_i and x_j were close, and they belonged to the same class.

2) Calculating scatter matrixes. In this step, between-class scatter matrix S_b and within-class matrix S_w were calculated using Equations 4.5 and 4.6.

3) LDA projection. Data points x_i were projected into the LDA subspace so that the matrix S_w was non-singular. The transformation matrix of LDA was presented here as W_{LDA}. After projection, S_b and S_w became

$$\widehat{S_b} = W_{LDA}^T S_b W_{LDA} \tag{4.9}$$

$$\widehat{S_w} = W_{LDA}^T S_w W_{LDA} \tag{4.10}$$

4) Computing the projection matrixes. After adding the Lagrange multiplier and some derivation steps, the following function was achieved. It is also called the Fisher Linear Discrimination.

$$S_w^{-1} S_b W = \lambda W \tag{4.11}$$

It can be seen that W is the eigenvector of matrix $S_w^{-1} S_b$.

5) Linear embedding. With the substitution of eigenvector, W_{best} is easy to find by the following equation:

$$\mathrm{W} = S_w^{-1}(\mu_1 - \mu_2) \tag{4.12}$$

where μ is the mean value (central point) of samples in each class.

$$u_i = \frac{1}{n_i} \sum_{x \in C} x \tag{4.13}$$

4.2.2 Back-propagation Network

The BP neural network is an abbreviation for the error back propagation algorithm, and is commonly used in the artificial network [49]. It consists of information forward propagation and error backward propagation. As shown in Fig. 4.3, the typical BP

network is a three layer network, which includes input layer, hidden layer and output layer. Information is introduced, weighted then passed to the output layer. The results are compared with the desired outcome (training phase) and the error term minimization used to adjust the weighting.

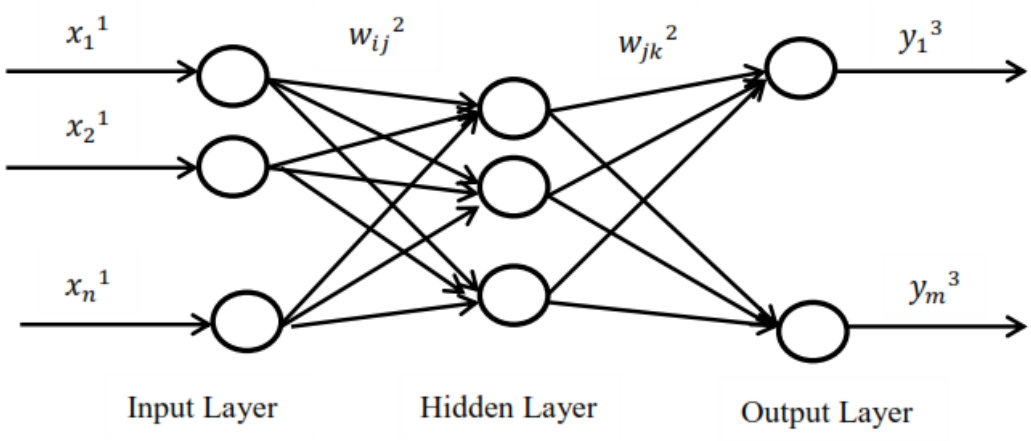


Figure 4.2: BP neural network

[1] $x_1, x_2, \ldots, x_n$: the input values of the BP network.
[2] w_{ij}, w_{jk}: the weight values.
[3] $y_1, \ldots, y_m$: the estimated values.

The BP network can also be seen as a non-linear function, which establishes a mapping relationship from n independent variables to m dependent variables.

4.2.2.1 Algorithm

1) Network initialization. According to the input and desired output values (X and Y) of the network, we can set n nodes in the input layer, l nodes in the hidden layer, and m nodes in the output layer. The weight values (w_{ij} and w_{jk}), the threshold value a in the hidden layer, the threshold value b in the output layer, the learning speed, and the activation functions should also be initialized.

2) Calculation of the hidden layer output. This output *H* can be achieved through $X, w_{ij},$ and a .

$$H_j = f\left(\sum_{i=1}^{n} w_{ij} x_i - a_j\right) \qquad j = 1,2,\dots,l \tag{4.14}$$

Here, l is the number of nodes in the hidden layer, f is an activation function. In this work, the activation function is chosen as

$$f(x) = \frac{1}{1+e^{x}} \tag{4.15}$$

3) Calculation of the output layer output. *O* is determined through $H, w_{jk},$ and b .

$$O_k = \sum_{j=1}^{l} H_j w_{jk} - b_k \qquad k = 1,2,\dots,m \tag{4.16}$$

4) Calculation of error. Based on the output O and the estimated output Y, we can obtain the prediction error e.

$$e_k = Y_k - O_k \qquad k = 1,2,\dots,m \tag{4.17}$$

5) Weight update. The values can be updated using the following equations:

$$w_{ij} = w_{ij} + \eta H_j\left(1 - H_j\right) x(l) \sum_{k=1}^{m} w_{jk} e_k \qquad i = 1,2,\dots,n; j = 1,2, \tag{4.18}$$

$$w_{jk} = w_{jk} + \eta H_j e_k \qquad j = 1,2,\dots,l; k = 1,2,\dots,m \tag{4.19}$$

in which, η is the learning speed.

6) Threshold update.

$$a_i = a_i + \eta H_j\left(1 - H_j\right) \sum_{k=1}^{m} w_{jk} e_k \qquad j = 1,2,\dots,l \tag{4.20}$$

$$b_k = b_k + e_k \qquad k = 1,2,\dots,m \tag{4.21}$$

7) If the iteration is not over, the algorithm goes back to the second step.

4.2.2.2 Implementation

The aim of the algorithm was to classify mammograms into two categories: cancerous or normal. Because the input features are 14 dimensional, and there are two kinds of mammograms to be classified, the construction of the BP network can be defined as “14-15, 2”. It means that there are 14 nodes in the input layer, 15 nodes in the hidden layer, and 2 nodes in the output layer. Furthermore, after random sorting of 670 mammograms, 520 of them were randomly selected as the training dataset, the remaining 150 were chosen to test the classification performance of the BP network.

4.2.3 Naive Bayes Classifier

The Naive Bayes classifier (NB), which is quite popular for its simplicity in implementation, is a probabilistic classifier using the Bayes' theorem [51]. It is also a supervised learning method, by using an approximation algorithm. Furthermore, the NB classifier has the following features [52]:

(1) Instead of assigning an instance to a certain category, it calculates the probability of this instance belonging to each category and chooses the largest one.

(2) All features are usually involved in the Bayes classification process. Only one or several features cannot determine the classification result.

(3) The features for the Naive Bayes classifier can be discrete, continuous, or hybrid.

4.2.3.1 Algorithm

Suppose that $A = \{A_1,\ A_2,\ldots A_n\}$ are the features for one dataset, and there are m classes, $C = \{C_1, C_2, \ldots C_m\}$. Given an instance, its feature is $X = \{X_1, X_2, \ldots X_n\}$, then the posterior probability that instance belongs to a class C_i is $P = P(C_i) = (X \mid C_i)$. The Bayes classifier can be represented as

$$C(X) = \underset{C_i \in C}{\arg max}\ P(C_i)\ P(X \mid C_i) \tag{4.22}$$

It indicates that the prediction accuracy reaches the maximum value when instance X has the largest posterior probability.

However, the posterior probability in Equation 4.22 is difficult to calculate. Therefore, the "Naive Bayes hypothesis", which assumes all features A_i are independent from each other, is introduced to the Naive Bayes classifier. Thus,

$$P\left(A_i \middle| C, A_j\right) = P(A_i|C), \quad \forall A_i, A_j, P(C) > 0 \tag{4.23}$$

The Naive Bayes classification algorithm can independently learn either the conditional probability of each feature A_i in the class C ($P(A_i|C)$), or the probability of each feature A_i. Replaced with a normalization factor "α", the posterior probability becomes

$$P(C = c|A_1 = a_1 \ldots A_n = a_n) = \alpha P(C = c) \prod_{i=1}^{n} P(A_i \mid C = c) \tag{4.24}$$

According to Equation 4.19, the optimal classification ($C = C_i$) should satisfy

$$P(C_i| < a_1 \ldots a_n >) = \frac{P(<a_1 \ldots a_n>|C_i)}{P(<a_1 \ldots a_n>)} P(C_i) \tag{4.25}$$

$$P(C_i| < a_1 \ldots a_n >) > P\left(C_j\middle| < a_1 \ldots a_n >\right),\ j \neq i \tag{4.26}$$

Based on the above analysis, the Naive Bayes classification process can be presented in Fig. 4.3.

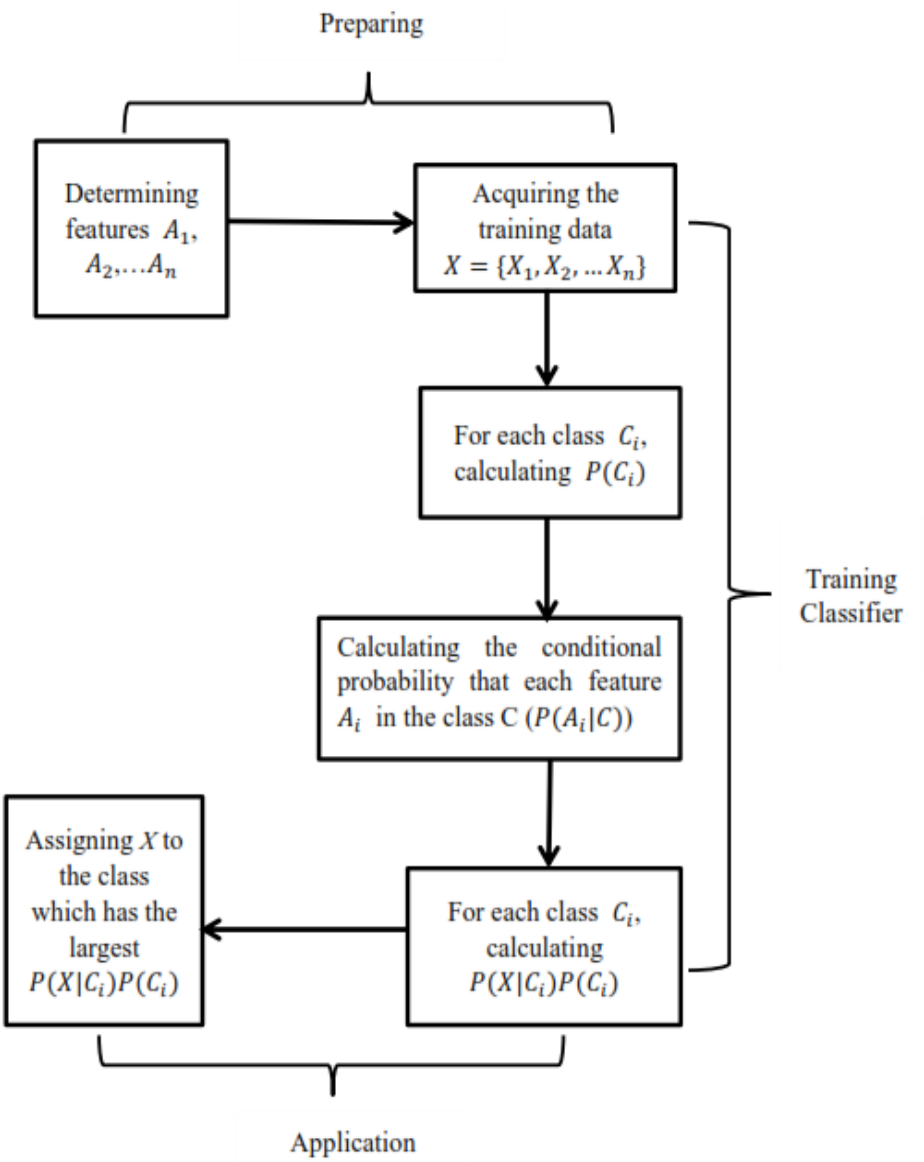


Figure 4.3: The Naive Bayes classification process

It can be seen from Fig. 4.3, the whole Naive Bayes classification process can be divided into three stages.

The first stage is preparing, which is mainly determining the characteristics of features, and allocating them appropriately into different classes. In this stage, the

unclassified data is set as an input, and the output is their features and training samples.

The second stage is training the Naive Bayes classifier. The main task is to calculate the prior probability of each class in the training data and each feature's conditional probability. The input is the features and the training data, the output is the trained classifier. This stage can be implemented by programming according to Equations 4.22-4.26.

The third stage is the application of the trained Naive Bayes classifier in testing data.

4.3 Voting Classification Scheme

Combining the above-mentioned classifiers (LDA, BP, NB), a voting classification scheme is further proposed for the mammogram analysis system in this research. Fig. 4.4 shows the voting classification scheme, where "1" represents cancerous mammograms from a classifier, and "0" represents normal mammograms. When classifying a mammogram image, the voting classification decision is made by taking opinions of the majority of the three classifiers.

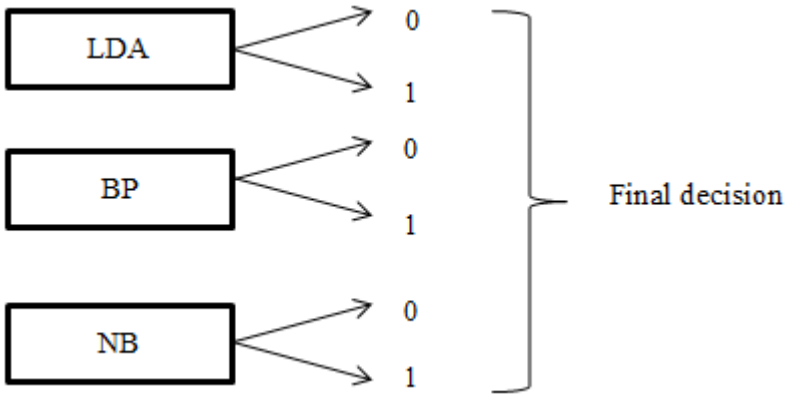


Figure 4.4: Voting classification scheme

4.4 Evaluation

The evaluation of the classifiers performance is essential in designing the proposed mammogram analysis system. In this paper, two evaluation methods were used.

The first method is based on the classification specificity of mammogram. In this binary classification problem, a classifier yields two results: positive and negative. According to the confusion matrix in machine learning, there could be four outcomes for the classifier in analyzing a sample. As shown in Fig. 4.5, the four results placed in a confusion matrix are true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Specifically, “TP” refers to a positive instance is classified correctly as positive; “FN” refers to the positive instance wrongly classified as negative. Similarly, “TN” refers to that a negative instance is correctly classified as negative; otherwise it is “FP”.

The classification specificity is defined as the proportion of instances deemed normal when breast cancer is absent. It is given in Equation 4.27. Besides, the classification accuracy and sensitivity, defined in Equation 4.28 and Equation 4.29, respectively, are also used to help the performance evaluation in this research.

		Actual Classes	
		Positive	**Negative**
Predicated Classes	**Positive**	**TP**	**FP**
	Negative	**FN**	**TN**

Figure 4.5: Confusion matrix

$$\text{Specificity } (SP): \ SP = \frac{TN}{TN+FP} \tag{4.27}$$

$$\text{Accuracy: } \ acc = \frac{TP+TN}{TP+TN+FP+FT} \tag{4.28}$$

$$\text{Sensitivity } (SN): \ SN = \frac{TP}{TP+FN} \tag{4.29}$$

The second method employs the Receiver Operator Curve (ROC) to facilitate comparison of different classifiers. More specifically, the area under the curve (AUC) can be used to simply and graphically evaluate the performance of classifiers [63, 64].

An ROC curve plots the X axis and Y axis using the false positive rate (FPr) and true positive rate (TPr) respectively, shown in Fig. 4.9. FPr and TPr can be calculated using Equations 4.30 and 4.31:

$$FPr = 1 - SP = \frac{FP}{TN+FP} \tag{4.30}$$

$$TPr = SN = \frac{TP}{TP+FN} \tag{4.31}$$

Generally, a good classifier would produce an ROC curve which is located closer to the upper left corner. The reason is straightforward: given the same false positive rate (FPr), a better classifier can obtain a larger true positive rate (TPr). In other words, the area under a better classifier curve is usually larger. For example, the classifier "a" outperforms the classifier "b" as shown in Fig. 4.6.

In general, the area under the ROC curve ranges between 0.5 and 1.0, and the closer to 1 the AUC is, the better the classifier is. The classifier achieves relatively low accuracy when the AUC is between 0.5 and 0.7; it has fairly good accuracy when the AUC is from 0.7 to 0.9, and excellent accuracy if the AUC is above 0.9. The classifier completely does not work if the AUC is equal to 0.5. The situation where the AUC is lower than 0.5 seldom appears in practice.

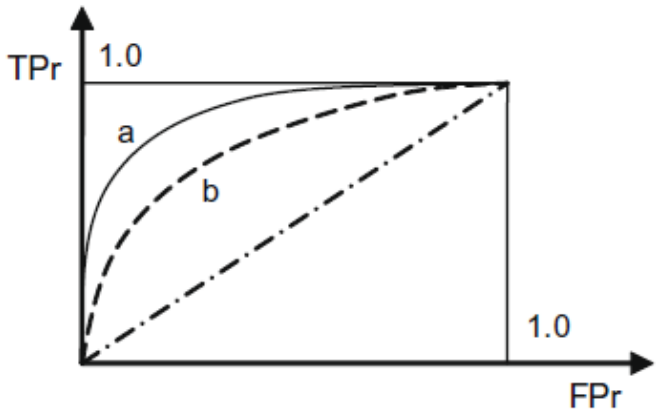


Figure 4.6: ROC curve for comparison between classifier a and b. (retrieved from Journal of Biomedical Informatics, www.elsevier.com/locate/yjbin, Jan., 2013)

Chapter 5 Results and Discussion

Features based on moments of the mean were computed from a training dataset containing 670 mammograms that had been normalized then Fourier and wavelet transformed (see Chapter 3 for details). Next an entropy based feature selection algorithm was applied to reduce the total number of features (See Chapter 4 for details). These features were passed through three unique classifiers and the results of that process will be discussed here. The performance of the classifiers was compared using a truth table and Receiver Operating Characteristic (ROC) curves. As a second test, the classifiers were tested in a testing dataset containing 817 normal mammograms with features extracted from the training dataset. The results of this experiment are also discussed.

5.1 Materials and Methods

5.1.1 Materials

The images to be analyzed in this work were a gift from the Eastern Health in Newfoundland and Labrador of Canada. One training dataset consisted of 670 mammogram images: 521 images of normal breasts and 149 images of malignant

breasts. These mammogram images were histologically confirmed by Newfoundland and Labrador breast screening program. The other testing dataset consisted of 817 normal mammogram images. Images from both datasets were all anonymous in the format of DICOM [61], which is a set of standard protocols in the medical image processing, storage, printing, and transmission. Their use was authorized by the Health Research Ethics Authority (HREA) in the reference number of 11312. All DICOM mammogram images were sampled to 1024×1024 pixels and reconfigured to PGM format.

5.1.2 Methods

The computer-aided mammogram analysis system is designed to process the original PGM images and automatically discriminate them as either normal or cancerous. As shown in Fig. 5.1, this system comprised three consecutive stages: the image processing, feature selection and image classification stage.

In the first image processing stage, a set of scalar features were extracted from an original image. This stage consisted of two steps: image pre-processing and data transform (including wavelet and Fourier transforms). In the image preprocessing step, the original digitized mammogram image is flipped, de-noised, and scaled to a common maximum value. In the data transform step, the normalized images are decomposed by three wavelet transforms with different bases (Daubechies db2, Daubechies db4, and Biorthogonal bior6.8) and the Fourier transform separately.

Multiple levels of decomposition were used, and four images are produced at each level of the decomposition. Finally, four statistic features, including the mean, standard deviation, skewness and kurtosis of the image intensities, were calculated.

In the second feature selection stage, the optimal features for the next stage were selected by an entropy-based feature selection algorithm, which reduces the number of features extracted from the transformed mammogram images. After calculating the information gain value of each feature, the features with high information gain value were selected.

In the final image classification stage, mammogram images were determined as either normal or cancerous based on the selected features. Three classifiers (Linear Discriminate Analysis, Back-propagation Network, and Naive Bayes classifier) were trained and tested. Furthermore, a combined voting classification scheme was proposed based on these classifiers.

In this experiment, the programs of the image processing and the classification were developed in Matlab 2010b, and the program of feature selection was developed in Eclipse using JAVA. The computer used was based on Windows 8.1, Intel dual core CPU i5-3210M @ 2.50GHz, 4GB RAM. It usually took 20 minutes to complete the process of feature extraction from 670 mammograms in Matlab. It took around 10 seconds to run the feature selection process in Eclipse. The final image classification process only required less than 5 seconds in Matlab.

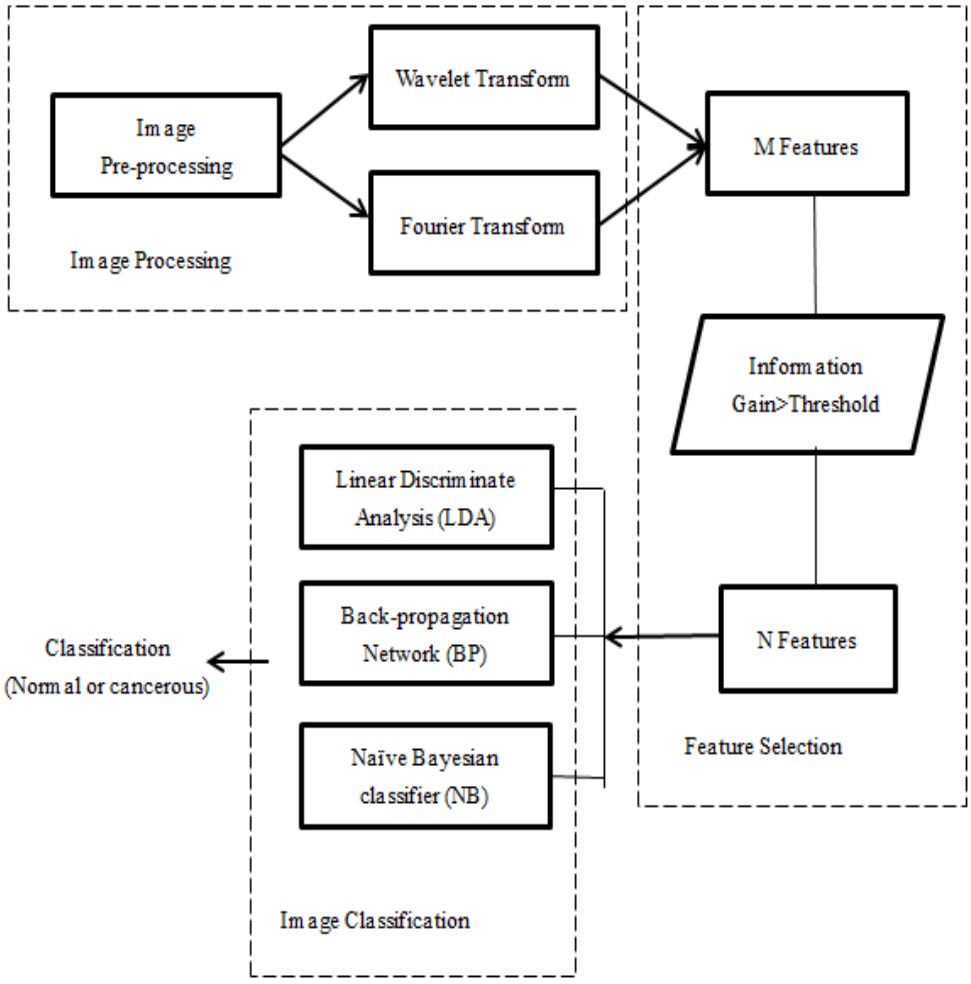


Figure 5.1: Block diagram of automatic mammogram analysis system

5.2 Feature Selection Results and Discussion

5.2.1 Results

When the aforementioned data transform methods in Chapter 3 were applied on the training dataset of 670 Mediolateral oblique (MLO) mammograms, 102 statistic features were extracted. Among the features, 6 features were extracted from the Fourier transform, and 96 features were calculated for each image from three kinds of

discrete wavelet transform (Daubechies db2, Daubechies db4 and the biorthogonal bior6.8 wavelets) separately.

The next step was to select the best features since including many features would almost certainly over specify the algorithm to the training dataset and this would impair its general application [60]. Therefore, the information gain ranking was used to select the best features and thus reduce the number of features enlisted for classification. As described in Chapter 4, Lei and Huan's research [58] found that 15 features could lead to the best classification accuracy. Therefore, 14 features in this research with information gain larger than 0.69 were empirically selected for the image classification after the usage of the sorting algorithm. If the information gain was set too low, the number of features would significantly increase, for example, the number of features was 23 when the information gain was set to 0.6. On the other hand, accurate classification results might not be obtained if the number of features is set too high. In the same way, the features selected from the db2 and Fourier transforms could be determined. In this case, information gain was set as 0.68 to keep consistent the number of features left with the db4 wavelet and Fourier transforms.

Using the aforementioned method, the features selected from the db4 wavelet and Fourier transforms are listed in Table 5.1.

This information gain value resulted in 15 statistic features, which are listed in Table 5.2.

Table 5.1: Information gain statistic for features calculated from the db4 wavelet and Fourier transform maps

Feature	IG
Level 3 kurtosis,h	0.74
Level 3 kurtosis,v	0.76
Level 3 kurtosis,d	0.74
Level 4 kurtosis,h	0.72
Level 4 kurtosis,v	0.75
Level 4 kurtosis,d	0.71
Level 5 kurtosis,h	0.71
Level 5 kurtosis,v	0.72
Level 5 kurtosis,d	0.69
Level 8 mean,a	0.70
Level 8 kurtosis,v	0.70
Fourier std	0.76
Fourier kurtosis	0.76
Fourier skewness	0.75

Table 5.2: Information gain statistic for features calculated from the db2 wavelet and Fourier transform maps

Feature	IG
Level 3 kurtosis,h	0.74
Level 3 kurtosis,v	0.74
Level 3 kurtosis,d	0.75
Level 4 kurtosis,h	0.71
Level 4 kurtosis,v	0.73
Level 4 kurtosis,d	0.73
Level 5 kurtosis,h	0.68
Level 5 kurtosis,v	0.71
Level 5 kurtosis,d	0.69
Level 7 mean,a	0.72
Level 8 mean,a	0.73
Level 8 std,a	0.69
Fourier std	0.76
Fourier kurtosis	0.76
Fourier skewness	0.75

Similarly, the features selected from the bior6.8 wavelet and Fourier transforms were obtained by setting the information gain value as 0.68, which are shown in Table 5.3.

Table 5.3: Information gain statistic for features calculated from the bior6.8 wavelet and Fourier transform maps

Feature	IG
Level 3 kurtosis,h	0.75
Level 3 kurtosis,v	0.76
Level 3 kurtosis,d	0.74
Level 4 kurtosis,h	0.73
Level 4 kurtosis,v	0.74
Level 4 kurtosis,d	0.71
Level 5 kurtosis,h	0.72
Level 5 kurtosis,v	0.72
Level 6 kurtosis,v	0.69
Level 7 mean,a	0.72
Level 8 mean,a	0.73
Level 8 std,a	0.71
Fourier std	0.76
Fourier kurtosis	0.76
Fourier skewness	0.75

In order to compare features from the three wavelet and Fourier transforms, these features were ranked in descending order according to their information gain. Table 5.4 shows the top 12 features with information gain higher than 0.74. The 12 features listed in Table 5.4 are named as the optimal features.

Table 5.4: Information gain statistic for the optimal features calculated from all wavelet and Fourier transform maps

Feature	IG
Fourier kurtosis	0.76
Db4 Level 3 kurtosis,v	0.76
Fourier std	0.76
Bior6.8 Level 3 kurtosis,v	0.76
Db2 Level 3 kurtosis,d	0.75
Db4 Level 4 kurtosis,v	0.75
Fourier skewness	0.75
Bior6.8 Level 3 kurtosis,h	0.75
Bior6.8 Level 3 kurtosis,d	0.74
Db2 Level 3 kurtosis,h	0.74
Db4 Level 3 kurtosis,h	0.74
Bior6.8 Level 4 kurtosis,v	0.74

5.2.2 Discussion

In the first three feature tables (Table 5.1, Table 5.2, and Table 5.3), features including standard diversion, kurtosis and skewness extracted from the Fourier transform are always obtained by the proposed entropy-based feature selection algorithm. It was also found that the kurtosis of the detail views (horizontal, vertical, diagonal) in level 3 and level 4, and the mean value of the approximate view of level 8 are obtained by the algorithm. Furthermore, the average information gain of features selected from the bior6.8 wavelet and Fourier transforms is larger than that of the db2

wavelet and Fourier transforms; and the information gain of features selected from the db4 wavelet and Fourier transforms is the highest.

In the selected features from the db4 wavelet and Fourier transforms listed in Table 5.1, kurtosis accounts for 78.6% (11 out of 14 features); it accounts for 66.7% (10 out of 15 features) in Table 5.2 (db2 wavelet and Fourier transform) as well as in Table 5.3 (bior6.8 wavelet and Fourier transform). Thus, kurtosis works best in the four features.

In regard to the optimal features of Table 5.4, three features of the Fourier transform, four features of the bior6.8 transform, three features of the db4 transform, and two features of the db2 wavelet transform were selected separately. Among those features, nine features are kurtosis, accounting for 75%. Additionally, features selected from the wavelet transform are mostly localized in the third level of wavelet decomposition (7 out of 9 features).

5.3 Image Classification Results and Discussion

5.3.1 Results

Three classifiers, the Linear Discriminate Analysis (LDA), the Back-propagation Network (BP), and the Naive Bayes Classifier (NB), were tested in the proposed automatic mammogram analysis system. Their performances were first trained and tested using the training dataset of 670 mammograms.

Table 5.5 shows the accuracy, sensitivity, and specificity of the three classifiers based on the above selected db2-Fourier, bior6.8-Fourier, db4-Fourier, and the optimal features.

Table 5.5: Classification performances of three classifiers for the training dataset

Classifier		LDA	BP	NB
db2-Fourier Features	Accuracy	80.69%	85.05%	81.03%
	Sensitivity	71.18%	83.05%	90.06%
	Specificity	89.03%	88.06%	70.08%
bior6.8-Fourier Features	Accuracy	81.01%	86.07%	83.03%
	Sensitivity	72.06%	84.65%	91.71%
	Specificity	90.32%	89.01%	72.15%
db4-Fourier Features	Accuracy	84.03%	89.06%	86.02%
	Sensitivity	74.24%	87.55%	93.20%
	Specificity	93.06%	92.05%	74.07%
The optimal Features	Accuracy	88.02%	94.14%	89.83%
	Sensitivity	78.26%	90.45%	96.21%
	Specificity	96.88%	95.03%	80.60%

Based on the obtained above performance results, the Receiver Operating Characteristics curves were plotted to facilitate comparison of the three classifiers (see Fig. 5.2). In this Figure, A denotes the LDA classifier, B denotes the BP classifier, and C denotes the Naive Bayes classifier.

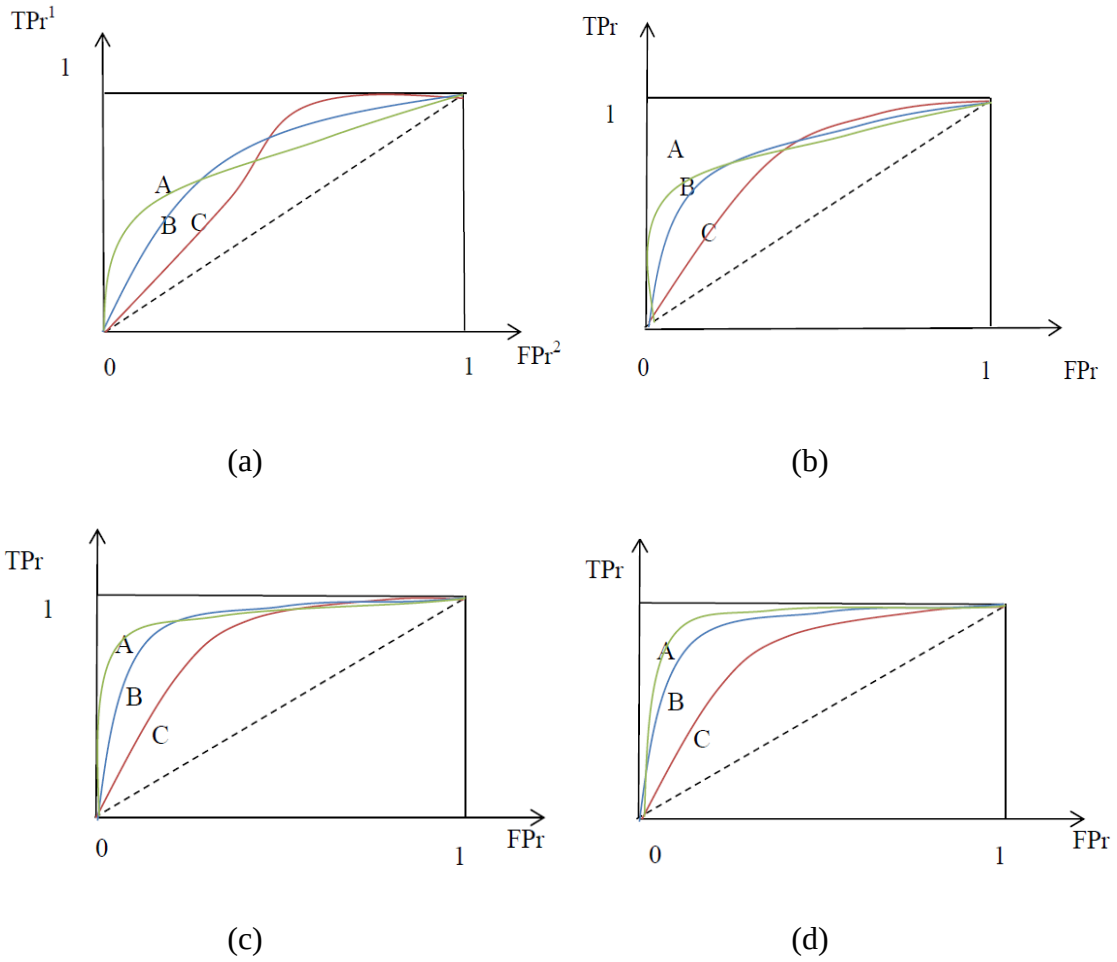


Figure 5.2: ROC curves with the classifiers: A. LDA; B. BP; and C. NB. (a), (b), (c), and (d): performances of classifiers based on db2-Fourier, bior6.8-Fourier, db4-Fourier, and the optimal features respectively.

[1] TPr: true positive rate;

[2] FPr: false positive rate.

According to the definition of specificity, it is the proportion of the cases deemed normal when breast cancer is absent. To further validate the performances of three classifiers and test if they were improved in specificity using the previous trained features, they were examined in the testing dataset, which consisted of 817 mammograms that were classified as normal by radiologists. When using features selected from the db2 and Fourier transforms, the LDA classifier suggested that 520

mammograms were normal out of the 817 mammograms, the BP network classified 410 normal mammograms, and the NB classifier showed that 325 were normal. Therefore, the specificity for these three classifiers for the testing dataset was 63.6%, 50.1%, 39.8%, respectively (see Table 5.6).

Table 5.6: Specificity of three classifiers for the testing dataset

	db2-Fourier	bior6.8-Fourier	db4-Fourier	the optimal features
LDA	63.6%	68.7%	70.1%	74.1%
BP	50.1%	55.5%	60.2%	64.2%
NB	39.8%	42.3%	47.8%	51.8%

Then, 817 mammogram images in the testing dataset were tested using the voting classification scheme introduced in chapter 4, which took majority vote of three classifiers. The results showed that there were respectively 520, 570, 594, and 690 normal mammograms out of 817 mammograms using selected db2-Fourier, bior6.8-Fourier, db4-Fourier and the optimal features (see Table 5.7).

Table 5.7: Specificity of different features using voting classification scheme

db2-Fourier	bior6.8-Fourier	db4-Fourier	The optimal features
63.6%	69.6%	72.7%	77.1%

5.3.2 Discussion

According to the results in Table 5.5, three classifiers achieved their highest classification performances using the optimal features, followed by the features selected from the db4 wavelet and Fourier transforms; whereas their performances are the lowest using features selected from the db2 wavelet and Fourier transforms. Specifically, Table 5.8 shows the detailed comparisons of accuracy, sensitivity, and specificity using the optimal features and other feature sets (the db2-Fourier, the db4-Fourier, and the bior6.8-Fourier feature sets).

Table 5.8: The performance increase of classifiers with the optimal features and separate feature sets

		Accuracy increase	Sensitivity increase	Specificity increase
optimal features VS db2-Fourier features	LDA	7.33%	9.09%	8.8%
	BP	7.08%	7.40%	6.15%
	NB	7.85%	6.97%	10.52%
optimal features VS bior6.8-Fourier features	LDA	7.01%	8.07%	6.80%
	BP	6.20%	5.80%	4.50%
	NB	6.56%	6.02%	8.45%
optimal features VS db4-Fourier features	LDA	3.99%	5.08%	3.81%
	BP	4.02%	2.90%	3.01%
	NB	3.82%	5.00%	6.53%

The reason that optimal features achieve the highest classification performance could be the information gain of the optimal features is the highest among the four different feature sets. In other words, the features in the optimal feature set are more correlated to the mammogram class than any of the other features.

It can also be found from Table 5.5 and ROC curves that the Naive Bayes classifier achieves the highest sensitivity using the default parameters in the proposed mammogram analysis system. On the other hand, the Linear Discriminate Analysis classifier achieves the highest specificity, and the Back Propagation network achieves the highest accuracy based on all the four feature sets. This result suggests that the NB classifier is more sensitive to classify cancerous mammograms, the LDA classifier gives better classification in normal mammograms, and the BP neural network works well in both of normal and cancerous mammograms.

From the classification results shown in Table 5.6, it could be found that the LDA classifier using the optimal features achieves the highest specificity. The specificity of the LDA classifier with the optimal features is 4.0% higher than that of the features of the db4-Fourier transform, 5.4% higher than that of the bior6.8-Fourier transform, and 10.5% higher than that of the db2-Fourier transform. The average specificity of the LDA classifier with the three feature methods (i.e., the db4-Fourier, db2-Fourier, and bior6.8-Fourier transform) is 12.2% higher than that of the BP classifier, and 24% higher than that of the NB classifier. These results are consistent with previous

theoretical analysis and experimental results based on the training dataset of 670 mammograms, though the performance for the testing dataset is not that satisfactory. This could be due to the lack of generality of the original training dataset.

In addition, in the examined testing dataset, 10 mammogram images were misclassified as cancerous by all classifiers. From the visual examination results, those mammograms contained a shadow of an unknown object, which probably led to the misclassification. 22 mammogram images were classified correctly as normal by the BP and NB classifiers, but they were classified as cancerous by the LDA classifier. This misclassification by the LDA classifier could be due to the small feature differences between subtle masses and some false-positive regions when trained in the training dataset, as the Linear Discriminate Analysis classifier does not work well in nonlinear classification.

After the comparison of Table 5.6 and Table 5.7, the proposed voting classification scheme works better than all three classifiers in specificity. Specifically, compared with the NB classifier, the voting classification scheme achieves 25.3% higher specificity with the optimal features, 24.9% higher specificity with the features of the db4-Fourier transform, 27.3% higher specificity with the features of the bior6.8-Fourier transform, and 24.8% higher specificity with the features of the db2-Fourier transform. Compared with the BP classifier, the voting classification scheme achieves 12.9%, 13.5%, 14.1%, and 13.5% higher specificity with four feature

sets, respectively. As to the LDA classifier, its specificity increases 3.0% with the optimal features, increases 2.6% with the db4-Fourier features, increases 0.9% with the bior6.8-Fourier features, and is the same with the db2-Fourier features.

Chapter 6 Conclusions and Future Work

6.1 Conclusions

Breast cancer has been the second leading cause of cancer-related death after lung cancer in women. X-ray mammography is a leading method for the early detection of breast cancer, which has effectively reduced the death rate since mid-1980s [3]. In this research, a computer-aided automatic mammogram analysis system, which consists of image processing, feature selection, and image classification, is proposed to improve the detection performances.

The image processing includes mammogram image pre-processing and data transforms (including wavelet and Fourier transforms). The pre-processing part was first developed for regularizing the appearance of normal and cancerous images with different orientations, background and intensity ranges. This was successfully done through masking and intensity normalization in tandem. All regularized mammogram images were then applied to wavelet and Fourier decompositions. Statistical features, including the mean, standard deviation, skewness, and kurtosis of the image intensities, were extracted from different decomposition levels. For each of the three wavelet bases (db2, db4, and bior6.8) used in this work, 96 features were extracted from level

3 to 8 of the wavelet decompositions. Adding 6 features from the Fourier transform, 102 potential features were obtained.

In order to select the most effective features for differentiating between normal and cancerous mammogram images, an entropy-based algorithm was employed to remove less significant features. This selection was achieved by sorting and selecting features with higher information gain values. The experimental results suggested that the information gain of features from the db4-Fourier transform was higher than that of features from the bior6.8-Fourier transform, and the information gain of features from the db2-Fourier transform was the lowest among the three feature sets. In the features from the db2, db4, bior6.8, and Fourier transforms, we selected the top 12 features (the optimal features) with their information gain values higher than 0.74.

Based on the four selected feature sets, three classifiers (NB, LDA, and BP) were first applied in the training dataset of 670 mammogram images. The highest sensitivity (96.21%) was achieved by the NB classifier, the highest specificity (96.88%) was achieved by the LDA classifier, and the highest accuracy (94.14%) was achieved by the BP network. All the best performances were achieved using the optimal features. ROC curves were then plotted to facilitate comparison of the three classifiers. The results showed that the NB classifier was more sensitive to classify cancerous mammogram images, the LDA classifier gave better classification in

normal mammograms, and the BP neural network worked well with both normal and cancerous mammogram images.

To further evaluate the performances of three classifiers and test if they could be improved in specificity using the previous feature sets, they were examined in the testing dataset of 817 normal mammogram images. The experimental results of each classifier and the proposed voting classification scheme showed that the LDA classifier using the optimal features achieved the highest specificity (74.1%), which was 10.5%, 5.4%, and 4.0% higher than using the db2-Fourier, bior6.8-Fourier, and db4-Fourier features, respectively. Similarly, the proposed voting classification scheme achieved the highest specificity (77.1%) using the optimal features. Compared with the LDA, BP, and NB classifier separately, the specificity of the voting classification scheme increased by 3.0%, 12.9%, and 25.3% using the optimal features.

In conclusion, the experiment on the training and testing datasets demonstrated that the proposed automatic mammogram analysis system could effectively improve the classification performances, especially using the voting classification scheme based on the selected optimal features.

6.2 Future Work

In this thesis, the proposed automatic mammogram analysis system was tested with three kinds of wavelet basis and three classifiers based on two datasets. However, more work in different aspects can be taken into consideration to deepen the research. For example, the evaluation of image signal-to-noise (SNR) can be further investigated regarding to its effect, consistency and sensitivity to the mammogram analysis system. In the future, more wavelet bases can be employed to see if they can extract valuable features with high information gain. At the same time, more classifiers can be investigated to compare their sensitivity, specificity and accuracy. In addition, more datasets of a larger volume of breast cancer cases can be used to train and evaluate the proposed analysis system. Moreover, the voting classification scheme achieves higher specificity compared with single classifier using any feature set. In the future, a sequential classification scheme could be taken into consideration to investigate its performance.

Bibliography

[1] National Cancer Institute of Canada (2004). *Canadian Cancer Statistics 2004.* Toronto, Canada.

[2] American Cancer Society (2014). *Global Cancer Facts & Figures*. Retrieved from http://www.cancer.org/acs/groups/content/@research/documents/webcontent/acspc-042151.pdf

[3] Canadian Cancer Society (2013). *Breast Cancer Statistics at a Glance*. Retrieved from:http://www.cancer.ca/en/cancer-information/cancer-type/breast/statistics/?region=nl

[4] R. L. Birdwell (2009). The preponderance of evidence supports computer-aided detection for screening mammography. *Radiology*,*253*(1), 9-15.

[5] S. Xia, W. Lv (2000). Research development of computer-aided detection in mammography. *Biomedical Engineering in Foreign Medical Science, 2000*(23), 1.

[6] S. Wu, et al. (1996). Application research of laser scanning microscopy for early diagnosis of tumors. *Proceedings of SPIE, 2887*, 190-192.

[7] E. Claridge, J.H. Ricbter (1994). Characterization of mammographic lessions. *Proc. 2rd Int. Workshop on Digital Mammography*, 241-250.

[8] J. Shi, et al. (2008). Characterization of mammographic masses based on level set segmentation with new image features and patient information. *Med. Phys. 35*(1), 280

[9] L. Wu (2000). Discuss of the value in early mammography detection and diagnosis. *Qi Lu Medicine Magazine, 9*(17), 3.

[10] X. Kang (2001). *Modern Medical Imaging Technology*. Tianjing: Tianjing Technology Translation Publication Company.

[11] C. M. Coulam (1981). *The Physical Basis of Medical Imaging*. New York: Appleton-Century-Crofts.

[12] A. Van Steen, R. Van Tiggelen (2007). Short history of mammography: a belgian perspective. *JBR–BTR, 90*(3), 151-153.

[13] Radiology Information organization (2012). Mammography. Retrieved from http://www.radiologyinfo.org/en/info.cfm?PG=mammo.

[14] F. Winsberg, M. Elkin, J. Macy, Jr., V. Bordaz, W. Weymouth (1967). Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis, *Radiology, 89*(2), 211–215. doi:10.1148/89.2.211.

[15] Physician Insurers Association of America: Breast Cancer Study (1995). Washington, DC.

[16] T.Balakumaran, Dr.ILA.Vennila, C.Gowri Shankar (2010). Detection of microcalcification in mammograms using wavelet transform and fuzzy shell clustering. *International Journal of Computer Science and Information Security. 7*(1),121-125.

[17] H.P. Chan, K. Doi, S. Galhptra, et al. (1998). Computer-aided detection of microcalcification in mammogram-methodology and preliminary clinical study. *Investigative Radiology. 23*(9), 664-671.

[18] J. Ren, Z. Wang (2013). Effective Classification of Microcalcification clusters using improved support vector machine with optimised decision making. *Image and Graphics (ICIG), Seventh International Conference.* 390-393.

[19] Mammographic Image Analysis Society (2005). *Mini-Mammography Database*. Retrieved from http://www.wiau.man.ac.uk/services/MIAS/MIASmini.html.

[20] M. Heath, et al. (2000). The digital database for screening mammography. *Proc. 5th International Workshop on Digital Mammography.*

[21] A. Kage, M. Elter, T. Wittenberg (2007). An evaluation and comparison of the performance of state of the art approaches for the detection of spiculated masses in mammograms. *29th Conf. IEEE Engineering in Medicine and Biology Society (EMBS)*, 3373-3376.

[22] W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillies, M. W. Riggs, M. L. Nipper (1994), Computer-aided mammographic screening for spiculated lesions. *Radiology 191*(2), 331-337.

[23] S. Liu, C.F. Babbs, E.J. Delp (2001). Multiresolution detection of spiculated lesions in digital mammograms. *IEEE Tran. Image Processing, 10*(6), 874-884.

[24] P. Conilione, D. Wang (2005). A comparative study on feature selection for E.coli promoter recognition. *Int. Journal of Information Technology, 11*, 54–66.

[25] A. Whitney (1971). A direct method of nonparametric measurement selection. *IEEE Tran. Computers, 20,* 1100-1103.

[26] I. Inza, P. Larranaga, R. Blanco, A.J. Cerrolaza (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif. Intell. Med. 31*(2), 91–103.

[27] A. Jain, D. Zongker (1997). Feature selection: evaluation, application, and small sample performance. *IEEE Trans. Pattern Anal. Mach. Intell. 19*(2), 153–158.

[28] P. Pudil, J. Novovičová, J. Kittler (1994). Floating search methods in feature selection. *Pattern Recognition Letters, 15*(11), 1119–1125.

[29] A. Kusiak (2001). Feature transformation methods in data mining. *IEEE Trans. Electron. Packag. Manuf. 24*(3), 214-221.

[30] Discrete Fourier Transform (DFT) (2004). Retrieved from http://home.eng.iastate.edu/~julied/classes/ee524/LectureNotes/l5.pdf .

[31] Q. Feng, Z. Yang (2000). *Practical Wavelet Analysis*. Xi'an: Xi'an electronic science and technology university press.

[32] I. Inza, P. Larranaga, R. Blanco, A.J. Cerrolaza (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif. Intell. Med. 31*, 91–103.

[33] Z. Chen, H. Guo (2005). *Wavelet and Discrete Transform Theory and Engineering Practice*. Beijing: Tsinghua university press.

[34] S. W. Smith (1997). The Scientist and Engineer's Guide to Digital Signal Processing. Chapter 8: The Discrete Fourier Transform. (Retrieved from http://www.dspguide.com/ch8/1.htm)

[35]Frequency Domain (2003). http://homepages.inf.ed.ac.uk/rbf/HIPR2/freqdom.htm

[36] Y. Quan, Q. Chen (2010). The characteristics of biorthogonal multivariate wavelets associated with a pair of biorthogonal scaling function vector. *2010 International Conference on Challenges in Environmental Science and Computer Engineering (CESCE). 1*, 262-365.

[37] Z. Xiao (2000). *Image Information Theory and Coding Technology*. Guangzhou: Sun yat-sen university press.

[38] G. Rong (2000). *Computer Image Processing*. Beijing: Tsinghua university press.

[39] Y. Zha (2004). Adaptive multiple threshold image denoising based on the wavelet transform. *Chinese Graphics Journal*, 567-570.

[40] S. Mallat (1988). A compact multiresolution representation: the wavelet model. *Proc. IEEE Workshop Comput.. Vision, Miami, FL.*.

[41] T. Liu, X. Zeng, J. Zeng (2006). *Introduction to Practical Wavelet Analysis*. Beijing: National defense industry press.

[42] Wavelet transforms on images (2003). Lecture notes. Retrieved from http://sundoc.bibliothek.uni-halle.de/diss-online/02/03H033/t4.pdf

[43] A. Skodras, C. Christopoulos, T. Ebrahimi (2001). The JPEG 2000 still image compression standard. *IEEE Signal Process. Mag.*, 36–58.

[44] B.E. Usevitch (2001). A tutorial on modern lossy wavelet image compression: Foundations of JPEG 2000. *IEEE Signal Process. Mag.*, 22–35.

[45] E. Visser, T. Lee, M. Otsuka (2001). Speech enhancement in a noisy car environment. *Proc. Intl. Conf. Independent Component Analysis & Blind Signal Separation, San Diego, CA*, 272–276.

[46] M.G.E. Schneiders (2001). *Wavelets in Control Engineering*. Master's thesis, Eindhoven University of Technology.

[47] M. Kumar, S. Pandit (2012). Wavelet Transform and Wavelet Based Numerical Methods: an Introduction. *International Journal of Nonlinear Science, 13*(3), 325-345

[48] Z. Bian (2000). *Pattern Recognition (second edition)*. Beijing: Tsinghua university press.

[49] Q. Huang (2007). The concepts, theory and applications to pattern recognition. Retrieved from courseware: www.glbook.cn , p15.

[50] L.G. Shapiro, G.C. Stockman (2002). *Computer Vision*. Prentice Hall. ISBN 0-13-030796-3.

[51] G.I. Salama, M.B. Abdelhalim, M.A. Zeid, (2012). Experimental comparison of classifiers for breast cancer diagnosis. *2012 Seventh International Conference on Computer Engineering & Systems (ICCES)*. 180-185

[52] A.R. Webb (2002). *Statistical Pattern Recognition*, 2nd Edition. John Wiley & Sons.

[53] L. Wei, Y. Yang, R.M. Nishikawa, and Y. Jiang (2005). A study on several machine learning methods for classification of malignant and benign clustered microcalcifications. *IEEE Trans. Medical Imaging. 24*(3).

[54] N. Otsu (1979). A threshold selection method from gray-level histograms. *IEEE Trans. Sys. Man. Cyber. 9* (1), 62–66.

[55]Matlab help files (2010), retrieved from http://www.mathworks.com/help/matlab/

[56] H. Liu, L. Yu (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. Knowl. Data Eng., 17*(4), 491-502.

[57] Y. Saeys, et al. (2007). Bioinformatics (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23* (19), 2507-2517.

[58] Y. Lei, L. Huan (2003). Feature selection for high-dimensional data: a fast correlation-based filter solution. *Proc. Twentieth International Conference on Machine Learning (ICML-2003).*

[59] N. Terki, N. Doghmane, A. Ouafi, Z. Baarir (2004). Study of effect of filters and decomposition level in wavelet image compression.

[60] D. Dave, J. Jiashun (2008). Higher criticism thresholding: optimal feature selection when useful features are rare and weak. Retrieved from http://statweb.stanford.edu/~donoho/Reports/2008/HCT-20080730.pdf

[61] Digital Imaging and Communications in Medicine (DICOM) (2005). DICOM Standards Committee, Working Group 6,1300 N. 17th Street Suite 1847,Rosslyn, Virginia 22209 USA

[62] E.J. Kendall, M. Barnett, and K. Chytyk-Praznik (2013). Automatic detection of anomalies in screening mammograms. *BMC Medical Imaging*, 13-43.

[63] E.J. Kendall, M.T. Flynn (2014). Automated breast image classification using features from its discrete cosine transform. *PLOS ONE, 9*(3), e91015. doi:10.1371/journal.pone.0091015

[64] C. Sang, J. Pu, B. Zheng (2009). Improving performance of computer-aided detection scheme by combining results from two machine learning classifiers. *Academic Radiology, 16*(3), 266-275.