An Investigation of Automatic Feature Extraction for Clustered Microcalcifications on Digital Mammograms

by

Aqilah Baseri Huddin

B. E. (Electrical & Electronic Engineering, First Class Honours)
The University of Adelaide, 2007

Thesis submitted for the degree of

Doctor of Philosophy

in

Electrical and Electronic Engineering
The University of Adelaide

2015

THE UNIVERSITY *of* ADELAIDE

Contents

Heading	Page
Contents	**iii**
Abstract	**ix**
Statement of Originality	**xi**
Acknowledgement	**xiii**
Thesis Conventions	**xv**
List of Publications	**xvii**
List of Figures	**xix**
List of Tables	**xxv**
Chapter 1. Introduction	**1**
1.1 Mammography	2
1.1.1 Abnormalities in Mammograms	4
1.1.2 Breast Cancer Screening by Radiologists	9
1.2 Computer-Aided Diagnostic System	10
1.3 Aims of Thesis	12
1.4 Contributions of Thesis	13
1.5 Organisation of Thesis	13

Chapter 2. Previous Work in CAD System for Mammograms **15**

2.1 CAD Detection of Microcalcification (CADe) 16

2.1.1 Pre-processing Stage 17

2.1.2 Segmentation 23

2.1.3 Feature Analysis 23

2.2 CAD Diagnosis of Microcalcification (CADx) 26

2.2.1 Feature Extraction 27

2.2.2 Classification 30

2.3 Neural Network in Biomedical Applications 33

2.4 Summary 35

Chapter 3. Mathematical Background **37**

3.1 Fourier Series and Fourier Transform 38

3.1.1 Short Time Fourier Transform 41

3.2 Multi-resolution Analysis 44

3.3 Wavelet Analysis 47

3.4 Steerable Pyramid Filtering Analysis 50

3.4.1 Steerable Filters 52

3.4.2 Steerable Pyramid Filtering 57

3.5 Summary 63

Chapter 4. Classifiers **65**

4.1 General Paradigm of Classification 66

4.2 Support Vector Machine 67

4.2.1 Linear SVM 68

4.2.2 Support Vectors and Optimizing Hyperplane 70

4.2.3 Non-Linear SVM Classification . . . 73
4.3 Neural Network . . . 74
4.3.1 Feed-Forward Neural Network . . . 75
4.3.2 Back-Propagation Neural Network . . . 77
4.4 Summary . . . 79

Chapter 5. Deep Belief Networks **81**
5.1 Deep Network . . . 82
5.2 Restricted Boltzmann Machine (RBM) . . . 84
5.3 DBN Architectures . . . 87
5.3.1 DBN Parameters Tuning . . . 89
5.4 Previous Work in DBN . . . 91
5.4.1 Autoencoder . . . 92
5.5 Summary . . . 96

Chapter 6. Feature Extraction Experiments **97**
6.1 Pre-processing and Data Acquisition . . . 98
6.1.1 Data Acquisition . . . 98
6.1.2 Image Format Conversion . . . 101
6.1.3 Segmentation of Region of Interest (ROI) . . . 102
6.1.4 Data Scaling . . . 104
6.1.5 Training and Testing Datasets . . . 105
6.1.6 System Evaluation . . . 106
6.2 Feature Extraction using Steerable Pyramid Filtering . . . 107
6.2.1 Experiment 1: Steerable Pyramid Topology Selection . . . 108
6.2.2 Experiment 2: Steerable Pyramid Feature Extraction with SVM Classifier for Microcalcification Classification . . . 114

6.2.3 Experiment 3: Data Dimensions Reduction using Principal Component Analysis (PCA) . . . 120
6.2.4 Experiment 4: PCA-SP Features with SVM Classifier . . . 124
6.3 Automatic Feature Extraction using DBN . . . 138
6.3.1 Experiment 5: DBN Architecture Selection . . . 140
6.3.2 Experiment 6: Feed-Forward DBN Feature Extraction for Microcalcification Classification . . . 148
6.3.3 Experiment 7: Microcalcification Classification using Multiple Orientation and Multiple Resolution DBN . . . 156
6.4 Summary . . . 163

Chapter 7. Conclusion **165**
7.1 Summary of Findings . . . 166
7.2 Suggestions for Future Work . . . 169
7.3 Concluding Remark . . . 170

Appendix A. Database **171**
A.1 Benign Microcalcification Clusters . . . 172
A.2 Malignant Microcalcification Clusters . . . 174

Appendix B. Detailed DBN experiment results **185**
B.1 Pre-Training Error in Layer One DBN . . . 186
B.2 Pre-Training Error in Layer Two DBN . . . 188
B.3 Pre-Training Error in Layer Three DBN . . . 190

Bibliography **193**

List of Acronyms **201**

Index **203**

Biography **205**

This page is blank.

Abstract

Mammography is a common imaging modality used for breast screening. The limitations in reading mammogram images manually by radiologists have motivated an interest to the use of computerised systems to aid the process. Computer-aided diagnosis (CAD) systems have been widely used to assist radiologists in making decision; either for detection, CADe, or for diagnosis, CADx, of the anomalies in mammograms. This thesis aims to improve the sensitivity of the CADx system by proposing novel feature extraction techniques. Previous works have shown that multiple resolution images provide useful information for classification. The wavelet transform is one of the techniques that is commonly used to produce multiple resolution images, and is used to extract features from the produced sub-images for classification of microcalcification clusters in mammograms. However, the fixed directionality produced by the transform limit the opportunity to extract further useful features that may contain information associated with the malignancy of the clusters. This has driven the thesis to experiment on multiple orientation and multiple resolution images for providing features for microcalcification classification purposes. Extensive and original experiments are conducted to seek whether the multiple orientation and multiple resolution analysis of microcalcification clusters features are useful for classification. Results show that the proposed method achieves an accuracy of 78.3%, and outperforms the conventional wavelet transform, which achieves an accuracy of 64.9%. A feature selection step using Principal Component Analysis (PCA) is employed to reduce the number of the features as well as the complexity of the system. The overall result shows that the accuracy of the system when 2-features from steerable pyramid filtering are used as input achieved 85.5% as opposed to 2-features from conventional wavelet transform, which achieves an accuracy of 69.9%. In addition, the effectiveness of the diagnosis system also depends on the classifier. Deep belief networks have demonstrated to be able to extract high-level of input representations. The ability of greedy learning in deep networks

provide a highly non-linear mapping of the input and the output. The advantage of DBN in being able to analyse complex patterns, in this thesis, is exploited for classification of microcalcification clusters into benign or malignant sets. An extensive research experiment is conducted to use DBN in extracting features for microcalcification classification. The experiment of using DBN solely as a feature extractor and classifier of raw pixel microcalcification images shows no significant improvement. Therefore, a novel technique using filtered images is proposed, so that a DBN will extract features from the filtered images. The analysis result shows an improvement in accuracy from 47.9% to 60.8% when the technique is applied. With these new findings, it may contribute to the identification of the microcalcification clusters in mammograms.

Statement of Originality

This work contains no material that has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published written by another person, except where due reference has been made in the text.

I give consent to this copy of the thesis, when deposited in the University Library, being available for loan, photocopying, and dissemination through the library digital thesis collection, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue, the Australasian Digital Thesis Program (ADTP) and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

8th December 2015

Signed

Date

This page is blank.

Acknowledgement

I would like to express my deepest gratitude and appreciation to my supervisors, Dr Brian Ng and Professor Derek Abbott for their guidance and support during the course of the work reported in this thesis. During the years I spent my living in Adelaide away from family, I received plenty of advice and motivation from my supervisors that encouraged me to keep going through a number of difficulties faced in living throughout the journey. Their contribution of valuable ideas and advice has inspired my enthusiasm to explore more in my study. Without them, the completion of this thesis would not have been possible.

I also would like to extend my gratitude to all academic and supporting staff at various centers in the University of Adelaide. I convey many thanks to the staff of the School of Electrical and Electronic Engineering for providing me a comfortable environment to stay in the department and the access to the use of various facilities during my study. Thank you for the warm reception by the International Student Centre (ISC), especially to Ms Jane Copeland and Mr Soufiane Rboub for their great concern in taking care of my candidature and also, to the Adelaide Graduate Centre (AGC) for the precious help in solving a number of administrative requirements for my study in Adelaide. My sincere thanks also go to the Ministry of Education Malaysia (MOHE) and the National University of Malaysia (UKM) for their financial support to this PhD study.

Special thanks to my family for their limitless support and sacrifices. To my father, Dr Hj Baseri Huddin and my mother, Hjh Zaharah, words cannot describe my gratitude for your prayers and support. Not to forget to my father-in-law, Hj Ibrahim, my mother-in-law, Hjh Jahani and all my siblings, thanks for your support. I also like to thank to all my friends, both in Malaysia and Australia.

Acknowledgement

I deliver very special words to my beloved husband, Mohd. Faisal; I thank you so much for always being by my side and for being unbelievingly supportive and understanding throughout our PhD journeys. The continuous encouragement that you gave, has kept me going striving to achieve this accomplishment. Last but not least, to my beautiful daughter, Aliya, you are the source of my strength and my unending joy and love. Thank you.

In all, I express Alhamdulillah.

Aqilah Baseri Huddin

Thesis Conventions

The following conventions have been adopted in this thesis:

1. **Notation.** The acronyms used in this thesis are defined in the List of Acronyms on page 201.

2. **Spelling.** Australian English spelling conventions have been used, as defined in the Macquarie English Dictionary (A. Delbridge (Ed.), Macquarie Library, North Ryde, NSW, Australia, 2001).

3. **Typesetting.** This document was compiled using LaTeX2e. TeXworks was used as text editor interfaced to LaTeX2e. Inkscape 0.91 was used to produce vector graphics of the figures.

4. **Mathematics.** MATLAB code was written using MATLAB Version R2009a.

5. **Referencing.** The Harvard style has been adopted for referencing.

6. **Punctuation.** The Oxford convention for commas has been used for punctuation.

This page is blank.

List of Publications

1. Baseri Huddin, A., Ng, B. W.-H., Abbott, D. (2011). Investigation of multiorientation and multiresolution features for microcalcification classification in mammograms. *Proceedings of the 7th International Conference on Intelligent Sensors, Sensor Networks and Information Processing* (ISSNIP 2011), Adelaide, Australia, December 6–9 2011: pp.52–57.

This page is blank.

List of Figures

Figure		Page
1.1	The mammograms images acquired from the MIAS database.	3
1.2	Mammograms views; mediolateral-oblique and cranio-caudal.	4
1.3	Types of masses based on their opacities.	5
1.4	Terminal ductal lobular unit (TDLU) in the breast tissue.	6
1.5	Types of calcifications.	7
1.6	Lobular microcalcification.	8
1.7	Ductal microcalcification.	8
1.8	Framework of a complete CAD system.	12
2.1	Flowchart of CAD detection system.	16
2.2	Matched Gaussian filter by Heinlein *et al.*	21
2.3	Flowchart of CAD diagnosis system.	26
3.1	Fourier series for square wave approximation.	39
3.2	Fourier transform of non-stationary sinusoid wave with different frequencies.	40
3.3	STFT frequency-time joint tiling.	43
3.4	Wavelet transform frequency-time joint tiling.	44
3.5	Nested vector spaces spanned by scaling function.	46
3.6	Image decomposition using 2D wavelet transform filterbanks.	50

3.7 Image decomposition using wavelet at two levels of resolutions. 51
3.8 Two-dimensional Gaussian function. 53
3.9 Derivative filters derived from Gaussian function. 55
3.10 Filtered images using filters derived from Gaussian function. 56
3.11 Image decomposition using steerable pyramids. 58
3.12 Decomposition subimages of the circle image using steerable pyramid filtering. 59
3.13 Basis filters: sp1Filters. 60
3.14 Basis filters: sp3Filters. 61
3.15 Basis filters: sp5Filters. 62

4.1 An illustrative 3-dimensional linear decision surface. 69
4.2 Hyperplanes. 70
4.3 Optimal hyperplane margin in SVM. 71
4.4 Non-linear SVM. 73
4.5 Neurons and neural network. 75
4.6 Three-layer feed-forward neural network. 76
4.7 Three-layer back-propagation neural network. 78

5.1 Deep belief network. 84
5.2 Restricted Boltzmann machine. 85
5.3 Gibbs sampling in Markov chain for learning in RBM. 87
5.4 A d-layer DBN. 88
5.5 A d-layer back-propagation DBN. 89
5.6 A d-layer associative memory DBN. 90

5.7 Random sample of MNIST handwritten digits. The images were generated from MNIST database (Lee *et al.* 1998). 93
5.8 Autoencoder DBN for MNIST digits reconstruction procedure. 93
5.9 The sum squared errors of pre-training in the autoencoder's layers. . . . 95
5.10 Original and reconstructed MNIST images. 95

6.1 Flow chart of data acquisition and segmentation process. 98
6.2 A sample of *overlay* file in DDSM database. 100
6.3 Ground truth marked by the radiologists on the mammogram in DDSM database. 102
6.4 Segmentation of ROI using the chain code. 103
6.5 Histogram of sizes of the segmented ROIs. 104
6.6 Boxplot of ROIs' sizes showing the median of 138112 pixels. 104
6.7 Images of the original, the reconstructed and the calculated loss after filtering is performed. 110
6.8 Signal to noise ratio (SNR). 110
6.9 Losses images from 2 randomly chosen ROIs. 111
6.10 Boxplot of SNR at 3-level of decomposition (image size 256×256). . . . 112
6.11 Boxplot of SNR at 4-level of decomposition (image size 256×256). . . . 112
6.12 Boxplot of SNR at 3-level of decomposition using sp3Filters (image size 128×128). 114
6.13 The topology of steerable pyramid. 115
6.14 An example of ROI containing microcalcification cluster. 115
6.15 Detailed subimages of microcalcification cluster produced from the steerable pyramid filtering. 116
6.16 Detailed subimages of microcalcification cluster produced from the wavelet transform filtering. 117

6.17 Boxplot of classifier accuracy with combination of energy and entropy as input features. . . . 118
6.18 Boxplot of classifier accuracy with energy as input features. . . . 118
6.19 Boxplot of classifier accuracy with entropy as input features. . . . 119
6.20 Eigenvalues plot obtained from PCA; combination of energy and entropy features. . . . 123
6.21 Eigenvalues plot obtained from PCA; energy features. . . . 124
6.22 Eigenvalues plot obtained from PCA; entropy features. . . . 124
6.23 Boxplot of classifier accuracy with PCA transformed energy features (image size 256×256). . . . 126
6.24 Boxplot of classifier accuracy with PCA transformed energy features (image size 128×128). . . . 127
6.25 Boxplot of classifier accuracy with PCA transformed entropy features (image size 256×256). . . . 129
6.26 Boxplot of classifier accuracy with PCA transformed entropy features (image size 128×128). . . . 130
6.27 Boxplot of classifier accuracy with 2 PCA transformed energy entropy features. . . . 131
6.28 Boxplot of classifier accuracy with 3 PCA transformed energy and entropy features. . . . 132
6.29 Boxplot of classifier accuracy with 4 PCA transformed energy and entropy features. . . . 134
6.30 Boxplot of classifier accuracy with 5 PCA transformed energy and entropy features. . . . 135
6.31 Comparison chart of classifier accuracy with different type of single input features. . . . 136
6.32 Comparison chart of classifier accuracy with different type of combination input features. . . . 137

6.33 Sum squared errors in pre-training the first layer RBM after 200 epochs. 143

6.34 Sum squared errors in pre-training the second layer RBM after 200 epochs. 145

6.35 Sum squared errors in pre-training the third layer RBM after 200 epochs. 146

6.36 The architecture of the DBN with 3 hidden layers. 148

6.37 ROC curve of feed-forward unsupervised DBN. 151

6.38 DBN as automatic feature extractor and unsupervised classifier. 151

6.39 DBN as feature extractor and SVM for classifier. 153

6.40 Flowchart of the DBN-SVM training and testing phase. 153

6.41 Boxplot of accuracy for feature extraction using DBN directly from the raw pixels of segmented ROI with SVM classfier. 155

6.42 Boxplot of accuracy for feature extraction using hybrid SP-DBN at resolution 1 with SVM classifier. 158

6.43 Boxplot of accuracy for feature extraction using hybrid SP-DBN at resolution 2 with SVM classifier. 158

6.44 Boxplot of accuracy for feature extraction using hybrid SP-DBN at resolution 3 with SVM classifier. 159

6.45 Multiple resolution and multiple orientation DBN with SVM classifier for microcalcification classification. 160

6.46 Boxplot of accuracy for feature extraction using hybrid SP-DBN at resolution 1, 2 and 3, with SVM classifier. 162

6.47 Comparison graph of accuracies achieved using different feature extraction approaches. 163

This page is blank.

List of Tables

Table		Page
1.1	Morphology and distributions features of microcalcification on mammograms.	8
3.1	Properties comparison between steerable pyramid with wavelet.	52
6.1	Number of cases for each type normal, benign and malignant mammogram in DDSM database.	99
6.2	Chain code values and their direction in x and y coordinate	100
6.3	Confusion matrix of the classifier's outcomes.	107
6.4	Mean SNR for three different sets of basis filters at 3 and 4 levels decomposition.	113
6.5	Comparison of mean and median accuracy for microcalcification diagnosis between features measured from steerable pyramid filtering and wavelet transform.	119
6.6	Statistical analysis: mean, median and standard deviation of accuracies obtained when using PCA energy features as input for microcalcification classification (image size 256×256).	128
6.7	Statistical analysis: mean, median and standard deviation of accuracies obtained when using PCA energy features as input for microcalcification classification (image size 128×128).	128
6.8	Statistical analysis: mean, median and standard deviation of accuracies obtained when using PCA entropy features as input for microcalcification classification (image size 256×256).	129

6.9 Statistical analysis: mean, median and standard deviation of accuracies obtained when using PCA entropy features as input for microcalcification classification (image size 128 × 128). 130
6.10 T-test analysis between steerable pyramid and wavelet transform, for image 256 × 256, with two PCA entropy and two PCA energy features. . 131
6.11 T-test analysis between steerable pyramid and wavelet transform, for image 128 × 128, with two PCA entropy and two PCA energy features. . 132
6.12 T-test analysis between steerable pyramid and wavelet transform, for image 256 × 256, with three PCA entropy and three PCA energy features. 133
6.13 T-test analysis between steerable pyramid and wavelet transform, for image 128 × 128, with three PCA entropy and three PCA energy features. 133
6.14 T-test analysis between steerable pyramid and wavelet transform, for image 256 × 256, with four PCA entropy and four PCA energy features. 133
6.15 T-test analysis between steerable pyramid and wavelet transform, for image 128 × 128, with four PCA entropy and four PCA energy features. 134
6.16 T-test analysis between steerable pyramid and wavelet transform, for image 256 × 256, with five PCA entropy and five PCA energy features. . 134
6.17 T-test analysis between steerable pyramid and wavelet transform, for image 128 × 128, with five PCA entropy and five PCA energy features. . 135
6.18 Pre-training error in the first layer up to epoch 100 in step of 5 epochs. . 143
6.19 Execution time for pre-training Layer 1 RBM after 200 epochs 144
6.20 Execution time for pre-training Layer 2 RBM after 200 epochs 145
6.21 Execution time for pre-training Layer 3 RBM after 200 epochs 146
6.22 Summarised results using different number of code layer nodes. 150
6.23 Summary results to compare the accuracy achieved with different input images to the DBN. 159
6.24 T-test analysis between different type of images at the input of the DBN features extractor. 162

Chapter 1

Introduction

BREAST cancer is a common threat to women's lives worldwide. The World Health Organization (WHO) has estimated 521 000 died from breast cancer in 2012 (*WHO Cancer Fact Sheet* 2014). The disease is known to cause death but with proper diagnosis at the early stage, the chance of survival increases dramatically. Although some risk factors such as gender, age, family history, genetics and lifestyle have been linked to breast cancer, the main general causes of breast cancer are still to be determined. It is likely that the incidences occur naturally. Early detection can lead to proper treatments before the cancer spreads and becomes much worse that can potentially be fatal. Thus, it is important to raise awareness among women to have an early breast detection by screening. The statistic of declining death rate in the United States since 1989 shows the positive impact by raising awareness among women regarding the disease.

1.1 Mammography

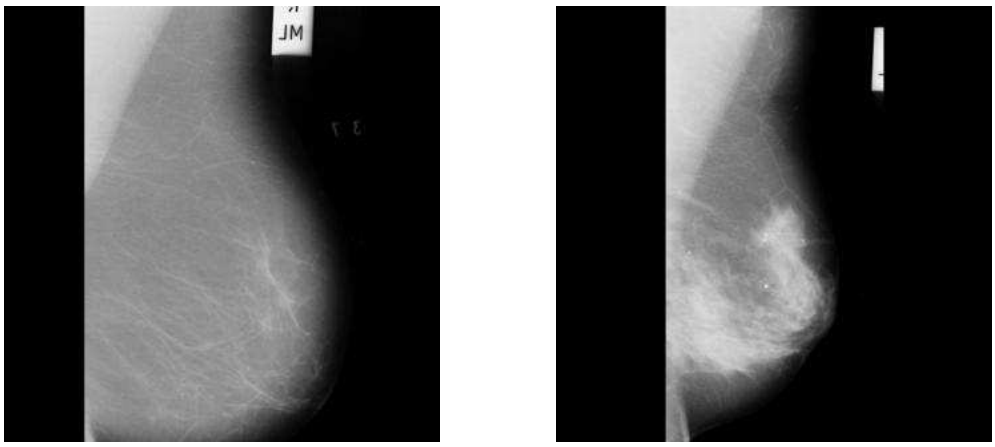
The main clinical technique for breast cancer diagnosis in current use is the use of mammography. Many programmes and campaigns have been introduced to raise awareness to undergo mammographic screening for the purpose of early detection. A study has shown that mammographic screening can help to reduce the number of fatalities and improve the success rate of treatments by 30–70% (Rangayyan *et al.* 2007). Australian Institute of Health and Welfare also reported that women undergone the diagnosis has 89% chance of surviving in 5 years (*Breast Screen Australia Monitoring Report 2009-2010* 2012).

Mammography is a low amplitude X-ray procedure to examine human breast tissue. The mammography is classified into two according to the types of mammograms it produces; standard mammograms and the digitally acquired mammograms (Tang *et al.* 2009b). The mammography was first invented in late 1960s which used standard radiographic X-ray systems. Here, the black and white image of breast tissue is printed on a photographic film. The major drawback of this system is the poor quality of the acquired images.

Digital mammography was then introduced, that produced images with improved dynamic ranges. With the digitized system, the images are recorded digitally on the computer and advanced imaging techniques can be applied to the images. Recently, the demand to use a full-field digital mammography (FFDM) is increasing. FFDM uses an improved image detector technology that improves the detection in the denser breasts compared to the conventional mammography. The performance of mammography is measured by its sensitivity; the percentage of the machine that correctly diagnoses positive cases. Current mammography technology is reported to have a sensitivity of 76% (Bevers *et al.* 2009). However the sensitivity varies depending on the breast density. In a denser breast, it is reported to have as low as 48% of sensitivity (Robson 2010).

The result from the procedure is used to detect any abnormality in breast tissue (*The Radiology Information Resource For Patients: Mammography* 2011). This process is

also called a screening process. During the process, an image of breast tissue is created using doses of ionizing radiation. Figure 1.1 shows example of a normal and an abnormal mammograms images acquired from the Mammographic Image Analysis Society Digital Mammogram Database (MIAS).



(a) (b)

Figure 1.1: The mammograms images acquired from the MIAS database; (a) normal breast tissue, and (b) an abnormality in breast tissue that has been diagnosed to be cancerous. *Image source: J. Suckling et al., The Mammographic Image Analysis Society Digital Mammogram Database, 1994*

Normally, two views of a breast will be captured separately, mediolateral-oblique (MLO) and cranio-caudal (CC) views as shown in Figure 1.2(a) and Figure 1.2(b), respectively. The MLO view captures the side-to-side, whilst the CC view captures the top-to-down of breast tissue.

Mammography has become a useful breast screening technique in medical practice. This technique has contributed to a reduction in breast cancer fatalities among women. Mainly because the early symptoms of breast cancer are not noticeable, until the tumor becomes larger and palpable. Thus the early screening by mammography plays an important role in screening and detecting any abnormality in human breast tissue. The Impact of BreastScreen SA reported that, the chances of diagnosed patients surviving an additional of 5 years is 98.4% and 10 years is 95.3% when the sizes of tumor detected is less than 10 mm. In contrast, when the tumor detected is larger than 40 mm, the rate of survival decreases dramatically to only 67.0% for 5 years and 54.4% for 10 years.

(a) (b)

Figure 1.2: Two views of mammograms acquired from DDSM database of normal tissue, volume Normal_01 and Case A_0002; (a) mediolateral-oblique (MLO) view, and (b) cranio-caudal (CC) view. MThe MLO view captures the side-to-side, whilst the CC view captures the top-to-down of breast tissue.
Image source: Breast - Heath M. et al., The Digital Database for Screening Mammography, 2001

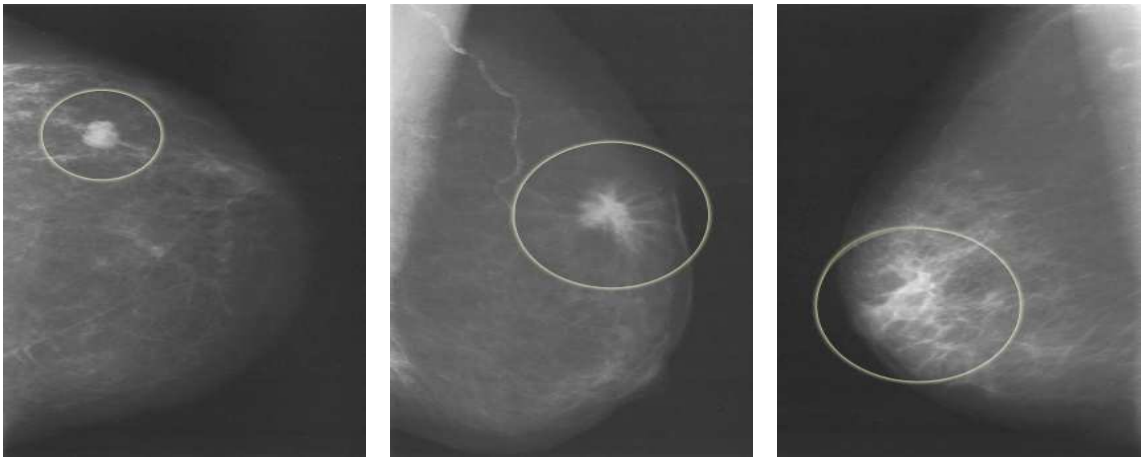
This shows that an early detection of these cancerous tissues is crucial for breast cancer treatment to be successful.

1.1.1 Abnormalities in Mammograms

There are several types of abnormalities found on mammograms by the radiologists. These abnormalities include asymmetric breast tissue and density, architectural distortions and appearance of masses or microcalcification clusters. A brief description of masses and microcalcification will be discussed in the following sections. The masses and microcalcification commonly can be categorised as either benign or malignant. A benign abnormality is normally a non-cancerous tumour, whilst a malignant is a cancerous tumour.

Masses

Masses or lumps are growths of tissue in breast. However, not all masses in breast tissue are malignant or cancerous. Masses are classified as either circumscribed, stellate or diffused. A circumscribed mass has well-defined morphology. It has curved shape, either round or oval. A stellate mass has irregular margins and a diffuse mass has a very poorly defined margin.



(a) (b) (c)

Figure 1.3: Types of masses based on their opacities; (a) circumscribed opacity, (b) stellate opacity, and (c) diffuse opacity.
Image source: Mammograpy, Guide to interpreting, reporting and auditing mammographic images, Lattanzio and Simonetti, 2010

Commonly, masses that have smooth, round margins are associated with benign masses and masses that has poorly defined margin are associated with malignant masses. Figure 1.3 shows three different type of masses commonly found in mammograms. Figure 1.3(a) shows mass with oval shape and is classified as circumscribed mass. Typically this type of mass is associated with benign mass. Figure 1.3(b) shows the stellate type of mass that has irregular margin. Stellate type of mass can be either benign or malignant depending on their other diagnosis features. Figure 1.3(c) shows a mass with poorly defined margin and commonly associated with malignant tumour.

Microcalcification

Microcalcification are tiny deposits of calcium that are formed in breast tissues. There are two types of calcifications based on their origins. The calcifications can originate from either within the ducts or within the lobules (Lattanzio and Simonetti 2010). Briefly, the terminal ductal lobular unit (TDLU) is illustrated in Figure 1.4, adapted from Smithuis and Pijnappel (2008).

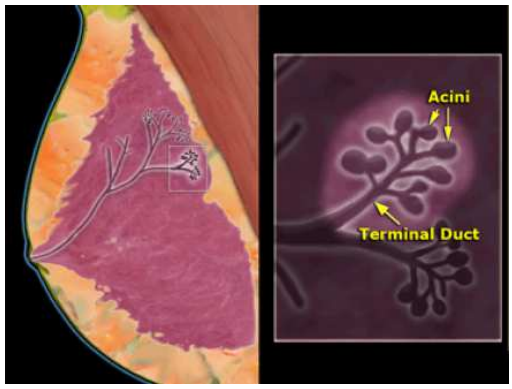
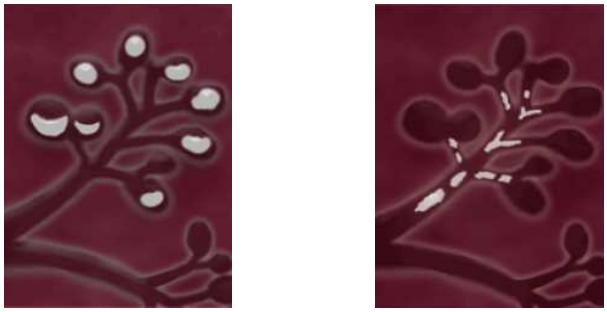


Figure 1.4: Terminal ductal lobular unit (TDLU) in the breast tissue. In TDLU, acinis are connected to the nipple through terminal duct.
Image source: Breast - Calcification Differential Diagnosis, Robin Smithuis and Ruud Pijnappel, 2008

Calcifications that originate within the lobules are mostly diagnosed as benign. These calcifications fill in the acini and often result in acini dilation. The shapes of the calcifications in this case are often uniform, homogeneous and sharply outlined. Calcifications that originate within the ductal often has irregular contours. This type of calcifications vary significantly in size and shape. Commonly they are diagnosed as being highly suspicious of malignancy. The differences between these two calcifications can be seen in Figure 1.5.

On mammograms, microcalcification are seen as bright spots. This is due to the higher X-rays absorption by the calcifications than by the tissues surrounding them (Spiesberger 1979). Often, the microcalcification on the mammogram appear in a group–at least 5 microcalcification per square centimeter (Nishikawa 2002). The sizes



(a) (b)

Figure 1.5: Two types of calcifications originate from; (a) lobular, and (b) ductal.
Image source: Breast - Calcifications Differential Diagnosis, Robin Smithuis and Ruud Pijnappel, 2008

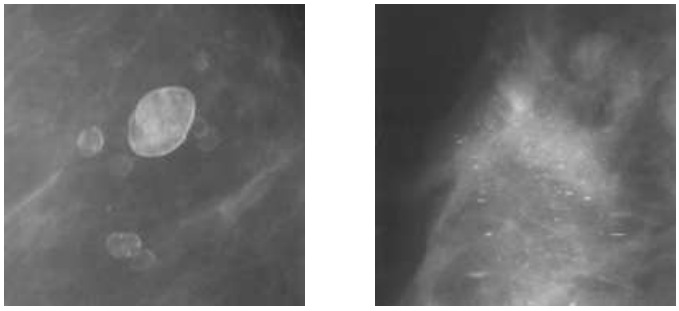
of individual microcalcification that form diagnostic images so far recorded by radiologists vary greatly from about 10 μm to up to a few millimeters. By examining the mammograms, the shapes of microcalcification can also be observed. Their shapes vary from spherical to elongated forms.

Clusters of microcalcification caused more clinical concern than individual isolated coarser calcification. An individual isolated calcification, also known as a macrocalcification, is normally caused by the degenerative changes in breast tissue, such as the aging of breast arteries, the inflammation in breast tissue or former breast injury. The macrocalcification normally does not multiply and therefore it is less likely to be malignant. For the microcalcification which exist in clustered form, they might be associated with different diagnosis results, ie. either benign or malignant.

Similarly with masses, not all detected microcalcification are malignant. Features such as shape, size, number of calcifications and their distribution are the common features used in identifying their natures. Table 1.1 summarises the above mentioned features and their corresponding nature ie. most probably benign, intermediate or probably malignant. Most probably benign has defined and regular shapes, such as round and rod-shapes. On the other hand, most probably malignant calcifications have irregular shapes and have casting or granular shapes. Figure 1.6 shows two samples of benign cases, where the calcifications originate from lobes. Whilst, Figure 1.7 shows two samples of malignant cases, where the calcifications originate from the ducts.

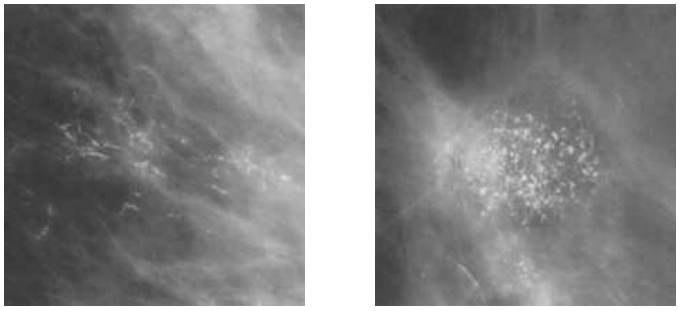
	Benign	Immediate Concern	Malignant
Morphology	Round Round lucent-centered Rod-shaped or tubular Coarse Skin "milk-of-calcium" "egg-shell"	Powdery Rounded Granular type	Casting Granular Mixed
Distributions	Diffuse regional	Clustered	Linear segmental

Table 1.1: Morphology and distributions features of 3 different cases; probably benign, intermediate or immediate concern and probably malignant.



(a) (b)

Figure 1.6: Two types of calcification originate from the lobe; (a) egg-shell, and (b) milk of calcium, which has been diagnosed as benign.
Image source: Mammograpy, Guide to interpreting, reporting and auditing mammographic images, Lattanzio and Simonetti, 2010



(a) (b)

Figure 1.7: Two types of calcification originate from the duct; (a) "casting" type, and (b) granular, which has been diagnosed as malignant.
Image source: Mammograpy, Guide to interpreting, reporting and auditing mammographic images, Lattanzio and Simonetti, 2010

The detection of an early sign of breast cancer is important to reduce the fatalities in women caused by breast cancer. The detection and diagnosis of microcalcification

in mammograms are crucial tasks. The challenge is to detect microcalcification that vary in size and shape. Due to their variable nature, it is often difficult to identify and distinguish the class of microcalcification even for trained specialists. Thus, it has become an interest in this thesis to implement a novel computer-aided diagnostic (CAD) system to improve mammogram diagnostic specifically on microcalcification.

1.1.2 Breast Cancer Screening by Radiologists

The images resulting from screening procedure will be interpreted by the experts or radiologists for further diagnostic purposes. The screening process confronting radiologists is very challenging. A study in 2002 has found that in 1000 screening cases, only 5 cases were detected to have breast cancer developed in the tissue (Rita E. Sohlich and Dee 2002). More than 50,000 screening data collected between 1997 and 2001 were used in their study. The task of finding a small number of cases from a large database is challenging and induces fatigue in the radiologists that can lead to an unacceptably high rate of false-negatives. In contrast, to increase the efficiency and reduce the false-negative rate, double reading has been advocated. Here, two radiologists will read the same mammograms. Warren and Duffy showed that this approach improves the sensitivity and effectiveness of the screening (Warren and Duffy 1995). However, this approach drastically increases the workload and is not cost effective.

Due to the manual nature of the task, low dynamic range and low contrast of film, together with the fuzziness shapes and sizes of microcalcification appeared on mammograms, radiologists often make inaccurate diagnosis of microcalcification clusters. Any subtle calcification in the mammograms are often imperceptible to human eyes. It has been reported that the mammograms can only display about 3% of the information they detected at normal view (Laine *et al.* 1995a). This contributes to the statistics of 10% to 30% of women being misdiagnosed and reported to be safe but actually they have breast cancer (Chen and Lee 1997, Ren *et al.* 2011). A recent work also has

showed that a double reading technique has a sensitivity of 88% compared to the detection technique with CAD assistance that achieved a sensitivity of 90.4% (Gromet 2008). This work has showed that CAD system has improved the sensitivity of detection on mammograms, and also reducing the cost in advocating two radiologists for double reading technique.

1.2 Computer-Aided Diagnostic System

Although there are highly trained and skilled radiologists, the incidences of human errors are unavoidable and may contribute to the misdiagnosis. Factors such as fatigue, boredom, and the changeable working environment may possibly indirectly affect the accuracy of the analysis. These factors are magnified if the task involves thousands of mammograms in the screening.

Historically, the first study to use computer for analysing mammograms was by Winsberg in 1967 (Winsberg *et al.* 1967, Nishikawa 2010). In their work, the mammograms films were digitised using facsimile machine with pixel size of 0.14 mm and produced image with grayscale of 32 levels. The digitised image was divided into 64 square blocks. These blocks were aligned and normalised in a standardised manner, so that each block was located at the same position of breast tissue for both breasts. This was to ensure the comparison between the two breasts are carried out consistently. Each of the square blocks were then measured for 4 different density features. Then, these features were compared between two breasts. This technique was rather primitive with the use of an unsophisticated digital scanner that allows 32 levels of grayscale as opposed to 256 or more levels of grayscale in more recent technology (Winsberg *et al.* 1967).

In 1987, Doi *et al.* proposed the concept of CAD system to assist radiologists in making decision. This technique differs from the earlier attempts because the CAD system diagnoses mammograms automatically. The findings by Getty *et al.* (1988) and Chan *et al.* (1990) were the major breakthroughs and initiated many new researchers to

develop CAD algorithm in mammogram diagnosis (Getty *et al.* 1988, Chan *et al.* 1990, Nishikawa 2007).

Getty *et al.* (1988) proposed a checklist from the radiologists as an input to the CAD system for malignancy classification. The checklist catalogs characterisation features of a lesion that the radiologists use to classify the cases into benign or malignant. Hence, the proposed CAD system has assisted radiologists to predict the malignancy of the lesion (Getty *et al.* 1988). On the other hand, Chan *et al.* (1990) introduced CAD system to detect clusters of microcalcification of mammogram. The study showed that with CAD assistance, many additional clusters were found on the mammograms (Chan *et al.* 1990). With these two earlier studies, many algorithms were developed by other researchers to improve the performance of the CAD system. Furthermore, with the emerging technology and an expanding knowledge in digital image processing, CAD has been proven to be able to contribute as an assistance to the radiologists in diagnosing breast cancer (Freer and Ulissey 2001).

The common approaches in CAD system are divided into two categories; CAD detection (CADe) and CAD diagnosis (CADx). The aim of CADe is to detect or identify any suspicious microcalcification revealed from the image in the mammograms. While the CADx aims to assist radiologists in the characterisation of the detected microcalcification clusters, which may be classified as benign or malignant. An overview framework of a CAD system in breast cancer diagnosis is illustrated in Figure 1.8.

Both the detection and diagnosis by CAD systems are not aimed to take over the role and the expertise of the radiologists but CAD systems can potentially provide a complimentary effort in resolving some of the medical challenges related to breast cancer diagnostic issues. As quoted by Nishikawa 2002, CAD systems are aimed to act as a second opinion to the radiologists (Nishikawa 2002).

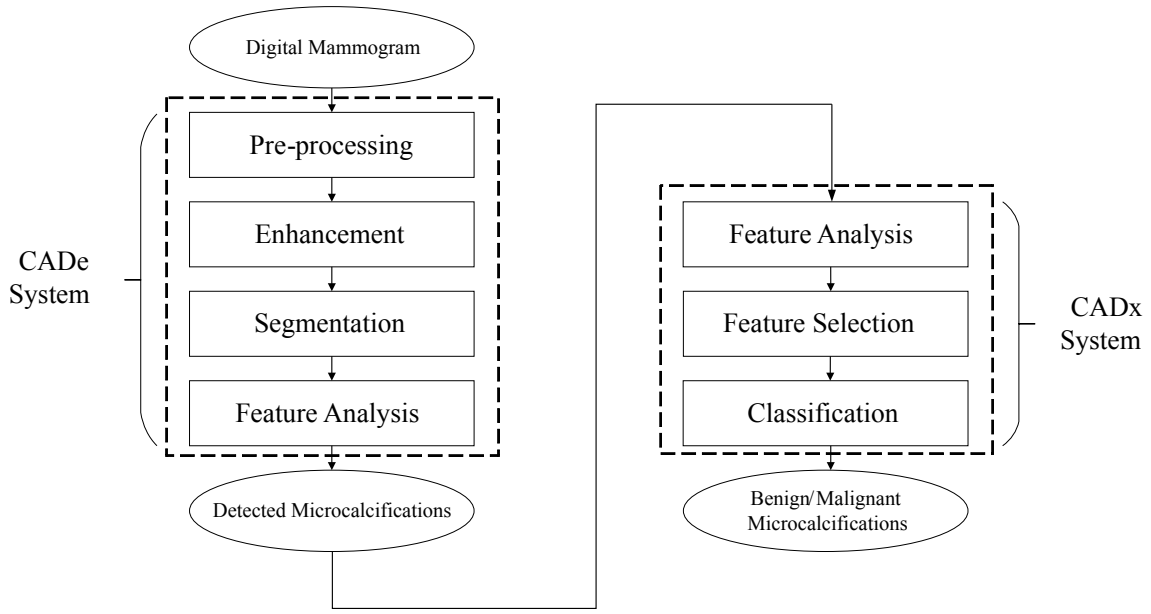


Figure 1.8: Overview framework of a complete CAD system in mammograms diagnosis.

1.3 Aims of Thesis

The main focus in this research is to implement a novel approach in improving CADx system, which is a part of the CAD system. As discussed earlier, CADx is designed to be able to recognise microcalcification clusters, hence the system is able to classify the microcalcification into benign and malignant classes. A study has reported that only 34% from the total microcalcification cases involved are actually malignant and required further biopsy (Knutzen and Gisvold 1993). By conventional practices, the radiologists have to screen all cases regardless benign or malignant classes. However, the recognition of microcalcification classes by CADx system ought to help the radiologists in classifying them. Hence, the improvement of CADx system is essential in improving the diagnosis of breast cancer.

Common steps involved in CADx include feature extraction and classification. In feature extraction, the aim is to extract dense features from the large images that will represent the pattern or characteristic of the detected clusters. These features necessarily contain useful information to be used next in classification steps. Commonly, there will be a features selection step in between feature extraction and classification steps.

The aim is to reduce the number of features into an acceptable number, and to reduce the complexity of the system. In classification step, the classifier will learn the pattern of these known samples for both benign and malignant cases. The system now has a supervised training pattern that can be used to classify the unknown sample based on the features extracted.

1.4 Contributions of Thesis

The contributions and findings of the thesis can be summarised as follows:

1. A multiple resolution and multiple orientation approach to extract features for microcalcification clusters classification. The work shows an improvement in classifying microcalcification into benign and malignant classes when multiple resolution and multiple orientation features were used.

2. An investigation on new techniques in deep belief networks for fully and semi automated feature extraction for mammogram diagnosis. Deep belief networks have proven to be able to extract features for classification. However, the network appears to face difficulties in extracting features directly from the raw images of microcalcification in mammograms. Hence, a novel technique such as in the next point is implemented for improvement.

3. Development of a multiple resolution and multiple orientation deep network to improve feature extraction of microcalcification clusters in mammograms. With this approach, the accuracy of the diagnosis is shown to be significantly improved.

1.5 Organisation of Thesis

This thesis consists of seven chapters including this introductory chapter as Chapter 1. Chapter 2 reviews the previous work in mammogram diagnosis that leads to the aim of

the thesis. Chapter 3 discusses the mathematical background of the filters commonly used in feature extraction and also introduces the steerable pyramid filtering that will be used to extract multiple resolution and multiple orientation features of microcalcification clusters. In Chapter 4, types of classifiers are reviewed. The new technique based on a deep belief network (DBN) is introduced in Chapter 5. The experiments conducted and the results of proposed approach of feature extraction to classify microcalcification in mammogram is presented in Chapter 6. Finally, conclusion and suggestions for future work are summarised in Chapter 7.

Chapter 2

Previous Work in CAD System for Mammograms

THIS chapter extensively reviews several previous work in CAD system for mammograms diagnosis. Due to the broad area of the field, it is customary to categorise the existing work into 2 streams of work, CAD detection (CADe) and CAD diagnosis (CADx). As its name suggests, CADe is for the detection purposes, whilst CADx is for the diagnosis or malignancy analysis purposes.

The following sections will discuss the chronological research development of each system, CADe in Section 2.1 and CADx in Section 2.2 for breast cancer diagnosis, specifically in detection and diagnosis of clusters of microcalcification in mammograms. Section 2.3 briefly reviews few work in biomedical applications that use neural network as part of their CAD system.

2.1 CAD Detection of Microcalcification (CADe)

The first stage in CAD system for mammography diagnosis is called detection or CADe. The aim of this stage is to detect or identify the appearances of microcalcification in mammograms. It does not classify the microcalcification as benign or malignant. The decision is made at a latter stage. The CADe acts as second reader to radiologists to spot any appearance that might be constituted with microcalcification. Some may called the system as screening tools (Nishikawa 2010). Work in automated microcalcification detection using computer analysis has attracted researchers from as early as in 1970s (Spiesberger 1979). Now, there are probably more than a hundred research groups taking part in developing method for automated clustered microcalcification detection.

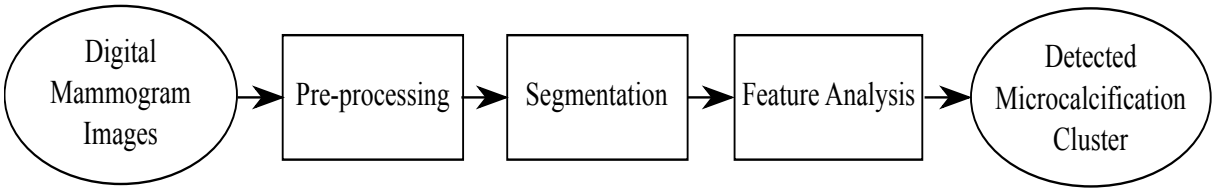


Figure 2.1: Flowchart of a CADe system commonly used for detection of microcalcification clusters. The system starts with the image acquisition from the mammograms. Typically the images are pre-processed to remove noise, suppress background and enhance the features on the mammograms. Next, the suspicious areas or features are segmented to reduce the computational cost. Features in the segmented area are extracted and analysed. The analysis is performed to reduce the false alarm by identifying the suspicious features as either microcalcification clusters or artifacts on the mammograms.

Most of the approaches in CADe systems follow the same scheme, as illustrated in Figure 2.1. After acquiring a digital mammogram, the first step to be taken is the pre-processing of the image. In this stage, the purpose is to make subtle microcalcification more visible by reducing the effects of normal tissues that act as camouflaging background to the clusters of microcalcification. The techniques include contrast enhancement, features enhancement and background noise reduction. The next stage is to segment an area in mammogram, which potentially has the microcalcification clusters in it. In the features analysis, the potential area that has been segmented is analysed

to reduce the false alarm by identifying the suspicious features as either microcalcification clusters or artifacts on the mammograms. At the final stage of the system, the area with clustered microcalcification is detected.

2.1.1 Pre-processing Stage

Different approaches have been proposed in the pre-processing stage of CADe, either by the enhancement of microcalcification on mammogram or the reduction of the background. Some of the work may integrate both approaches for the pre-processing purposes. The next subsections briefly discuss each method for pre-processing stage in the CADe system.

Enhancement of Microcalcification

The idea of microcalcification enhancement is to make their small structures to be more visible and thus can either help the radiologists to make decision or can be further analysed at the next stage in the CADe system. In previous work, some of the enhancement techniques in this area include the conventional enhancement techniques (Cheng *et al.* 2003), the region-based techniques (Chan *et al.* 1990, Morrow *et al.* 1992, Cheng *et al.* 2003) and the multi-scale-based techniques (Laine *et al.* 1994, Heinlein *et al.* 2003, Cheng *et al.* 2003, Scharcanski and Jung 2006, Tang *et al.* 2009a).

Conventional enhancement techniques The conventional techniques for enhancement, as described in Cheng *et al.* (2003) are contrast stretching, histogram equalisation, convolution mask and fixed or adaptive-neighbourhood statistical enhancement (Cheng *et al.* 2003). The first two techniques are performed by image intensity manipulation. This can be conducted since the graylevel histogram of an image gives a global impression of different image intensity over the dynamic range of the image. Thus, the simplest contrast enhancement can be performed by stretching the distribution of the

graylevel to achieve a higher separation between the foreground and the background. This technique is usually employed when an image has narrow graylevel distribution.

Another simple image intensity manipulation technique is called histogram equalisation. In this technique, the histogram of the image's graylevel distribution is transformed into a better histogram distribution (Rangayyan 2005). By distributing the image's histogram, it allows the low contrast area to achieve higher contrast and thus an enhancement of an image is obtained. For an example, a uniform graylevel distribution in mammogram is said to contain more information and the maximum entropy can be obtained (Cheng *et al.* 2003). However, these conventional techniques also enhance some normal tissues and noise, which may contribute to the false detection.

One of the convolution mask enhancements, the unsharp mask, was reported in work by McSweeney and Sprawls (1983) and Chan *et al.* (1987). By applying unsharp mask, the low-frequency details are reduced and the high-frequency details are amplified. This was achieved by subtracting the original image with its blurred version of image. Thus, the technique can only be employed on images with high density or overexposed mammograms. This limits the efficiency when applying the technique to the mammogram image with lower density (McSweeney and Sprawls 1983, Chan *et al.* 1987).

The previously applied conventional enhancement techniques use spatial information. Statistical enhancement methods extract local statistical information such as local mean and local standard deviation to estimate the background and suppress it to increase the contrast. This is useful especially when the mammogram has an inhomogenous background (Gordon and Rangayyan 1984, Cheng *et al.* 2003). The contrast is enhanced by finding the differences between the neighbouring pixels. The technique can either be performed by finding the differences in fixed or adaptive neighbourhood. A fixed-neighbourhood or standard-neighbour in image processing has a center of one pixel surrounded by eight pixels in the 3×3 configurations (Gordon and Rangayyan 1984). On the other hand, adaptive-neighbourhood gives more flexibility of the local area. The adaptive-neighbourhood has $m \times m$ center and surrounds of $3m \times 3m$ pixels,

where m is an odd number. This means, the contrast enhancement function can be applied accordingly to the size of the object on the image, ie. in this case, the clusters of microcalcification in mammograms. However, since the contrast function used in the technique was the square-root function, it also enhanced the noise and the background. As a result, the enhanced microcalcification were not distinguishable from the noise and the background.

Region-based enhancement technique The region-based enhancement technique for mammograms has been used by Rangayyan *et al.* in 1992 (Rangayyan 2005). Here, a method called growing seed was used to enhance the mammograms adaptively. One pixel, or they called a seed is used to grow a region by measuring the neighbourhood intensity. This produced an overlapping region of neighbourhood that had similar pixel values. The advantage of an overlapping region over the non-overlapping region is that it would not introduce an artifact at the edge of disjoints regions. An enhancement function is then applied to the grown region. This resulted in the enhancement of the region that contained the seed while other regions remained the same values. The region-based enhancement technique has shown an improvement in microcalcification visibility in the denser breast tissue on mammograms. Using this technique, the region's contrast can be enhanced without significantly enhancing other artifacts in the mammograms. However, the selection of seeds in the image is a crucial task and needed to be carefully chosen so that the region seed will contain only microcalcification and not the noises or artifacts.

Features-based enhancement technique Conventional enhancement has the disadvantage of indiscriminately enhancing microcalcification along with other artifacts in the mammograms. One method to overcome this disadvantage is by enhancing the microcalcification adapted to its features in mammograms. This can be conducted by using multi-scale analysis where the analysis of the mammogram is performed in the

frequency-domain. The aim is to enhance the high-frequency contents of the mammograms. Frequency-content analysis using wavelet-based techniques has attracted several researchers in improving the contrast of the mammograms (Laine *et al.* 1994, Strickland and Hee Il 1996, Sakellaropoulos *et al.* 2003, Heinlein *et al.* 2003, Salvado and Roque 2005, Tang *et al.* 2009a).

A similarity among these works is that they decomposed the mammograms using wavelet transform and then modified the coefficients of the decomposed images by applying enhancement operators. The enhancement operators can be either non-linear (Laine *et al.* 1994) or linear (Sakellaropoulos *et al.* 2003, Tang *et al.* 2009a) functions. The modified decomposed images with enhanced coefficients were then reconstructed to produce an image that has better visualisation of microcalcification clusters on the mammograms.

The first reported work that used multi-scale technique for contrast enhancement of mammograms was the work by Laine *et al.* in 1994. They used dyadic wavelet transform and Deslauriers-Dubuc interpolation wavelets to compute a multi-scale representation of the mammograms. The mammograms were then enhanced by reconstructing the image from the transform coefficients, which has been modified at each level by a non-linear operator. This technique has shown good contrast enhancement results. However, the levels of analysis, threshold values and the gain parameter values used in this technique must be determined before applying enhancement processing technique and it can be challenging to obtain the desired results (Laine *et al.* 1994).

The multi-scale-based enhancement was also adopted by Heinlein *et al.* in 2003. In his work, he introduced a novel technique for construction of filterbanks that produced wavelet decomposition called integrated wavelet. The interesting aspect of this work is that they have introduced a new algorithm to construct wavelet to adaptively fit various sizes and shapes of microcalcification (Heinlein *et al.* 2003). Figure 2.2 illustrates the adaptive-filter to model various microcalcification for enhancement purposes. This gives an extra flexibility in scale when compared to the conventional dyadic wavelet

transform in enhancing the microcalcification (Heinlein *et al.* 2003). However, this technique only considers an empirical specification of size range of the features to be enhanced. Thus, it is not appropriate for most general mammograms images, since the sizes and shapes of microcalcification in mammograms can vary significantly and are not known a *priori*.

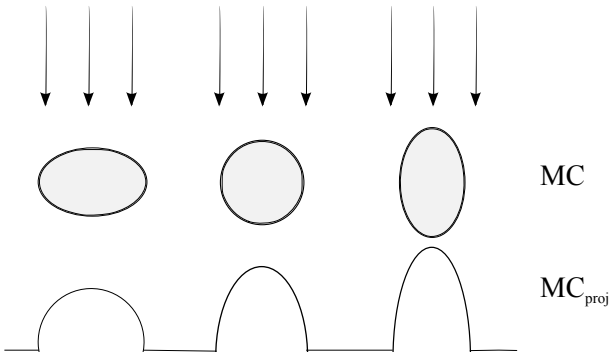


Figure 2.2: Matched Gaussian filter for modelling microcalcification proposed by Heinlein *et al.*

Another recent study that employed the multi-scale analysis for microcalcification enhancement is by Tang *et al.*, who proposed an application of wavelet transforms to CAD detection system. Their work has shown progressive results when the contrast of the images was enhanced in the wavelet domain (Tang *et al.* 2009a). They applied a uniform enhancement factor to the different scales in the mammograms using an image contrast measure as proposed by Pu and Ni (2000). The advantage of this technique is that only one parameter is needed to be modified by the radiologists when they need to perform the contrast enhancement of the mammograms. However, suitable parameter values must be carefully chosen as the visual quality of the enhanced images will be affected. If the parameter values exceeded a certain threshold, the noise and background variations will possibly be enhanced, both of which are undesirable. Moreover, different scales and different regions of the mammograms might need a different enhancement factor. This direct contrast enhancement technique only allows one fixed parameter for the enhancement factor. The method was evaluated by calculating its region contrast value (RCV) and by using this method, they have achieved

a better RCV compared to other conventional methods, ie. the histogram equalisation and unsharp masking, for contrast enhancement.

De-noising method

Jacob Scharcanski *et al.* addressed the contrast problem on mammograms by proposing a new detection system with a combination of de-noising and enhancement methods. In this work, redundant wavelet transform was employed. By applying redundant wavelet transform to the image, it produced horizontal and vertical detailed images with the same dimensions of the original image. For the de-noising purposes, wavelet shrinkage method was applied. Wavelet shrinkage is the noise reduction method which retains the high magnitude coefficient values and suppresses the low magnitude ones. At each resolution of mammograms, the coefficients represented the edge were modeled by Laplacian random variables, while the coefficients represented the noise were modeled by Gaussian random variables. Edge associated coefficients were then linearly enhanced (Scharcanski and Jung 2006). However, they claimed this as preliminary work and no quantitative result was reported apart from the figures showing the improvement in image contrast of microcalcification in mammograms.

Work by Gorgel *et al.* improved the work by Scharcanski by combining their noise reduction method with a homomorphic filtering (Gorgel *et al.* 2009). In this method, the images were pre-processed by the wavelet transform decomposition to produce the detailed coefficients of the images. These detailed coefficients revealed the horizontal, vertical and diagonal patterns. Next, the decomposed approximation coefficients were passed through a homomorphic filter to distinguish between the coefficients associated with the edges and the coefficients associated with noise. Similar to the work by Scharcanski *et al.*, noise and edges in mammograms were modeled by an additive zero-mean Gaussian noise and Laplacian random variables, respectively (Gorgel *et al.* 2009), (Scharcanski and Jung 2006). The modified coefficients were next adaptively thresholded to separate them from the background. The local threshold was chosen instead of the global threshold because it can locally determine the threshold value for each

pixel, which contributed to improved mass detection. The experimental result showed that this method greatly improved the contrast of the mammograms.

2.1.2 Segmentation

After the pre-processing has been made to the mammograms, a better visualisation of mammogram features is achieved. Next, the enhanced area in the mammograms is segmented. In general, the purpose of the segmentation stage is to extract an object or region of interest (ROI) from the image. The robustness of the segmented area depends on how well the structure is separated from the background by the chosen pre-processing techniques. In CADe system, the term segmentation carries multiple definitions. Several ways of segmentation of a mammogram image include the segmentation of an area containing microcalcification and the segmentation of an individual microcalcification. One of the simplest conventional methods is thresholding. This method is most suitable for cases with images that have uniform background. By observing their histograms, the image will show two distinct peaks that represent the object and the background. A threshold value between the two distinct peaks can be determined, which separates the object from the background.

However, the choice of the segmentation method depends on the type of features that will used for the detection purposes. Hence, the work in segmentation can also be described in the features analysis step that will be discussed in Section 2.1.3.

2.1.3 Feature Analysis

After the segmentation step is carried out, the next stage is to analyse the features from the segmented region. This analysis determines whether the segmented area contains individual microcalcification or clusters of microcalcification, or may be possibly some noises or artifacts. Hence, the aim of the analysis stage is to confirm the appearance of microcalcification in the mammograms. One of the analysing methods is by utilising

the machine learning method. Generally, several sets of features extracted from the images are presented as an input to the learning machine, such as a classifier. The features extracted are useful and contain information that are used to detect the clusters of microcalcification in mammograms. Two common features that are extracted for detection are the individual microcalcification features (Veldkamp and Karssemeijer 1999) and the texture-based features.

Individual Microcalcification Features

Features of individual microcalcification are normally segmented directly from the mammograms. In work by Veldkamp *et al.*, statistical measures of the extracted individual features were calculated (Veldkamp and Karssemeijer 1999). The parameters were: perimeter, area, compactness, elongation, eccentricity, thickness, orientation, direction, the mean of line/edge detector, mean intensity of the background, mean intensity of the detected microcalcification, distance of each microcalcification in a cluster and contrast measure. In addition to the individual microcalcification features, cluster features were also been extracted. The statistical measures of the cluster features include: mean, median, standard deviation, minimum and maximum value of local features in the clusters. They also added the size of cluster's area and the number of microcalcification in the clusters as the features. These features were then classified using a k-nearest neighbour (kNN) classifier. However, the downside of the technique is that the microcalcification clusters needed to be accurately segmented. The task is challenging since the microcalcification clusters in mammogram are subtle and variable.

Region-based Features

To effectively extract features of the individual microcalcification, it is necessary to have a robust segmentation method to ensure high accuracy of the measurements. However, to avoid this requirement, the techniques that extract the features from the textures of regions containing microcalcification, such as their statistical-texture

features and multi-scale-texture features can be performed. One of the most common statistical-texture feature extraction is called the surrounding region-dependence method (SRDM). It is based on the second order histogram in the two surrounding regions. In the SRDM, the regions were transformed into a surrounding dependent matrix. The coarseness or fineness of an image can be interpreted as a distribution of elements in the matrix. The graylevel will be similar between pixel and its surrounding if the images were smooth and vice versa. If the area has a matrix that was distributed diagonally or in other means there is difference between pixel and surrounding, then the area is considered to potentially contain suspected microcalcification. The method was tested on 140 regions of interest and achieved sensitivity of more than 90%. This method outperformed the individual feature extraction method as it does not need to carefully segment the microcalcification individually. However, because the method depends on the graylevel of mammograms images, it is sometimes hard to distinguish microcalcification in the denser breast tissues.

Another method to extract texture-based features is by using multi-scale method. Wavelet-based approaches have shown promising results in assisting the radiologists in the overall system's performance (Laine *et al.* 1994, Laine *et al.* 1995b). This is due to the multi-resolution property of wavelet transforms that enables zooming into any irregularities of an image and thus allows us to characterise them locally. The properties offered by wavelet transform will be further discussed in Section 3.3.

Although many methods have been proposed in the CADe system, the development of new algorithms is still an active research topic in the field. The previous proposed methods in CADe include the basic image or features enhancement, statistical measurements, multi-scale decomposition methods and machine learning methods. There are also some of the research that integrate two or more of the methods for CADe purposes. For example, Yu *et al.* combined the method of multi-scale decomposition for feature extraction and also statistical measurements features with neural network machine learning for the detection classification (Songyang and Ling 2000).

Section 2.2 extensively reviews the previous work conducted for malignancy analysis of the microcalcification clusters. Most of the approaches in the CADx system however, adapt almost similar techniques as in the CADe system.

2.2 CAD Diagnosis of Microcalcification (CADx)

Computer-aided diagnosis or CADx systems aim to assist radiologists in the characterisation of the detected microcalcification clusters into benign or malignant classes. The CADx acts as a diagnostic tool. The input of the CADx is normally obtained from the output of the CADe system, ie. the detected microcalcification clusters. The CADe systems will only detect or enhance the abnormalities in mammograms without classifying them into benign or malignant cases. This is because the CADx task is to assist radiologists in classifying them.

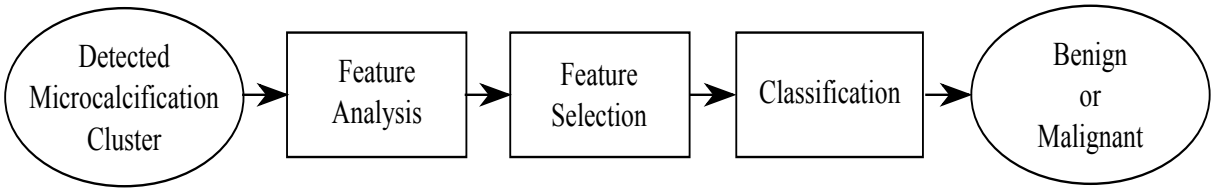


Figure 2.3: Flowchart of a CADx system commonly used for malignancy diagnosis of microcalcification clusters. Most of previous work adapted the same process. Firstly, the input of detected clusters is obtained from CADe system or from the experts. The features are then extracted from the area. If the number of features extracted are enormously high, the feature selection process is carried out for dimensionality reduction, thus the dashed box represents an optional process. These features are then fed into a classifier for classification and malignancy analysis purposes.

Figure 2.3 illustrates the common steps used by CADx systems for malignancy diagnosis. Given the input from the CADe system, the next stage is to extract features that can be used to represent the malignancy type of the microcalcification clusters. The features extracted by the CADx system are used to classify the clusters into benign or malignant classes. The final stage is that the classification of the extracted features is performed by the classifier.

This section reviews methods used in previous work for microcalcification diagnosis. Some of the methods especially for feature extraction are almost similar with the methods in the CADe system as presented in Section 2.1.3. However, the features extracted in the CADx system are now used for malignancy analysis. Since the feature extraction techniques have been described extensively in Section 2.1.3, the following subsections briefly summarise the previous work in CADx system and the performance of the methods for microcalcification diagnosis. The effectiveness of the feature extraction techniques can be evaluated by the results obtained from the classification step. Commonly, the performance evaluation is made based on the true positive rate (TPR) and the false positive rate (FPR). The true positive rate is the rate of correctly diagnosed case to be malignant. Whilst, the false positive rate is the rate of incorrectly diagnosed case to be malignant cases. This will further be discussed in Section 6.1.6. Another way to evaluate the classifier performance is by determining the accuracy obtained by the classifier, ie. the rate of correct classification of the microcalcification clusters into their respectively classes.

2.2.1 Feature Extraction

Before diagnosis or classification can be performed, a useful set of features that represents the characteristics of the clustered microcalcification needs to be extracted. Several methods, such as multi-scale, neural network and morphology-based are among the most useful feature extraction methods for microcalcification classification purposes. These features are then fed into a classifier. The following subsections discuss in detail the previous work in feature extraction and classification methods that have been implemented and tested by previous researchers in this field.

Individual Features

The features of individual microcalcification are directly extracted from their shapes. The features include the size, area, compactness, and the number of microcalcification in a cluster. Zadeh *et al.* extracted 17 shape features from each mammogram to characterise the clusters of microcalcification, including the physical features such as the number of microcalcification in the cluster and the maximum size of the calcifications, as well as the statistical features such as the average, standard deviation and the compactness of the individual microcalcification. A feature selection experiment was conducted using genetic algorithm, GA-based global search method and then a kNN classifier is used to obtain the performance of the overall system. The highest performance achieved by the classifier is recorded with sensitivity of 0.82 (Soltanian-Zadeh *et al.* 2001). Again, this approach requires higher accuracy of segmentation of microcalcification from the background to ensure the robustness of the features classifications step. Region-based feature extraction offers an advantage by reducing the demand to have a very accurate segmentation of microcalcification. This is particularly significant because the microcalcification are small and subtle and due to the low contrast of mammography, it is possible to have inaccuracies when segmenting the microcalcification individually.

Region-based Texture

Features extracted from the region-based techniques can reduce the demand of highly accurate segmentation. Two of the approaches in region-based are statistical texture features and multi-scale texture features. Statistical texture features measure the statistical calculations based on the pixel information of the images. One of the methods that can be used to extract the statistical texture features is the surrounding region dependence method (SRDM). This method also has been explained in Section 2.1.3, where the SRDM features were used to detect the microcalcification clusters on the mammograms.

On the other hand, multi-scale texture features are extracted from the spatial information of the images. The information was obtained by transforming images into sets of coefficients that contain the information of the images. Multi-scale analysis has shown to be advantageous for texture analysis, because of this framework's ability to analyse an image at different resolutions. This is known as a zooming property, and has advantage in analysing mammograms, due to the fuzzy shapes and small sizes of microcalcification. One of the methods that can be used to perform multi-scale analysis is the wavelet transform. An image is decomposed using discrete wavelet transform into set of detailed images at different resolutions (refer to Section 3.3 for more details). Common features that are extracted from these detailed images are the energy and entropy of the coefficients.

Dhawan *et al.* used two features, energy and entropy, extracted from the multiple resolution images as the local texture features to discriminate between malignant and benign cases. The images were decomposed using wavelet transform. Together with other global texture features, the performance of the method yields a sensitivity of 0.83 (Dhawan *et al.* 1996).

The work by Soltanian-Zadeh *et al.* (2004) has shown that features extracted from the wavelet and multiwavelet achieved the highest classification sensitivity rate when compared to the other features, including shape features and statistical features. In this work, images were decomposed using three wavelet packets, ie., Daebuchies 6, 10 and 12. Energy and entropy of each of the decomposed subband images were calculated and used as the features to determine the malignancy class of the microcalcification. Comparative study showed that classification by using wavelet-based features achieved the sensitivity rate of 0.85, which was higher than the shape and statistical features techniques that achieved sensitivity rates of 0.82 and 0.77 respectively.

2.2.2 Classification

The classifier plays an important role in categorising the microcalcification clusters into benign or malignant classes. A set of features that represent the clusters was used as the input patterns to the classifier. The aim is to generate a pattern recognition algorithm that can differentiate between these two classes. Some of the classifiers that have been used in classifying microcalcification include support vector machine (SVM) and artificial neural network (ANN). The topic of classifiers will be discussed separately in Chapter 4.

In 2005, Wei *et al.* compared the performance of several machine learning methods that have been used in classification of microcalcification (Wei *et al.* 2005). The machine learning methods were the support vector machine, relevance vector machine and the kernel Fisher discriminant. These supervised classifiers learn the patterns of the clusters of microcalcification in order to develop the classification algorithm. The set of features in this study were the 8 shapes features: 1. the number of microcalcification in the cluster, 2. the mean effective volume (area times effective thickness) of individual microcalcification, 3. the area of the cluster, 4. the circularity of the cluster, 5. the relative standard deviation of the effective thickness, 6. the relative standard deviation of the effective volume, 7. the mean area of microcalcification, and 8. the second highest microcalcification-shape-irregularity measure. These features were extracted from the dataset collected by the Department of Radiology at the University of Chicago. There were 386 cases with 75 being malignant and 311 benign. This study considered the problem of learning the patterns from the multiple views of mammograms, MLO and CC of each breast. Three different methods were used for these multiple views; direct, averaging and joint methods. In this work, it was shown that kernel-based method for supervised classifier outperformed the feed-forward neural network, FF-NN (Wei *et al.* 2005). The outcome of this research motivated us to investigate the kernel-based supervised classifier, ie. SVM for microcalcification classification. The

advantage of the SVM is that the machine is relatively well understood and has fast implementations.

Meanwhile, the use of artificial neural networks (ANNs) also has attracted many researchers in the area of microcalcification diagnosis. ANN has the ability to model the complex relationship in classifying microcalcification clusters given the set of features as input. Thus, the ANN has become one of the most frequently used as a classifier. These include the previous work by Tsujii *et al.* (1999), Verma (1998), Verma and Zakos (2001) and De Melo *et al.* (2010).

In the work by Verma (1998), a feed-forward neural network (FF-NN) with single hidden layer was used to classify the microcalcification clusters. Two types of training methods were tested: error backpropagation with momentum and the direct solution method. The Nijmigen database was used with 105 areas containing the suspected microcalcification (76 malignant and 29 benign). The database was divided into 90 data cases for training and 15 data cases for testing. The parameters that needed to be adjusted during the training and testing were the number of hidden units, learning rate, momentum and the number of iterations. The result obtained was 87% of correct recognition (Verma 1998).

The work by Tsujii *et al.* (1999) also adopted FF-NN with a single hidden layer as the classifier. The paper reported that the aim was to overcome the overfitting problem in the radial-basis function neural network (RBF-NN). This problem arises due to the selection of three parameters in the radial-basis function neuron. The overfitting occurred when these parameters were chosen by each set of data. Therefore the number of degrees of freedom in the network was equal to the number of the data points. Thus, the network will learn all the details, including the unrepresentative peculiarities of individual data instead of the patterns in the data. Hence, to overcome this problem, a method named trend-oriented RBF-NN was proposed. The modification was made upon the cost function where in this function, the approximation of the center y_j and width σ_j were found and at the same time to reduce the number of degrees of freedom.

The mini-MIAS database was used with 128 areas contain suspected microcalcification (47 benign and 81 malignant). Before putting the ROIs into the neural network classifier, feature extraction and feature selection procedures were conducted. Features selection was carried out using Karhunen-Loeve transform. The training and testing were performed using the round-robin method. In round-robin method, one data was tested on the remaining data as training dataset. The result compared the performance obtained from the proposed RBF-NN with the unmodified RBF-NN (Tsujii *et al.* 1999). It was shown that the proposed system achieved higher sensitivity with the true positive rate against the false positive rate of 0.76, whilst unmodified system achieved a lower rate of 0.73.

There was also work proposed to improve the mammography diagnosis by improving the method for features selection before putting into a FF-NN classifier. Melo *et al.* ranked features of both individual and clusters of microcalcification using Fishers Discriminant Ratio (FDR). The neural network with a single hidden layer was used as classifier in this work. Eight different sets of features; each consisted of 5, 6, 7 or 8 higher ranked individual and cluster features were fed into the neural network as inputs. The hidden layer was varied to have two nodes for microcalcification cluster features or three nodes for individual microcalcification features. The highest accuracy achieved by the classifier when the microcalcification cluster features were ranked by FDR was 86.19% (De Melo *et al.* 2010). Whilst the highest accuracy achieved when using individual microcalcification as input features was 72.81%. From this work, it was observed that the cluster features provide more malignancy information compared to the individual microcalcification features, hence the better accuracy rate.

Most of the previous work that used neural network for microcalcification classification focus on improving the input vectors and the parameters of the learning algorithm in neural networks. However, the architecture of neural network is flexible, ie. the number of hidden nodes and the number of hidden layers in the neural network is not fixed and can be manipulated to increase the efficiency of the network. Yet, the conventional neural network training algorithms are known for a propensity to get stuck

at poor local minima, commonly when the networks have more than two hidden layers. This decreases its overall performance. Nevertheless, a deep network is claimed to have the ability to extract more features of a complex structure, and this will further discussed in Chapter 5.

2.3 Neural Network in Biomedical Applications

Sections 2.1 and 2.2 discuss extensively the previous work in CAD systems for microcalcification detection and diagnosis. From the discussion, it emerged that neural networks are one of the common methods used for features extraction and classification. In recognition of their prominence, this section reviews the use of neural networks in related applications in the biomedical field.

One application of neural networks is in segmentation of brain magnetic resonance (MR) images (Demirhan *et al.* 2015). Brain MR images were segmented into several classes such as tumor, edema, white matter (WM), gray matter (GM and cerebrospinal fluid (CSF). A neural network is used to learn the segmentation process using self-organizing map (SOM) and fine-tuning the weight vectors with the labeled data for classification. Shen *et al.* (2005) used neural networks to determine the parameters in segmenting brain MR images. The main technique in segmenting the images was fuzzy c-means clustering (FCM). The work improved the FCM by considering the neighboring pixels intensities, and the degree of neighbourhood attraction is determined using a neural network. The proposed method has shown to have the lowest incorrect segmentation percentage of 0.025% when compared to the conventional FCM that achieved an incorrect segmentation percentage of 14.24%.

In work by Neofytou *et al.* (2015), a neural network is used for early detection of endometrial cancer. Texture features, such as the statistical features (SFs), spatial gray-level dependence matrices (SGLDM) and gray- level difference statistics (GLDS) were extracted from the images. A probabilistic neural network with RBF kernel is used to classify the features. The work also compares the performance when the SVM is used

as the classifier. However, it was found out that SVM performs better than neural network in diagnosing endometrial cancer with extracted texture features as the input for classification.

Neural network also has been used to detect and classify masses in mammograms (Azevedol *et al.* 2015, Wong *et al.* 2012, Islam *et al.* 2010). In Azevedol *et al.* (2015), an algorithm that is called Morphological Extreme Learning Machines (mELMs) is developed for this purpose. This algorithm is based on an artificial neural network but with improved characteristic such as training speed and data prediction. To overcome the problem of being stuck in local minima, the learning in the algorithm is based on the Moore-Penrose generalised inverse, where the learning is performed by batch and the output weights are calculated. To validate the performance of the developed classifier, four different features were used as input to the network, ie. Haralick, wavelet and Haralick, Zernike, and wavelet dan Zernike. Three different kernels in the network were used for comparison, ie. sigmoid, dilation and erosion kernels. It was shown that the morphological kernels, ie. dilation and erosion kernels improved the accuracy of masses classification rate in mammograms.

In Wong *et al.* (2012), 12 features from the gray level co-occurrence matrix (GLCM) were extracted before neural network is used to classify the masses and achieved a classification accuracy of 86%. In Islam *et al.* (2010), the statistical texture features of the mammograms images were used as the input to a multi-layer perceptron (MLP). The neural network that used to classify each mass as benign or malignant has 3 layers: an input layer, a hidden layer with 5 units and an output layer. The neural network-based system achieved a classification accuracy of 83.87% and 90.91% for benign and malignant, respectively (Islam *et al.* 2010).

Another diagnosis of disease that has applied neural network is cardiovascular diseases (Karabulut and Ibrikci 2012). This work used a neural network with Levenberg-Marquardt back propogation algorithm as the classifier. The neural network used in

this work consists of three layers; one input, one hidden and one output layer with sigmoid transfer function as the activation function. The work by Karabulut and Ibrikci (2012) achieved an accuracy of 91.2% in diagnosing the disease. Other than cardiovascular disease, heart valve defect has also been diagnosis with the use of neural network (Uguz 2012). Neural network is used to classify the features of the heart sound to detect the defectiveness in the heart valve. The heart sounds were captured using the stethoscope. The features of the signals were then extracted by applying the discrete Fourier transform to the signal. Neural network was then used to classify them into three different states of the heart diseases, ie. mitral stenosis, aortic stenosis and ventricular septal defect. It was shown that the method achieved an accuracy rate of 92% (Ghwanmeh 2012).

The reviewed works above are some of applications in biomedical field that relied on neural networks for crucial functions in their systems. Neural networks can accommodate highly non-linear relationships between the inputs and output, which partly explains why they have been popular for modelling complex images, especially for biomedical applications. Details of neural networks, along with more recent advances in this area, will be the subject of Chapter 4.

2.4 Summary

This chapter discusses the previous work in both CADe and CADx systems for mammograms diagnosis, specifically for diagnosing microcalcification clusters. Works in CADe system have made great progress and many approaches have shown promising results to detect microcalcification within mammograms. On the other hand, results from previous work in CADx revealed that there is room for improvement that is needed in the system for classifying microcalcification clusters. The complex structures of microcalcification clusters require a robust feature extraction method and an intelligent learning algorithm in order to successfully classify their classes.

2.4 Summary

Multi-scale analysis approaches have shown promising results to produce useful features for classification. However, multiple orientation features have not yet been fully exploited for classification purposes. Hence, the work in this thesis will further investigate the usefulness of multiple orientation features together with the multiple resolution features, to classify the microcalcification clusters. The method proposed is a natural extension of wavelet-based techniques which offers an improved sensitivity to orientations. Thus, the technique may possibly extract more useful features from multiple orientation that are associated with malignancy analysis. In Chapter 3, the properties of commonly used multi-scale method, the wavelet transform and the combination of multiple resolution with multiple orientation method, called the steerable pyramid filtering will be further discussed.

The other focus in the study of the CADx system is to investigate the best classifier for microcalcification clusters classification. The support vector machine (SVM) is one of the classifiers that has shown an improvement in diagnosis compared to neural networks (Wei *et al.* 2005). The neural network (NN) is one of the machine learning algorithms that is based on the multilayer perceptron with many hidden layers. However, conventional NNs failed to perform with an increasing number of hidden layers. It was shown that NN can only work with small number of hidden layers. The NN efficiency degrades when performing a deep learning of complex structures, such as the structure of microcalcification clusters. Recently, Hinton *et al.* (2006) has introduced the deep learning architectures that can be used as a classifier. Hence, the deep neural network for the improvement of the microcalcification diagnosis will be further investigated in Chapter 5.

Chapter 3

Mathematical Background

THIS chapter provides the mathematical background that commonly serves as the backbone in most signal processing applications. Signal processing is any activity that involves the acquisition, representation and analysis of a signal. Some of the examples of signal include audio, video, image, and speech. Many of the real-world applications deal with signals. Thus, signal processing has assumed central importance to various applications in several fields such as communications, biomedical engineering, echolocation technologies and image processing.

The rest of this chapter is organised as follows: Section 3.1 briefly describes the history of Fourier analysis, which remains the most often used analytical framework in signal processing. Section 3.2 describes the multiple resolution analysis and how the wavelet transform achieves multiple resolution analysis is described in Section 3.3. Steerable pyramid filtering is an extension of wavelet analysis with an improved directionality for image analysis and is described in Section 3.4.

3.1 Fourier Series and Fourier Transform

In early 1800s, Fourier analysis was discovered by a French mathematician, Joseph Fourier, and has been widely used as one of the important tools in various signal processing applications. He has discovered that every periodic signals can be decomposed into sinusoidal components with different harmonic frequencies. The Fourier series of the periodic signal $f(t)$ is defined as (Bracewell 1999):

$$f(t) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi t}{T} + b_n \sin \frac{n\pi t}{T} \right), \tag{3.1}$$

where T is the function period and the coefficients a_n and b_n determine the relatively weight of each sinusoidal component.

A function is approximated by a truncated Fourier series with a finite number of N terms in the function. As expressed in the Equation 3.1, the sine and cosine functions in the Fourier expansion form the bases for the space of periodic functions. Figure 3.1 illustrates how the sinusoids are used in approximating a square wave using Fourier series expansion. From the figure it was observed that as the number of coefficients, N, increases, the Fourier series approximates the square wave with an ever greater accuracy.

Fourier transform was later formulated as a generalisation of Fourier series for non-periodic functions. By Euler's formula of $e^{j2\pi\theta} = \cos(2\pi\theta) + j\sin(2\pi\theta)$, the bases of Fourier, ie. the sine and cosine functions can be written in term of complex exponential, $e^{j2\pi\theta}$. The Fourier transform, which transforms the time domain signal $f(t)$ into the frequency domain $F(\omega)$, is defined as:

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt, \tag{3.2}$$

and the inverse Fourier transform is

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{j\omega t} d\omega, \tag{3.3}$$

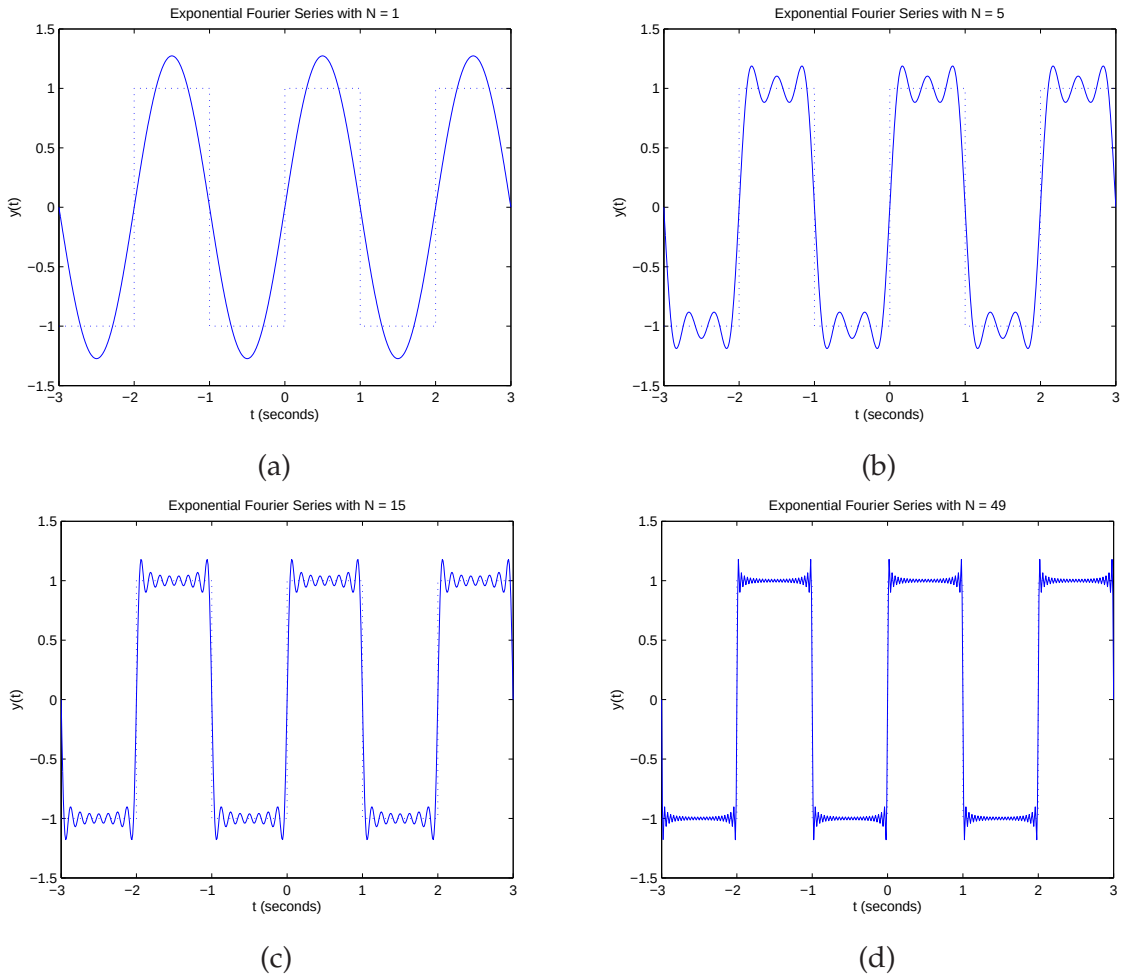


Figure 3.1: Fourier series for square wave approximation with number of coefficients (a) $N = 1$, (b) $N = 5$, (c) $N = 15$ and (d) $N = 49$, respectively.

where ω is the angular frequency.

By applying Fourier transform to a signal, the result gives information of the power distribution across frequencies, or the spectral density $S_x(\omega)$ of the signal $f(t)$. The power spectral density is defined as:

$$S_x(\omega) = |F(\omega)|^2. \tag{3.4}$$

The main drawback of Fourier transform stems from its use of smooth periodic sinusoids in the analysis, which gives rise to inaccuracies in the vicinity of signal discontinuities, a phenomenon known as the Gibbs phenomenon. Referring to Figure 3.1,

Fourier series well approximates a square wave with $N = 49$ coefficients. However, at the edge of the square wave there is an overshoot in the approximation function. The phenomenon is attributed to the poor time localisation of the sinusoidal basis functions. Thus Fourier transform is unsuitable for representing signals that contain discontinuities which often occur in many real-world signals.

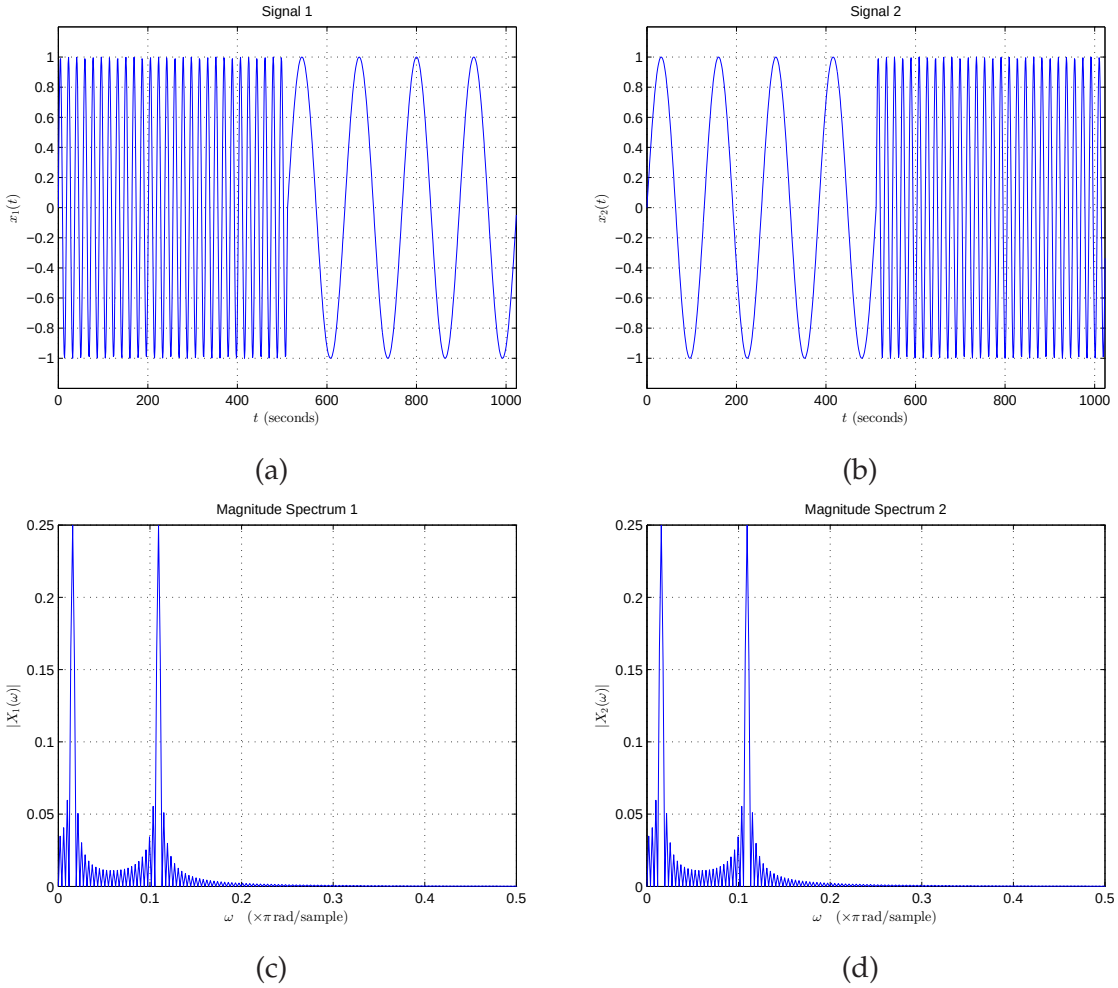


Figure 3.2: (a)-(b) Non-stationary signals with different frequency that change after 500 seconds. (c)-(d) Fourier transforms of the signals in (a) and (b) respectively. The Fourier transforms of both signals produce the same magnitude spectrum, without giving the time information when the change in frequency occurred.

Viewed in another way, the Fourier transform provides localisation in the frequency domain only. Hence, for any signal whose characteristics change with time, known generally as a non-stationary signal, its Fourier transform embeds such changes in its phase only. This implies that, Fourier transforms are not ideally suited to applications

that need both time and frequency localisations (Cohen and Kovacevic 1996). An illustrative example is shown in Figure 3.2. In the Figure 3.2(a), the signal has higher frequency in the first 500 seconds, and changing to lower frequency afterwards. Hence the magnitude spectrum in Figure 3.2(c) shows two distinct peaks that represent the frequencies of the signal. In Figure 3.2(b), the signal starts with lower frequency, and after 500 seconds, the signal changed to have higher frequency. However, the magnitude spectrum in Figure 3.2(d) shows the similar result. The Fourier transform unable to give the time information of each frequency of the signal. Thus, Fourier transform is said to be lack in time localisation.

With the shortcoming of the Fourier transform, it results an inefficiency in representing functions that contain spikes and discontinuities. The Fourier transform also lacks time localisation and so is unsuitable for non-stationary signals. This makes this tool less powerful for certain applications analysis such as signal compression.

3.1.1 Short Time Fourier Transform

A method known as time-frequency representation describes the signal represented over both time and frequency axes. The time localisation function, $\delta t(\psi)$ of a time-varying function $\psi(t)$ is defined as:

$$\delta t(\psi) = \frac{\int_{-\infty}^{\infty}(t-\bar{t})^2|\psi(t)|^2dt}{E}, \tag{3.5}$$

and the frequency localisation function, $\delta\omega(\psi)$ of its Fourier transform $\Psi(\omega)$ is defined as:

$$\delta\omega(\psi) = \frac{\int_{-\infty}^{\infty}(\omega-\bar{\omega})^2|\Psi(\omega)|^2d\omega}{E}, \tag{3.6}$$

where $\bar{t}$ and $\bar{\omega}$ are the time and frequency central values respectively. They are defined as:

$$\bar{t} = \frac{\int_{-\infty}^{\infty} t|\psi(t)|^2 dt}{E} \tag{3.7}$$

$$\bar{\omega} = \frac{\int_{-\infty}^{\infty} \omega|\Psi(\omega)|^2 d\omega}{E}, \tag{3.8}$$

where

$$E = \int_{-\infty}^{\infty} |\psi(t)|^2 dt = \int_{-\infty}^{\infty} |\Psi(\omega)|^2 d\omega. \tag{3.9}$$

Qualitatively, δt measures the degrees of spread of the function $\psi(t)$ at time axis, whilst $\delta\omega$ measures the spread of its spectrum at frequency axis.

The time-frequency localisation property is subjected to the uncertainty principle. The principle limits the product of time and frequency uncertainties by:

$$\delta t(\psi)\delta\omega(\psi) \geq \frac{1}{4\pi}. \tag{3.10}$$

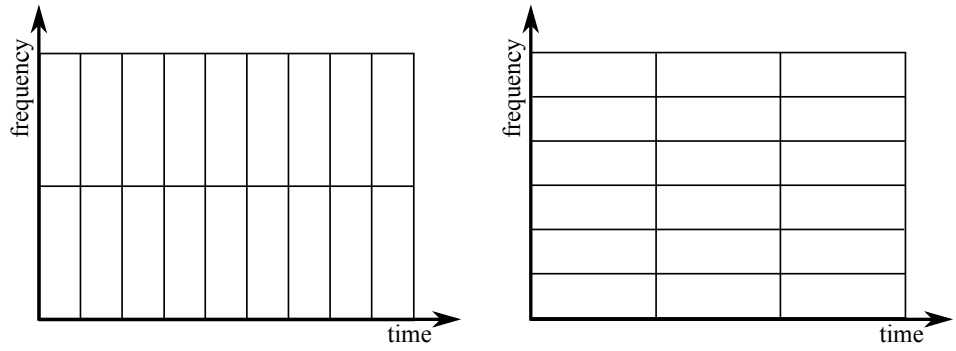
The short time Fourier transform (STFT) was introduced to overcome the drawbacks of the conventional Fourier transform. It modifies the Fourier transform such that it can be used to represent signal with localisation in both time and frequency.

Continuous-time STFT in 1-dimension is defined as:

$$\mathbf{STFT}\{x(t)\}(\tau,\omega) \equiv X(\tau,\omega) = \int_{-\infty}^{\infty} x(t)W(t-\tau)e^{-j\omega t}dt, \tag{3.11}$$

where $W(t)$ is a window function centered on $t = 0$.

In STFT, a window function $W(t)$ is used to divide the signal function $f(t)$ into (overlapping) sections and is concentrated over a finite time range of t. The signal division is carried out by sliding the window function $W(t)$ along the time axis of the signal.



(a) Narrow window (b) Wide window

Figure 3.3: STFT frequency-time joint tiling; (a) shows narrow STFT window has good time resolution but poor frequency resolution, and (b) shows wide STFT window has good frequency resolution but poor time resolution.

Next, the frequency content of each section is analysed individually with a Fourier transform. The STFT thus provides localisation in both frequency and time domains.

However, the limitation of STFT is the size choice of the window function $W(t)$. The size choice of window function $W(t)$ is crucial, since once the window function has been chosen, the time-frequency localisation is fixed. With these limitations, STFT analysis can either possesses good time resolution or good frequency resolution, but not both (Cohen and Kovacevic 1996). Thus, the resolutions of different basis functions are the same at all frequencies and times (Graps 1995, Rioul and Vetterli 1991). Figure 3.3 illustrates how the window function has either good frequency localisation or good time localisation. In Figure 3.3(a), the window is narrower, thus it provides excellent time resolution but poor frequency resolution. Whilst, in Figure 3.3(b), the window is wider, thus it has excellent frequency resolution but poor time resolution.

For function that has steady-state response, the wider window would suit, whilst function with spike or transient response, the narrower window would suit. However, if a function has both steady and transient response, the analysis using STFT with single window may not be carried out appropriately. This problem can be addressed by an analysis that can performed with variety of window widths. This is called multiple resolution analysis and will be discussed in Section 3.2.

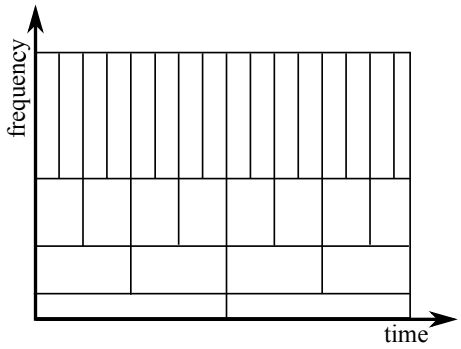


Figure 3.4: Wavelet frequency-time joint tiling shows that wavelet can provide both good time and frequency resolution.

3.2 Multi-resolution Analysis

Motivated from the drawbacks in Fourier transform and STFT, multiple resolution analysis (MRA) has been introduced to provide an alternative analysis that addresses these shortcomings. This section discusses the theory of MRA, which provides the main framework for constructing compactly supported wavelets. This idea was originally proposed by Meyer and Mallat in 1986.

The idea behind MRA is to provide a transform that produces both varying time and frequency resolutions through the MRA framework depending on the signal characteristics. Figure 3.4 illustrates the joint frequency-time tiling provided by MRA. The joint frequency-time tiling shows that the MRA can provide good time resolution but poor frequency resolution at high frequencies and good frequency resolution but poor time resolution at low frequencies. This characteristic proves to be beneficial for applications where the signals often contain both high frequency components for short durations (transient response) and low frequency components for long durations (steady-state response).

A MRA on the space of finite energy functions, $L^2(\mathbf{R})$ can be illustrated as in Figure 3.5. The basic requirement of MRA is to achieve nested spanned spaces such that the space containing the high resolution signals will also contain all the lower resolutions (Burrus *et al.* 1998). This is mathematically expressed as a sequence of nested

subspaces, $V_j, j \in \mathbf{Z}$ to satisfy

$$\ldots V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \ldots \tag{3.12}$$

where $\mathbf{Z}$ is the set of integers (Graps 1995, Vetterli and Kovacevic 1995, Burrus *et al.* 1998). The set of subspaces is subjected to the following conditions:

1. Upward completeness
$$\overline{\cup V_j} = L^2(\mathbf{R}). \tag{3.13}$$

2. Downward completeness
$$\cap V_j = \{0\} \tag{3.14}$$

3. Self-similarity
$$x(t) \in V_j \iff x(2t) \in V_{j-1}. \tag{3.15}$$

4. Translation invariance
$$x(t) \in V_0 \iff x(t-n) \in V_0 \, , \forall n \in \mathbf{Z}. \tag{3.16}$$

Equation 3.16 implies that V_0 has an orthonormal basis consisting of all integral translates of a single function $\phi_{0,n}(t) : \phi(t-n) \forall n \in \mathbf{Z}$.

Illustration of the above mathematical expression is shown in Figure 3.5. The figure shows that the space V_2 is a proper subspace of V_1, and V_1 is a proper subspace of V_0 and so on. This also means that the hierarchical structure of multiple resolution analysis approximates a function at different scales or resolutions. As the term j decreases, the finer the approximation to a function $f \in L^2\mathbf{R}$.

Notice also there are orthogonal complement components in the space labelled W_1 and W_2 in the Figure 3.5. The complement components are the error spaces between

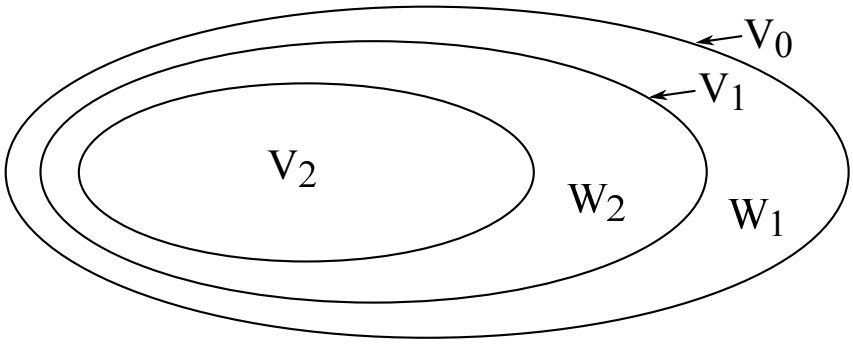


Figure 3.5: Nested vector spaces spanned by scaling function.

different resolutions.

$$V_{j-1} = V_j \oplus W_j \tag{3.17}$$

These are the orthonormal basis subspace in the MRA structure. The W_j exists with the same conditions of V_j holds, which are:

1. Completeness

$$\oplus W_j = L^2(\mathbf{R}). \tag{3.18}$$

2. Self-similarity

$$x(t) \in W_j \iff x(2t) \in W_{j-1}. \tag{3.19}$$

3. Translation invariance

$$x(t) \in W_0 \iff x(t-n) \in W_j \, , \forall n \in \mathbf{Z}. \tag{3.20}$$

Similarly in the previous paragraph, Equation 3.20 implies that W_0 has orthonormal wavelet basis consisting of all integral translations of a single function $\psi_{0,n}(t) : \psi(t - n) \forall n \in \mathbf{Z}$.

Thus, from the MRA analysis, the scaling and wavelet bases which are equivalent with the orthornormal basis for subspace in V_j and W_j can be obtained. Section 3.3 discuses

the wavelet analysis that was constructed in multiple resolution analysis and its relationship with iterated filterbanks, which can be used to decompose an image into multiple resolution subbands.

3.3 Wavelet Analysis

Previous section has discussed how wavelet and scaling functions are related to MRA. In this section, a further discussion on analysis using wavelet transform, mainly for image decomposition, is provided (Graps 1995, Vetterli and Kovacevic 1995, Burrus *et al.* 1998). A wavelet analysis uses a family of functions made up of the dilations and translations of a template function called the mother wavelet, $\psi(t)$. The (dyadic) dilated and translated version of $\psi(t)$ are its wavelet basis, which is expressed mathematically as:

$$\psi_{k,n}(t) = 2^{\frac{n}{2}}\psi(2^{-n}t - k) \tag{3.21}$$

where the variable k and n in Equation 3.21 represent the wavelet location index and width of the wavelet basis, respectively.

To further explain Equation 3.21, refer back to the nested vector spaces in previous section (as in Figure 3.5). Let V_0 be a base scale space. This space is spanned by a scaling function $\phi(t)$ and its translated versions $\phi(t-n)$ for integers n. The scaling function at this scale has a given width; by convention, this is defined to be the scale in which the set of all integer shifts of $\phi(t)$ is a basis for V_0. Thus, the space describes signals of a particular resolution, or is labeled as at resolution 0.

In contrast, V_{-1} is space spanned by scaling function $\phi(2t)$. The signals in this space has twice as fine a resolution as V_0. Similar as in previous paragraph, the functions $\phi(2t)$ and its translation version $\phi(2t-k)$ form a basis for V_1 . This pattern applies as the scaling function changes the time scale, $\phi(2^j t)$, for all integers j. In conclusion, for $j < 0$, the space spanned is larger and the resolution is higher, while for $j > 1$, the space spanned is smaller and the resolution is lower.

Since, each of the vector space V_j can be expanded by changing the time scale of scaling function, the scaling function, $\phi(t)$ now can be written as a sum of weighted shifted $\phi(2t)$:

$$\phi(t) = \sum_k h_0(k) 2^{\frac{1}{2}} \phi(2t - k). \tag{3.22}$$

The characteristic of a signal function is not only determined by the set of dilations and shifts of the scaling function. It is also described by a different set of function $\psi(t)$. This set of function is called the wavelet function. The wavelet functions span the difference between the two successive resolutions. That is, the wavelet functions of V_1 are the functions that are included in V_0 but not in V_1, thus containing the missing details between the successive scales. These are the wavelet functions that span a space W_1.

$$V_0 = W_1 \oplus V_1. \tag{3.23}$$

For the case where wavelet function $\psi(t)$ is orthogonal with the scaling function, the wavelet function, $\psi(t)$ can be written as a sum of weighted shifted of scaling function $\phi(2t)$ with appropriate coefficients $h_1(n)$ as in Equation 3.24:

$$\psi(t) = \sum_k h_1(k) 2^{\frac{1}{2}} \phi(2t - k). \tag{3.24}$$

As the requirement for wavelet is to span the orthogonal spaces to the scaling function spaces, the coefficient $h_1(n)$ must be related to $h_0(n)$ by $h_1(n) = (-1)^n h_0(1 - n)$. From this equation, the prototype mother wavelet is generated as in Equation 3.21. Using this analysis, a function set $g(t)$, can be written as a series of combination of scaling and wavelet function as:

$$g(t) = \sum_n c_{k0}(n) \phi_{k0,n}(t) + \sum_n \sum_{k=k0} d_j(n) \psi_{k,n}(t), \tag{3.25}$$

where $c_{k0}(n)$ and $d_j(n)$ are the inner product between $g(t)$ and the various scaling and wavelet functions, respectively.

Conceptually, the signal is iteratively decomposed into approximation (scaling function) and detailed (wavelet function) components. The relationship of coefficients $h_1(n)$ and $h_0(n)$ shows that $h_1(n)$ corresponds to a discrete time high-pass filter whilst $h_0(n)$ corresponds to a discrete time low-pass filter.

For a two-dimensional signal such as image, the scaling function is defined as:

$$\phi_{m,i,j}(x,y) = 2^{-m}\phi\left(2^{-m}x - i\right)\phi\left(2^{-m}y - j\right). \tag{3.26}$$

Similar with one-dimensional wavelet transform, the wavelet function for two dimensional signal must satisfy the orthogonal basis, which can be expressed as:

$$\psi^k_{m,i,j}(x,y) = 2^{-m}\psi^k\left(2^{-m}x - i\right)\psi^k\left(2^{-m}y - j\right), \text{ for } 1 \leq k \leq 3. \tag{3.27}$$

Since the signal is two-dimensional (x,y), the wavelet transform is achieved by performing row transform followed by column transform. In other words, the signal function is described to being decomposed into approximation component by applying scaling function or low-pass filter, G and wavelet function or detailed components by high-pass filters, H. This two-dimension wavelet transform by filterbanks is illustrated in Figure 3.6.

The scaling function $\phi(x,y)$ and three wavelet functions $\psi^1(x,y)$, $\psi^2(x,y)$ and $\psi^3(x,y)$ produce four separable bases functions of:

$$\begin{aligned}
\phi(x,y) &= \phi(x)\phi(y), \\
\psi^1(x,y) &= \psi(x)\phi(y), \\
\psi^2(x,y) &= \phi(x)\psi(y), \\
\psi^3(x,y) &= \psi(x)\psi(y).
\end{aligned} \tag{3.28}$$

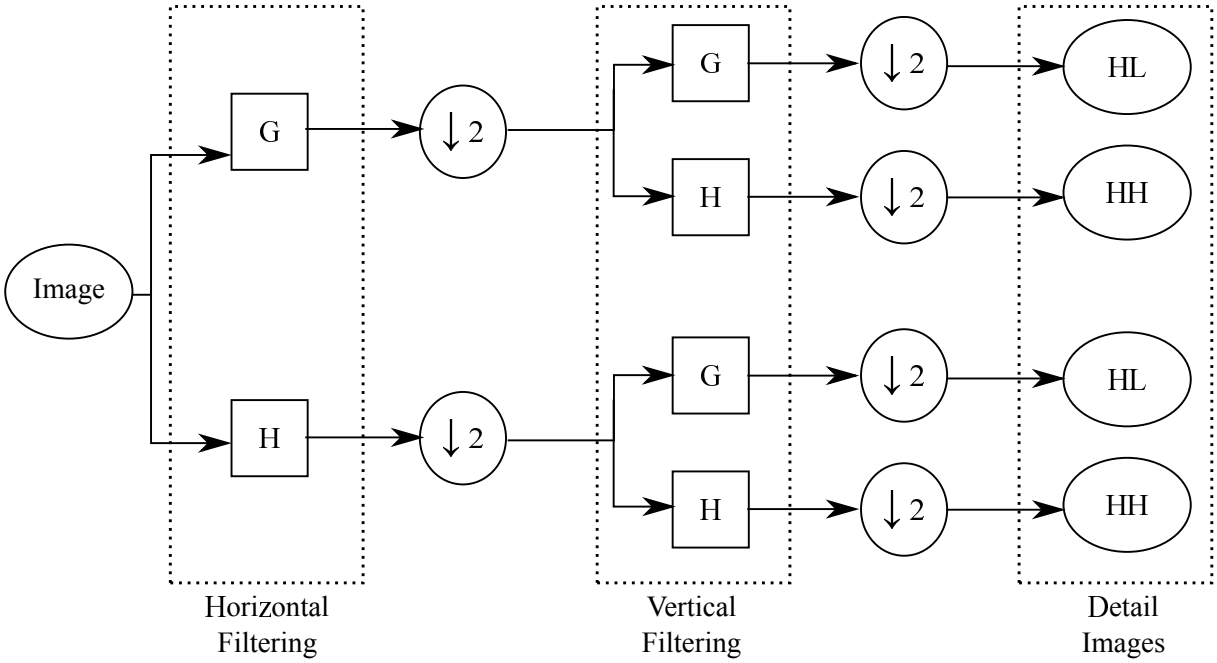


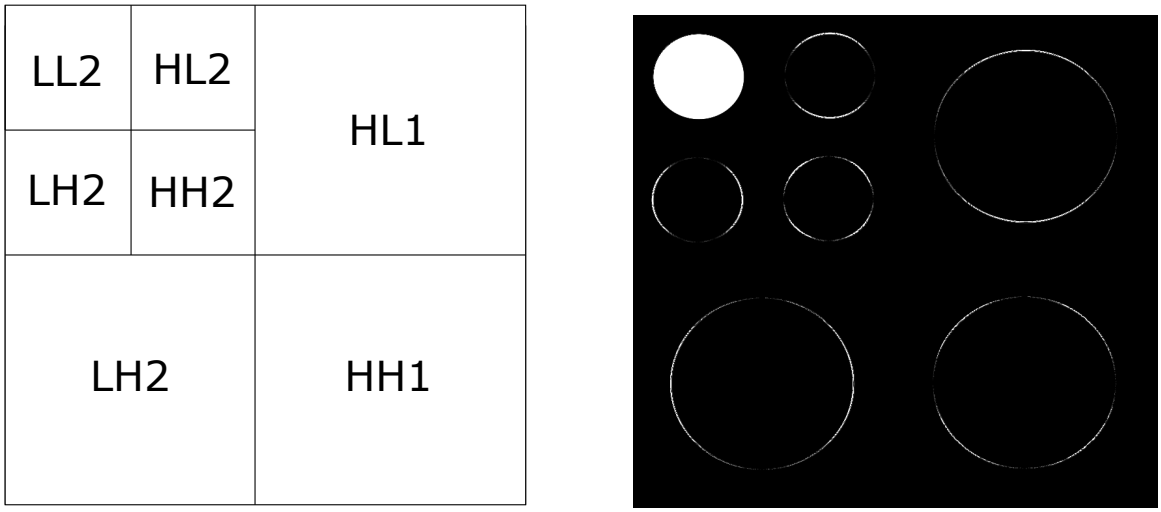
Figure 3.6: Image decomposition using 2D wavelet transform filterbanks, where G is a low-pass filter and H is a high-pass filter.

Since these functions are separable, this add up the appropriate values of wavelet transform that have separability property. As a result of this property, a wavelet transform is able to decompose images into three different orientations of detailed images, which are; horizontal, vertical and diagonal. To achieve multiple resolution, the approximation version of the detailed image is further decomposed by applying the same high-pass filter to obtain the detailed components, and the same low-pass filter to obtain the approximation detailed for the next resolutions.

One of the examples of two-dimensional wavelet transform for image decomposition is illustrated in Figure 3.7. In this example, an image of a circle is decomposed using Daubechies wavelet, db4, and is decomposed at two level of resolutions.

3.4 Steerable Pyramid Filtering Analysis

As described in previous section, it has shown that wavelet analysis is able to overcome the lack of time-frequency localisation of conventional Fourier transform.. However,



(a) (b)

Figure 3.7: Image decomposition using wavelet at two level of resolutions. (a) Each of subband contains details components, for next resolution, the approximation component is further apply to the filterbanks to get the details components at the next sequence resolution. (b) Decomposition of circle image into two level of resolution using Daubechies wavelet, db4.

there is one distinctive disadvantage of wavelet analysis when applied to image processing; the two-dimensional subbands exhibit three preferred orientations only. This is shown in Figure 3.7, where the image is decomposed into horizontal, vertical or diagonal orientations only. This implies that two-dimensional wavelet transforms are unable to differentiate between other orientations. As such, in this thesis, an approach towards multiple resolution decomposition paradigm by addressing a steerable pyramid filtering analysis will be explored.

One of the great advantages of steerable pyramid filtering is its ability to combine multiple resolution decompositions with multiple orientation differential measurements to increase the sensitivity to arbitrary orientations. However, steerable pyramid has over-completeness of $\frac{4k}{3}$ whilst wavelet has over-completeness of 1. Table 3.1 summarises the main advantages of steerable pyramid decomposition over wavelet transform and several other approaches for texture feature extraction.

	Steerable pyramid	Wavelet	Laplacian pyramid	Gabor
Self-inverting	Yes	Yes	No	No
Overcompleteness	$\frac{4k}{3}$	$1\frac{4}{3}$	1	N/A
Aliasing in subbands	No	Yes	No	Yes
Rotated orientation bands	Yes	No	N/A	No

Table 3.1: Properties comparison between steerable pyramid with wavelet.

Historically, in 1994, Simoncelli designed a flexible architecture called steerable pyramid that decomposes an image into multiple resolution and multiple orientation subbands (Simoncelli and Freeman 1995). The combination of multiple resolution decomposition with multiple orientation decomposition eliminates the aliasing problem in the wavelet decomposition (Simoncelli 1994) and, crucially, also allows users to specify their preferred orientations. The term steerable filter refers to a class of filters that are steerable at arbitrary orientations which are formed by a linear combinations of basis filters (Freeman and Adelson 1991). The detailed work of steerable pyramid is discussed in the next section.

3.4.1 Steerable Filters

The simplest example of designing the filters of steerable pyramid can be discussed using a two-dimensional Gaussian function, G as in Equation 3.29:

$$G(x,y) = e^{-(x^2+y^2)}. \tag{3.29}$$

It is well-known that the directional derivative operators of the two-dimensional Gaussian function are steerable (Freeman and Adelson 1991). Thus, the basis filters used in this steerable pyramid are the directional derivative operators of the Gaussian functions. The first x derivative of a Gaussian, $G_1^{0^\circ}$ is

$$G_1^{0^\circ} = \frac{\partial}{\partial(x)} e^{-(x^2+y^2)} = -2xe^{-(x^2+y^2)}, \tag{3.30}$$

and the first derivative Gaussian function with respect to y is the function rotated by 90 degrees:

$$G_1^{90^\circ} = \frac{\partial}{\partial(x)} e^{-(x^2+y^2)} = -2ye^{-(x^2+y^2)}. \tag{3.31}$$

The Gaussian function and its derivatives are illustrated in Figure 3.8. Figure 3.8(a) shows the two-dimensional Gaussian function, $G(x, y)$. Figure 3.8(b) and Figure 3.8(c) are the first derivatives of Gaussian function with respect to x and y, respectively.

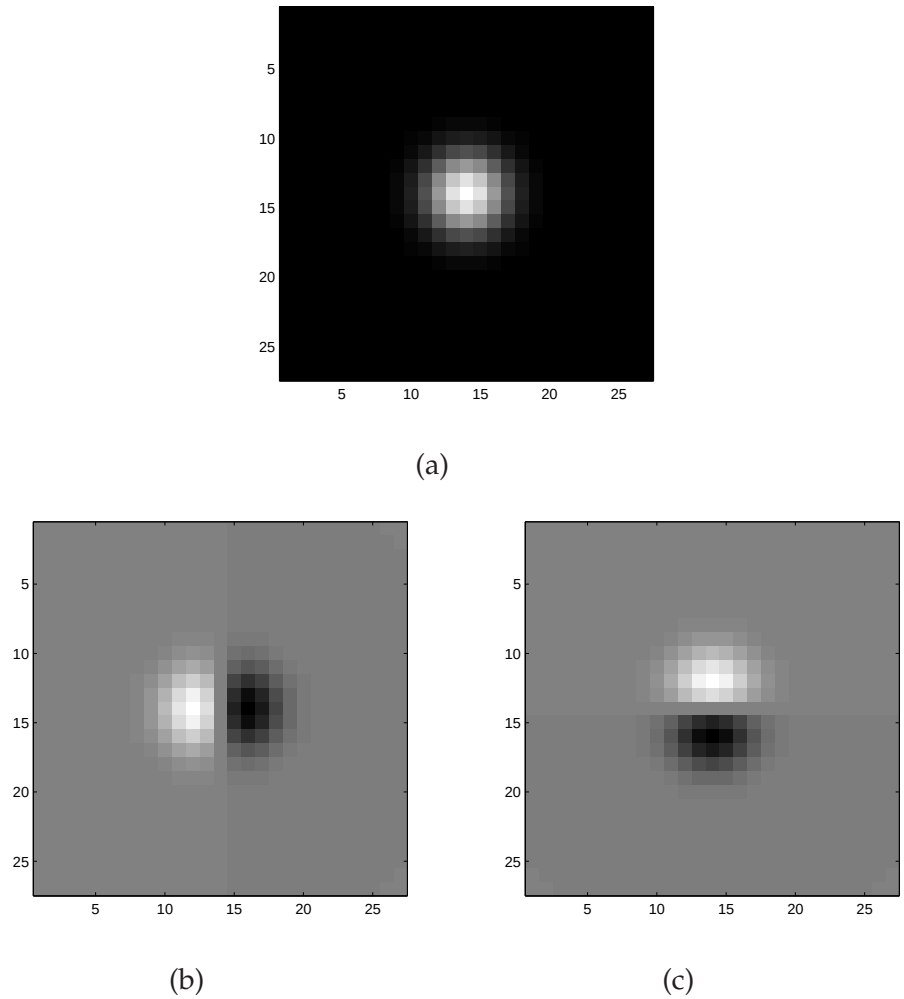


Figure 3.8: (a) Two-dimensional Gaussian function, $G(x, y)$; (b) basis function of first derivative of Gaussian function $G_1^{0^\circ}$, and (c) basis function of first derivative of Gaussian function $G_1^{90^\circ}$.

Linear combination of basis filters as shown in Equation 3.30 and Equation 3.31 can form a filter G_1 that has arbitrary rotation, θ. The linear combination function now can be written as:

$$G_1^{\theta} = \cos(\theta)G_1^{0^\circ} + \sin(\theta)G_1^{90^\circ}. \tag{3.32}$$

From Equation 3.32, the basis filters that form steerable filters, G_1^{θ} are $G_1^{0^\circ}$ and $G_1^{90^\circ}$. Whilst $\cos(\theta)$ and $\sin(\theta)$ are their interpolation functions, respectively. The filter with preferred orientation can be implemented by changing the θ to the desired orientation in Equation 3.32. As an example, Figure 3.9 shows the filter that has orientation of 60° implemented with linear combination of basis filters. As the convolution is a linear operation, the image, I filtered at arbitrary orientation can be obtained by combining image filtered at both basis functions:

$$\begin{aligned} R_1^{0^\circ} &= G_1^{0^\circ} * I, \\ R_1^{90^\circ} &= G_1^{90^\circ} * I, \end{aligned} \tag{3.33}$$

By combining the resultant filtered images from both basis filters:

$$R_1^{\theta} = \cos(\theta)R_1^{0^\circ} + \sin(\theta)R_1^{90^\circ}. \tag{3.34}$$

The result of the image convolution is illustrated in Figure 3.9(f), where the resultant filtered image at orientation of 60° is obtained by combination of filtered images by basis functions.

With the implementation of steerable filter from the derivatives of Gaussian function, an image filtered at arbitrary orientations as shown in Figure 3.10 can be obtained, where the image of a circle is filtered at multiple orientation, ranging from 10° to 90°.

The above steerable filter is formed from the derivatives of Gaussian function, where its derivatives are known to be steerable. To generalised the result in designing variety of filters, Simoncelli *et al.* has outlined few guidelines. A set of functions is defined as

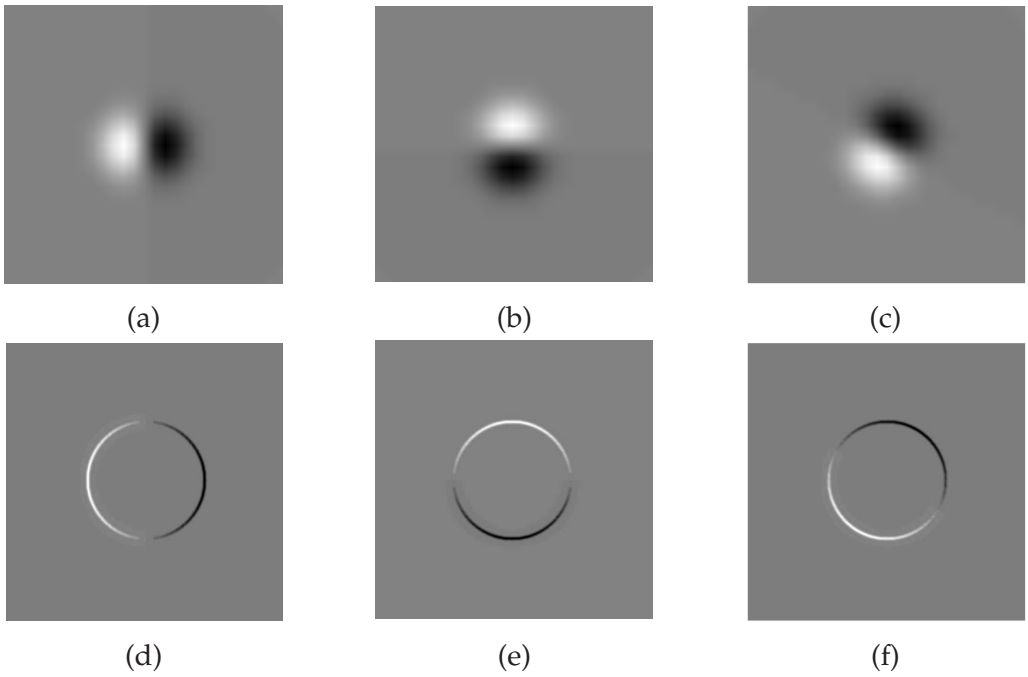


Figure 3.9: (a) Derivative filter at 0°. (b) Derivative filter at 90°. (c) Derivative filter at 60° that was form by linear combination of basis in (a) and (b). (d-e) The filtered image of circle from the filter in (a-b) respectively, and (f) The filtered image synthesised of image in (d) and (e).

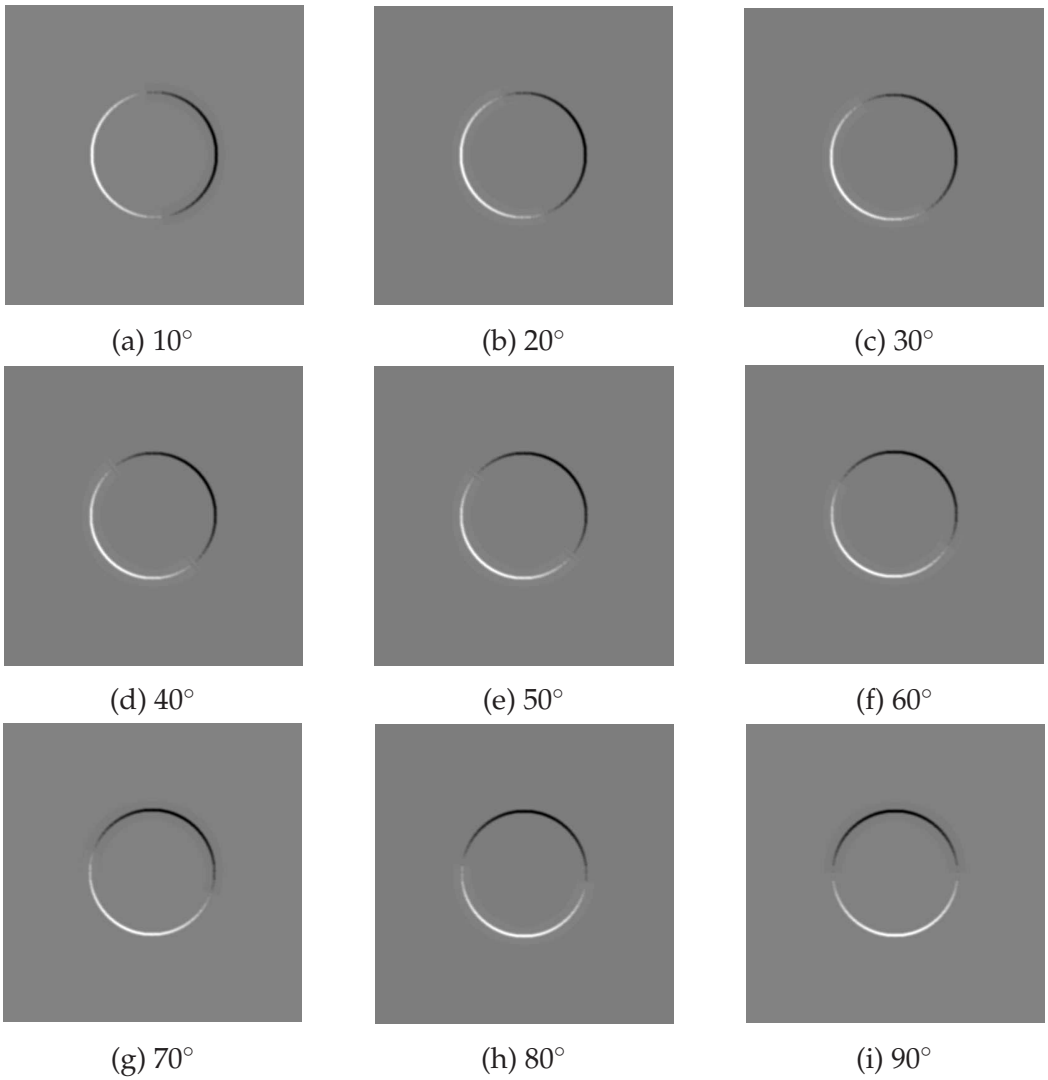
steerable filters when it can be written as a linear sum of its rotated version. To design a steerable filter, the number of basis filters that sufficient to form a steerable filter needed to be found. The interpolation function of each basis function also needs to be determined. Thus, the steering constraint is used as a guideline to find the minimum number of basis functions:

$$f^{\theta}(x,y) = \sum_{j=1}^{M} k_j(\theta) f^{(\theta)_j}(x,y), \tag{3.35}$$

where M is the number of basis filters and $k_j(\theta)$ are the interpolation functions.

The equation can be solved by working in polar coordinates, where $r = \sqrt{x^2 + y^2}$ and $\phi = arg(x,y)$. The function $f(x,y)$ is expandable in Fourier series in polar angle, ϕ:

$$f(r,\phi) = \sum_{n=-N}^{N} a_n(r) e^{in\phi}. \tag{3.36}$$



(a) 10° (b) 20° (c) 30°

(d) 40° (e) 50° (f) 60°

(g) 70° (h) 80° (i) 90°

Figure 3.10: Filtered image at 10° to 90°, using steerable filters derived from derivative of Gaussian function.

The interpolation function of each basis function can be determined from the solution of:

$$\begin{pmatrix} 1 \\ e^{i\theta} \\ \vdots \\ e^{iN\theta} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ e^{i\theta_1} & e^{i\theta_2} & \cdots & e^{i\theta_M} \\ \vdots & \vdots & \vdots & \vdots \\ e^{iN\theta_1} & e^{iN\theta_2} & \cdots & e^{iN\theta_M} \end{pmatrix} \begin{pmatrix} k_1(\theta) \\ k_2(\theta) \\ \vdots \\ k_M(\theta) \end{pmatrix}. \tag{3.37}$$

The solution of Equation 3.37 can be illustrated by steering the first derivative of the Gaussian function, as given in Equation 3.30. In polar coordinates, the equation is

written as:

$$\begin{aligned} G_1^{0^\circ}(r,\phi) &= -2re^{-r^2}\cos(\phi) \\ &= -re^{-r^2}(e^{i\phi}+e^{-i\phi}). \end{aligned} \tag{3.38}$$

From the polar coordinate form, and by the constrain sets in Equation 3.35 the first derivative of the Gaussian function has at least two nonzero coefficients. Hence, the minimum number of basis function is two. Thus, to find the interpolation function associated with the basis function, the Equation 3.39 below is solved:

$$(e^{i\theta}) = \begin{pmatrix} e^{i\theta_1} & e^{i\theta_2} \end{pmatrix} \begin{pmatrix} k_1(\theta) \\ k_2(\theta) \end{pmatrix}. \tag{3.39}$$

The interpolation functions are then found to be:

$$\begin{aligned} k_1(\theta) &= \cos(\theta), \\ k_2(\theta) &= \sin(\theta). \end{aligned}$$

The steerable filter is then formed from the linear combination of these two basis filters and their respective interpolation functions, which matches Equation 3.32.

3.4.2 Steerable Pyramid Filtering

Steerable filtering has shown to be useful in image analysis. Due to the very limited orientation sensitivity of wavelet transform, steerable filtering has the advantage of being able to analyse image at various orientations. On the other hand, pyramid filtering has an ability to analyse image at multiple resolution. The combination of multiple resolution and multiple orientation to decompose and analyse images will then be beneficial for extracting more useful features from the produced subimages. The multiple resolution and multiple orientation architecture is shown in Figure 3.11.

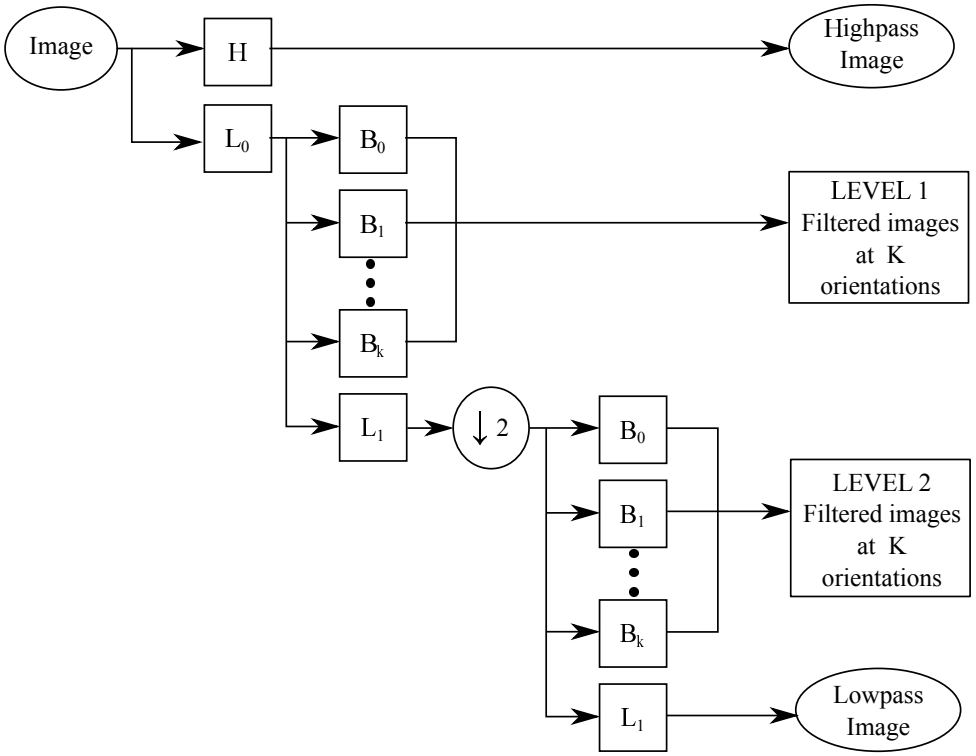


Figure 3.11: Image decomposition using steerable pyramids. In steerable pyramid filtering, the image is decomposed into highpass, H and lowpass, L_0 subbands. The low-pass filtered image is further decomposed by several sets of bandpass filters B_k to produce multiple orientation images. To achieve the next resolution, the image from lowpass subband is downsampled and the result will undergo the same set of bandpass filtering.

In steerable pyramid filtering, an image is first decomposed into low-pass and high-pass subbands by low-pass filter L_0 and high-pass filter H. The image from low-pass subband is further decomposed by several sets of bandpass filters B_K to produce sub-images at different orientations. The image from the low-pass subband will be further downsampled and the resultant image will undergo the same set of bandpass to produce multiple orientation subimages at finer resolution. The recursive algorithm is continued until the desired number of resolutions is reached.

Figure 3.12 shows the resultant subimages after an image of a circle is filtered by steerable pyramid at 3 resolutions and 4 orientations at each resolution. It is observed that the sizes of the filtered images are reduced to half as the resolution decreases. This is

because of the downsampling process, represented by a down arrow in Figure 3.11; to produce the next resolution of the image.

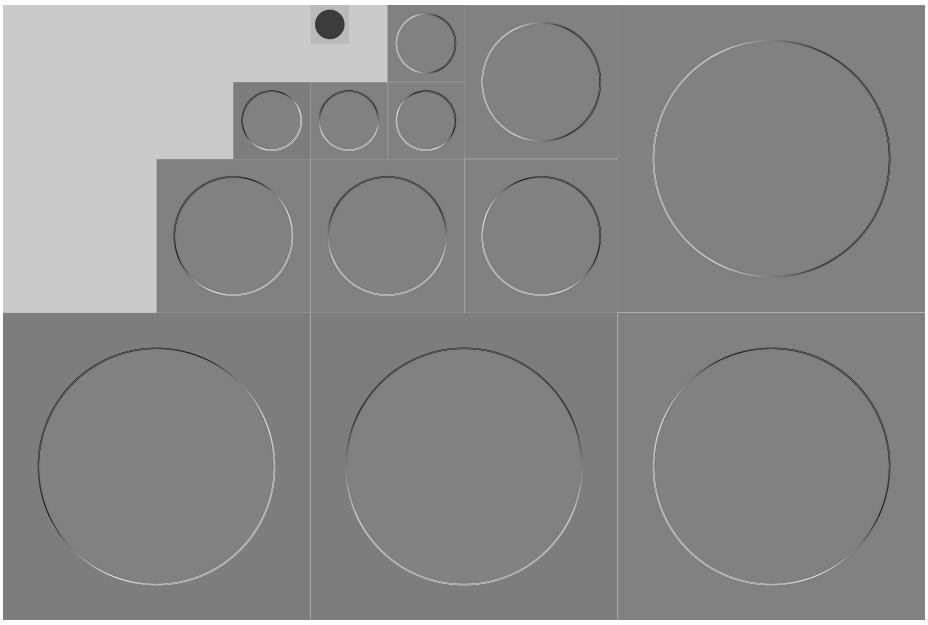
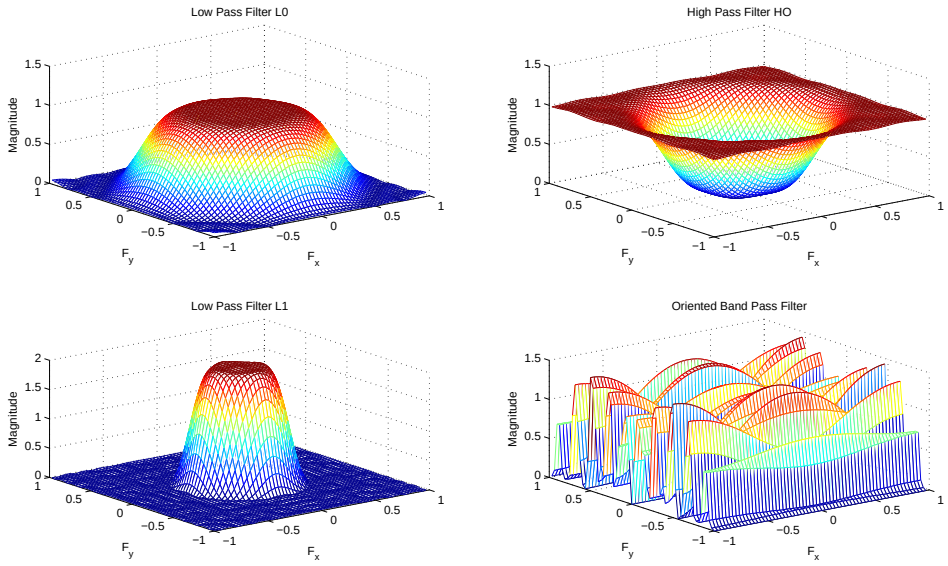


Figure 3.12: Decomposition subimages of the circle image using steerable pyramid filtering.

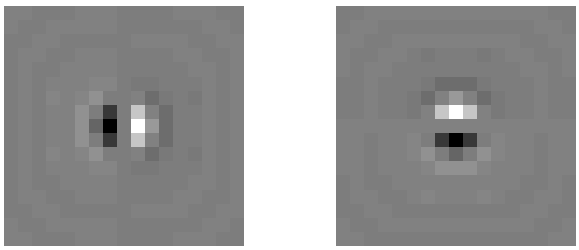
To reduce the computational load in designing steerable pyramid filters, the filters designed by Simoncelli *et al.* will be used. Based on the guidelines in designing set of low-pass, high-pass and band-pass filters, Simoncelli has designed 3 different set of steerable filters, where the steerable filter is made up from 2, 4 and 6 separable basis filters. These filters are also widely used in many other applications, such as speech recognition and classifications (El Aroussi *et al.* 2009, Benjelil *et al.* 2009).

Figures 3.13, 3.14 and 3.15 show the filters designed by Simocelli *et al.* Figure 3.13(a) illustrates the low-pass, high-pass and the set of the band-pass filters of the `sp1Filters`. The `sp1Filters` is formed by 2 separable basis filters. The frequency responses of these basis filters are shown in Figure 3.13(b). Figures 3.14(a) and 3.14(b) show the low-pass, high-pass and band-pass filters in `sp3Filters`, and their basis filters' frequency responses, respectively. The `sp3Filters` is formed by 4 separable basis filters.

Figures 3.15(a) shows the `sp5Filters`, consisting of the low-pass, high-pass and band-pass filters. In Figure 3.15(b), the plots of the frequency responses are obtained from the 6 separable basis filters.

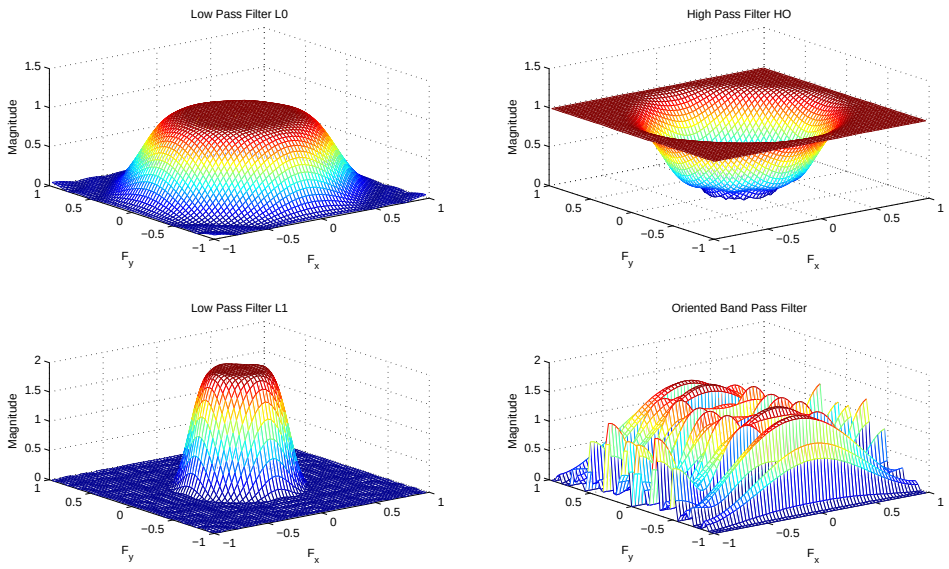


(a)

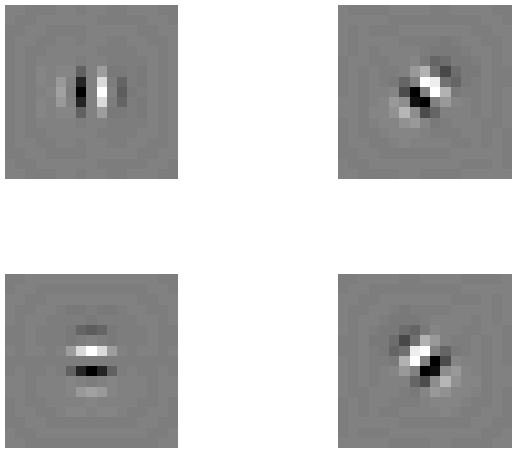


(b)

Figure 3.13: sp1Filters. (a) First row, left to right: Low-pass filter L_0, high-pass filter. Second row, left to right: Low-pass filter L_1, Set of band-pass filters, (b) Frequency response of two basis filters of sp1Filters steerable filter.



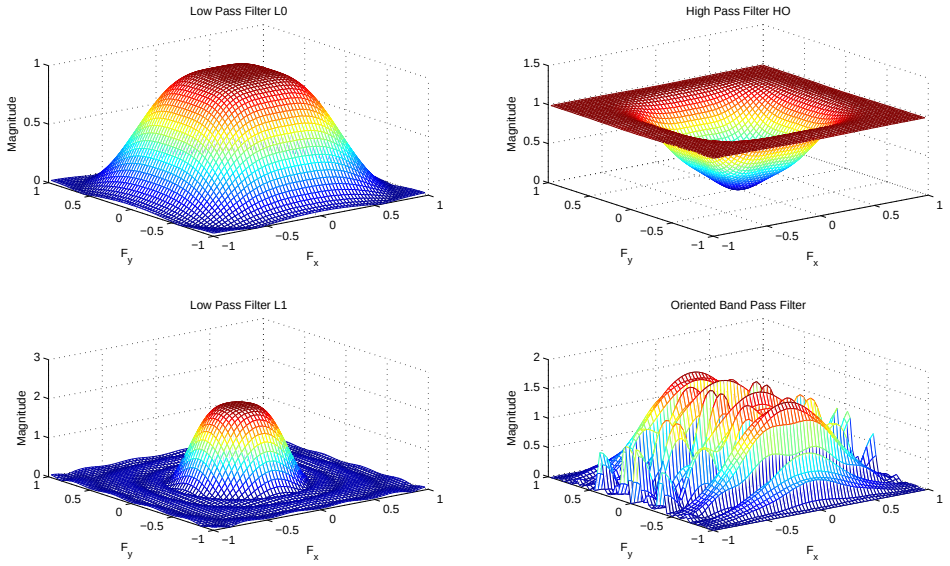
(a)



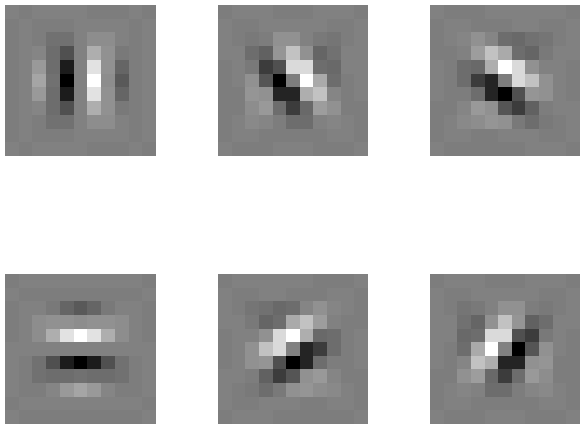
(b)

Figure 3.14: sp3Filters. (a) First row, left to right: Low-pass filter L_0, high-pass filter. Second row, left to right: Low-pass filter L_1, Set of band-pass filters, (b) Frequency response of four basis filters of sp3Filters steerable filter.

Since, the microcalcification clusters in mammogram vary in sizes, shapes and orientations, the information from arbitrary orientations may contain information that can be extracted and represented as features for the diagnosis. Thus, the concern raised



(a)



(b)

Figure 3.15: (a) First row, left to right, low-pass filter L_0, high-pass filter. Second row, left to right, low-pass filter L_1, set of band-pass filters. (b) Frequency response of six basis filters of sp5Filters steerable filter.

is to have a method that is capable of decomposing an image into multiple resolution and multiple orientation. The steerable pyramid filtering is an approach that offers both multiple resolution and multiple orientation decomposition of an image. For that

reason, steerable pyramid filtering technique is chosen to be the primary method for the feature extraction in this thesis.

3.5 Summary

This chapter discusses the commonly used methods for multiple resolution image decomposition. The relatively new technique, the wavelet transform has been adapted to many engineering application. However, wavelet transform lacks orientation sensitivity. This thesis proposes a method that has both resolutions and orientations flexibility in decomposing an image, that is the steerable pyramid filtering technique.

The steerable pyramid filtering technique has an ability to decompose images at multiple resolution and multiple orientation. The information from arbitrary orientations of microcalcification images may contain useful features that can be used for malignancy analysis. However, the analysis can only be carried out by integrating the feature extraction technique with a classifier. In Chapter 4, several types of classifiers that are commonly used for microcalcification diagnosis will be discussed.

This page is blank.

Chapter 4

Classifiers

THIS chapter discusses two types of classifiers that are commonly used as part of a CADx system. In the system, the classifier plays the role of assigning a class label to a mammogram. The classification is performed with the presentation of a set of features that have been extracted from the image into the classifier. Usually, the extracted measured features are organised into the form of a vector, and can be interpreted as a point in a multi-dimensional feature space. The classifier divides the feature space into regions corresponding to different classes by specifying the decision boundary between classes.

The rest of this chapter is organised as follows: Section 4.1 describes a general paradigm of classification steps commonly adopted in most pattern recognition systems. In the following sections, the two types of classifiers that have been used to classify the microcalcification clusters in mammograms are discussed, ie. support vector machine (SVM) in Section 4.2 and artificial neural network (ANN) in Section 4.3.

4.1 General Paradigm of Classification

Classification is a part of pattern recognition systems. Commonly, a complete pattern recognition system consists of pre-processing of the data samples, feature extraction, feature selection and lastly, classification. Classification is conducted to classify samples into different classes based on the pattern or the features measured from the samples.

Pre-processing is an essential step prior to classification to reduce the great variability in the input data. Common pre-processing steps include data scaling and segmentation processes. In data scaling step, an image is scaled to have pixels values between $[0, 1]$. Next, the image is cropped or segmented into one standard size. Feature extraction is then carried out to transform input data into new features space variable, with expectation that the new features space lends to easier categorisation of samples into associated classes. Classification step at the end of the pattern recognition system uses the extracted features to classify them into classes.

In this work, the features extracted from the microcalcification clusters as proposed in Section 3.4.2 are used for classification. The classification in CADx system assigns a detected ROI containing microcalcification clusters into two classes, benign or malignant. Since there are only two target classes, the task is called binary classification, where the classification has the target of $y_i = \pm 1, \forall 1 \leq i \leq N$.

The terms and notations that will be used from this point onwards will be briefly described. The collection of samples used for training is called training set and their associated classes are collected as target vector. Assume there is N number of samples, the training set is a large matrix $x_1, x_2, ..., x_N$, and the target vector is $y_1, y_2, ..., y_N$. Hence, for the case discussed in this thesis, the collection of samples are the feature vector extracted and the target vector is their diagnosis of benign or malignant.

Normally, a classifier can be categorised as supervised or unsupervised learning machine. A machine is called supervised machine when the training data samples are provided and used to adjust the parameters in the machine. Whilst, an unsupervised

machine does not use labeled training data, and the clustering is usually based on the patterns of the features extracted.

In this work, there are a finite number of training samples with unknown probability function describing the distribution of the features for each class. There are two ways to do the classification. First, these density functions can be estimated by using the training samples; an example of this approach would be the construction of parametric classifier such as the Bayesian Decision Theory. Alternatively, the second classification technique requires direct estimation of the density function; examples of this approach include the linear discriminant analysis or support vector machine (SVM), the k-nearest neighbour (kNN) and the back-propagation neural network (BP-NN). However, since non-parametric classifiers do not involve the estimation of parameters of statistics, it has an advantage of relatively lower computation requirements. Thus, non-parametric classifiers are chosen as the classifier in this work; ie. the SVM.

In this thesis, SVM will be used as the main classifier to evaluate the features that have been extracted by the method proposed in Section 3.4.2. Thus, the concept of SVM as a supervised classifier will be further discussed. In addition, the other commonly used classifier in this particular problem such as the neural network (NN) will also be discussed to serve as an illustrative comparison.

4.2 Support Vector Machine

Support vector machine (SVM) has been widely used in many classification applications. The SVM is an example of a supervised classifier; the machine is called supervised as it takes input data and their associated label to learn the pattern of the model, which is used to obtain the model for classification purposes. The SVM algorithm constructs a hyperplane or set of hyperplanes to separate data into their associative categories. The algorithm is first introduced in Vapnik's work on statistical learning theory (Vapnik 1995).

4.2.1 Linear SVM

Consider a simple case where the two dimensional training data $\mathcal{D}$ is to be classified into two classes, $y = \pm 1$, in region R^n (O.Duda *et al.* 2000),

$$\mathcal{D} = \{(x^1, y^1), \ldots, (x^l, y^l)\}, x \in R^n, y = \pm 1, \tag{4.1}$$

with a straight line separating function, $g(\mathbf{x})$,

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 \tag{4.2}$$

where $\mathbf{w}$ is the weight vector, the superscript T represents the transpose operator and w_0 is the bias or threshold weight.

For the classifier with separation function of $g(\mathbf{x})$ as in Equation 4.2, the data is separated by thresholding the inner product of the input, $\mathbf{x}$ with the weight, $\mathbf{w}$. If the inner product $\mathbf{w}^T\mathbf{x}$ exceeds the threshold, the data is assigned to one class and to the another class if does not exceed the threshold. For an example, if $g(\mathbf{x}) > 0$, the data falls into region, R_1. Otherwise, if $g(\mathbf{x}) < 0$, the data falls into region, R_2. However, if $g(\mathbf{x}) = 0$, the data can be assigned to either class. The $g(\mathbf{x}) = 0$ then defines the decision surface that separates the classes. If $g(\mathbf{x})$ is linear, the decision surface is a hyperplane. This is as illustrated in Figure 4.1, where the hyperplane separates the features space into two regions, R_1 for $g(x) > 0$ and R_2 for $g(x) < 0$.

From the discriminant function $g(\mathbf{x})$, the distance of $\mathbf{x}$ from the hyperplane can be found. The distance from $\mathbf{x}$ to the hyperplane can be expressed as:

$$\mathbf{x} = \mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}, \tag{4.3}$$

where $\mathbf{x}_p$ is the normal projection of $\mathbf{x}$ onto the hyperplane and r is the distance between the two classes.

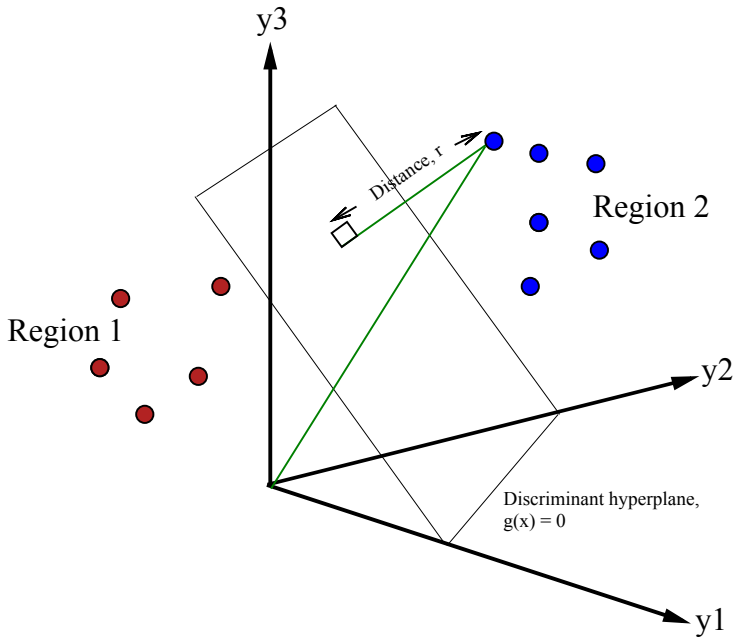


Figure 4.1: An illustrative 3-dimensional linear decision surface to separate features into two regions, R_1 for $g(x) > 0$ and R_2 for $g(x) < 0$.

Since the $g(\mathbf{x}_p) = 0$, and for $\mathbf{x}$ to be on the desired regions, ie. R_1 for $g(x) > 0$ and R_2 for $g(x) < 0$, the distance, r can then be expressed as:

$$r = \frac{g(\mathbf{x})}{\|\mathbf{w}\|}. \tag{4.4}$$

The above example is applied for the simplest case where the data is separated into two classes. The classifier is also able to separate data into multiple classes by employing linear discriminant functions. However, this may not be part of interest in this work, as the main focus is to classify the cases into two classes, benign or malignant.

The separation function $g(\mathbf{x})$ can be generalised as:

$$g(\mathbf{x}) = \sum_{i=1}^{\hat{d}} a_i y_i(\mathbf{x}) \tag{4.5}$$

or

$$g(\mathbf{x}) = \mathbf{a}^T \mathbf{y}, \tag{4.6}$$

where $\mathbf{a}$ is $\hat{d}$ dimensional weight vector and $y_i(\mathbf{x})$ can be any arbitrary functions of $\mathbf{x}$.

With the generalised function of $g(\mathbf{x})$, the polynomial discriminant function that will suit any other application of practical interests can be obtained. This generalisation offers an approach to accommodate broader classification problems including multi-category cases and also non-separable cases.

4.2.2 Support Vectors and Optimizing Hyperplane

In previous subsection, the hyperplane that separates the features space into two regions is discussed. In general, assuming the dataset is separable, there can be an infinite number of possible hyperplanes as illustrated in Figure 4.2. In this simple illustration, the data of two classes, red and blue dots can possibly be separated by hyperplane a, hyperplane b or hyperplane c. Thus, the main concern here is to find a hyperplane that separates the two classes optimally.

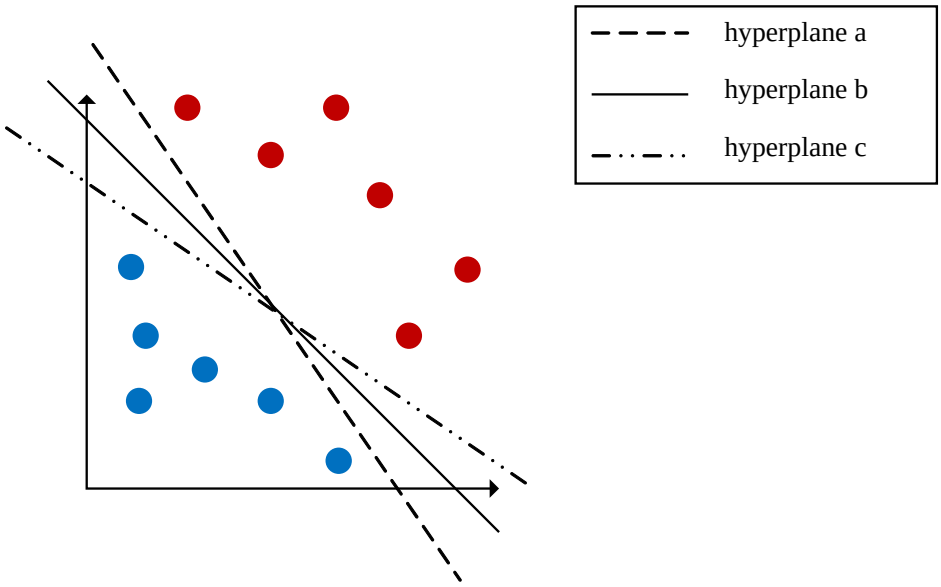


Figure 4.2: Example of three possible hyperplanes; hyperplane a, hyperplabe b and hyperplane c, to separate the data into two classes, red and blue dots.

The goal in SVM training is to find the optimum hyperplane with the largest margin; where the margin is the distance from any hyperplane to any pattern. The input data

is defined as pattern, since the SVM uses a pre-processing transformation prior to the training; where the data is mapped into higher dimensional space using appropriate non-linear mapping function, $\varphi(\mathbf{x})$. Hence, $\mathbf{y}_k = \varphi(\mathbf{x}_k)$, for each $k = 1, 2, \cdots, n$. The non-linear mapping function, $\varphi(\mathbf{x})$ will be further discussed in Section 4.2.3.

The objective of the SVM training is to find a hyperplane that separates the two classes optimally. The first step in the SVM training is identifying the support vectors. Support vectors are the critical elements in the SVM training, which are the training patterns that are closest to the hyperplane. Margin b is defined as the Euclidean distance of the support vectors from hyperplane, and mathematically expressed as (O.Duda *et al.* 2000):

$$b = \frac{g(\mathbf{y}_k)}{\| \mathbf{a} \|}. \tag{4.7}$$

Figure 4.3 illustrates the margin b that is the perpendicular distance of the hyperplane to the support vectors.

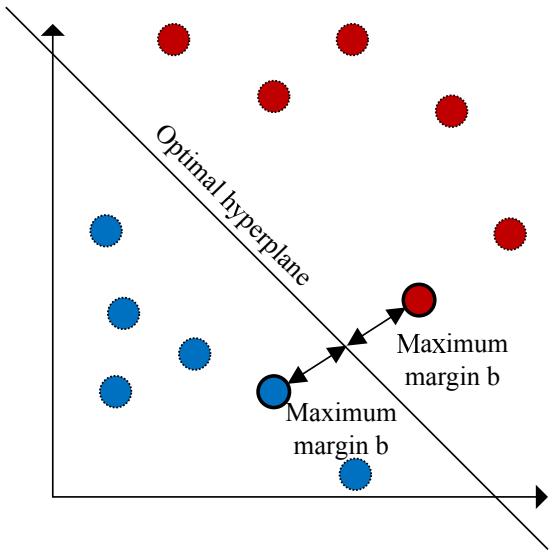


Figure 4.3: Finding optimal hyperplane margin b by training support vector machines. The support vectors are the elements drawn in solid line. The distance of support vectors in both regions is equally close to the hyperplane.

During the training, these support vectors will change the position of the hyperplane until the optimum or maximum margin is obtained. To briefly describe the training process in SVM, let us suppose to have data of $\mathbf{x}_i = \{x_1, \cdots, x_n\}$ and the labels for the two classes are $y \in \{1, -1\}$. The decision boundary that ought to classify all the data points correctly is:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \forall i. \tag{4.8}$$

In finding optimum hyperplane that separates between the two classes, the goal is to find weight vector $\mathbf{w}$ that obtains the greatest margin b. In order to maximise the margin, the optimisation problem can be solved by taking the Lagrangian formulation which seeks to minimise the following:

$$L = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{n} \alpha_i y_i(\mathbf{x}_i \cdot (w) + b) + \sum_{k=n}^{l} \alpha_i \tag{4.9}$$

where $\alpha_i \geq 0$, $i = 1, 2 \cdots n$.

The solution for the above formulation can be solved subject to Kuhn-Tucker construction constraints:

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i \tag{4.10}$$

$$\sum_{i=1}^{l} w_i x_i = 0 \tag{4.11}$$

$$\alpha_i(y_i(\mathbf{x}_i \cdot \mathbf{w}) - 1) = 0, \forall i. \tag{4.12}$$

The hyperplane obtained from the Lagrangian formulation is the optimum hyperplane. The larger the margin, the better the generalisability of the classifier. In other words, the support vectors are the most informative elements in finding the optimal hyperplane for classification task. The SVM algorithm generates weight vector in such

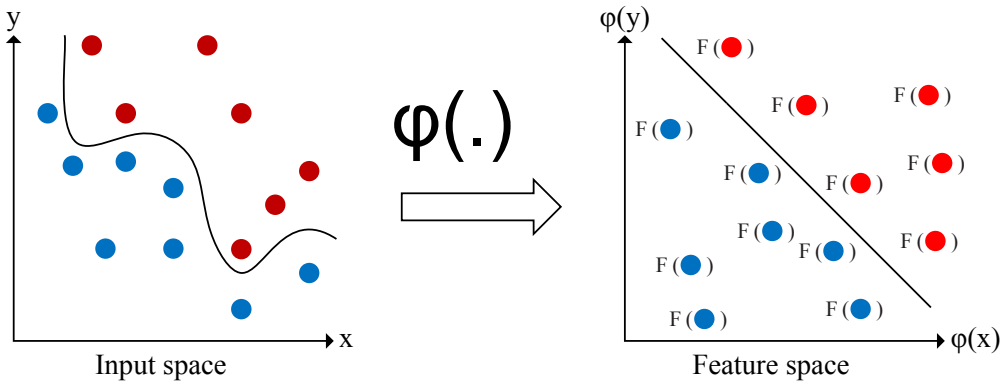


Figure 4.4: Mapping non-linearly separable data into higher dimensional features space by kernel function $\varphi(\cdot)$ so that they can be linearly separable.

a way that the weights are only affected by the support vectors. The non-support vectors do not affect the decision boundary even when they are removed. Hence, the machine is named as the support vector machine.

4.2.3 Non-Linear SVM Classification

In the above discussion, the input is assumed to be linearly separable. Because of the input can be separated linearly, then the weight vector obtained from the SVM algorithm can successfully classify the data by error-correcting procedure. However, for input which is linearly non-separable, there is no linear weight vector that can classify all input vectors correctly.

The input can be transformed by applying non-linear mapping function, $\phi(\mathbf{x})$ to each of training input. This will create features space for the training input. Then it may become possible to separate the input data by SVM in features space. Figure 4.4 illustrates how a kernel can be applied to map the input data into higher dimensional space to allow for classification. In the figure, $\phi(\mathbf{x})$ is the non-linear mapping function, where the input vectors are mapped into a new high dimensional space by the function. However, the features space now have very high dimensions. Hence, it is inefficient to calculate each of the $\phi(\mathbf{x})$.

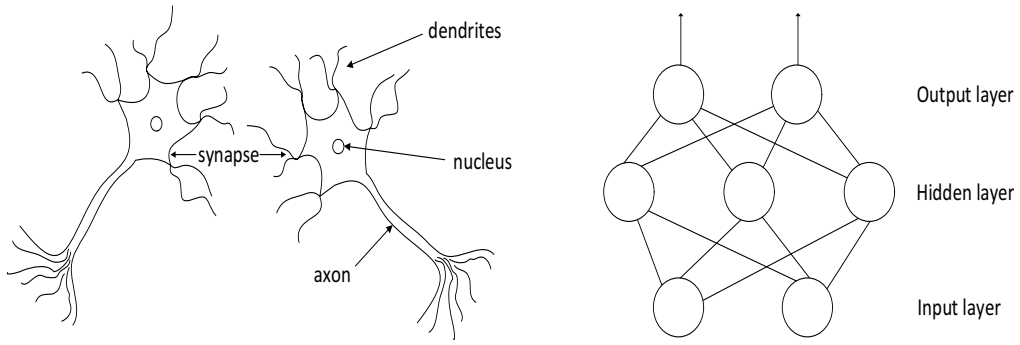
By using kernel functions to directly compute the inner products in the features space, it is possible to avoid explicit mapping computations. This is also known as the kernel trick. Common kernel functions $K(\mathbf{x},\mathbf{y})$ include:

- linear $K(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{y}$,
- polynomial $K(\mathbf{x},\mathbf{y}) = (\mathbf{x}^T\mathbf{y}+1)^p$,
- Gaussian RBF $K(\mathbf{x},\mathbf{y}) = \exp(-\frac{||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2})$.

By applying the kernel trick, the linearly non-separable data in a dimension space, $\mathbb{R}^N$ is possibly transformed into separable data in higher dimension space $\mathbb{R}^M$, where $M > N$ and N is the dimension of original features space. The kernel function $K(\mathbf{x},\mathbf{y})$ computes the inner product of the transformed SVM inputs $\mathbf{x}$ and $\mathbf{y}$ in the higher dimensional space $\mathbb{R}^M$. Hence, the computational cost is reduced as the SVM work in higher dimension space, without explicitly creating a higher dimension representation. The result is a non-linear decision boundary in dimension $\mathbb{R}^N$ corresponds to the linear decision in higher dimension $\mathbb{R}^M$.

4.3 Neural Network

The neural network (NN) is another commonly used learning machine in this particular area of interest. Historically, neural networks were inspired by human information processing systems, ie. neurons in the brain. In neural network system, the neurons are interconnected by weights, similar to the mechanism in brain system where these weights are known as synaptic weights. Figure 4.5(a) and Figure 4.5(b) show the connection of input to the node connected by weights in brain and neural network, respectively.



(a) (b)

Figure 4.5: (a) Neurons in brain are connected by synaptic weights in order to process the information received at dendrites. (b) Three layers neural network consists of input layer, one hidden layer and output layer. Each layer is connected by modified weights that are learned from training process.

4.3.1 Feed-Forward Neural Network

The minimum network architecture for neural network consists of three layers which are input layer, hidden layer and output layer. These three layers are interconnected by modified weights as shown in Figure 4.6. Each node in the input layer is connected with every node in the hidden layer. This connection also applies to the next layer, the nodes in a layer are fully connected with every node in the next layer. In addition, there is also a single bias unit that is connected to each node in the network except for input nodes.

By appending features value $x_0 = 1$, and also the weight factor w_0, when feeding an input at the i^{th} layer of the network, each neuron in the next hidden layer computes the weighted w_{ij}, sums up of its input and forms net activation net_j (O.Duda *et al.* 2000).

$$\text{net}_j = \sum_{i=1}^{d} x_i w_{ji} + w_{j0} = \sum_{i=0}^{d} x_i w_{ij} \equiv \mathbf{w_j^t x}, \tag{4.13}$$

where x_i is the input at i^{th} node and w_{ji} is weight that connects the i^{th} node to the j^{th} node in next layer.

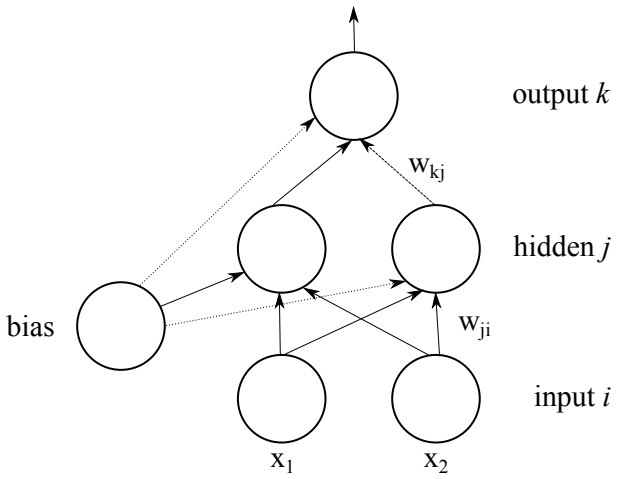


Figure 4.6: Three-layer feed-forward neural network. The first layer is directly connected to the input data. Each of input node, i is connected to every hidden node j in the hidden layer by the weights $\mathbf{w}_{ji}$. These nodes are activated and connected to the output node k by the weights $\mathbf{w}_{kj}$. The output node is activated by the activation function at the top of the network, to give the output of the network.

The term *hidden* is used because the output of the layer is not observable. The output of a hidden layer will be passed as an input to the next hidden layer. An output at a node y_j will be a function of a net activation, net_j, ie. $y_j = f(\text{net}_j)$. The function $f(\text{net}_j)$ is the activation function. There are several types of activation function. One of the commonly used activation functions is the sigmoid, $\sigma(x)$, which is written as in Equation 4.14:

$$f(\text{net}) = \sigma(\text{net}) \equiv \frac{1}{(1 + e^{-\text{net}})}. \tag{4.14}$$

The top layer of the network is the output layer. The nodes at the output layer represent the overall output of the network. For binary pattern classification or two-class classification problem, there is usually one node at the output layer. The node will be activated to produce an output of the network. Each of the output nodes will calculate its net activation similar to the hidden unit's net activation, which is mathematically expressed in general as (O.Duda *et al.* 2000):

$$\text{net}_k = \sum_{j=1}^{n_H} y_j w_{kj} + w_{j0} = \sum_{j=0}^{n_H} y_j w_{kj} \equiv \mathbf{w_k^t y} \tag{4.15}$$

where y_k is the output at k^{th} node and w_{kj} is weight that connects the node at j^{th} node to the k^{th} node in output layer. n_H denotes the number of hidden units. For $k = 0$, y_0 is the bias unit. Similarly, this equation is applied for the case where the bias unit always has a value of $y_0 = 1$.

4.3.2 Back-Propagation Neural Network

Neural network with back-propagation learning scheme (BP-NN) is categorised as one of the supervised classifiers. The back-propagation algorithm is employed to train the network. The back-propagation procedure involved in the training is briefly described as follows; an input vector is presented at the input layer, and the weights of each connected node are normally set randomly at initial, and the output at the output layer is determined. The squared difference between the actual output and the desired target is the sum squared error. During the learning process, the weights will be adjusted to minimise this error. This process is iterated until the minimum error is achieved. When the error is minimised, the output is said to approximately match the desired output. Hence, the network is now trained and ready for classification task. Figure 4.7 briefly illustrate the process of BP-NN.

Since the error is used to adjust the weights, the error criterion can be expressed as a function of weight. Thus, the error criterion to be minimised, $J(w)$ is the sum of the squared difference between the output and the desired target, and mathematically $J(w)$ is expressed as (O.Duda *et al.* 2000):

$$J(w) \equiv \frac{1}{2}\sum_{k=1}^{c}(\mathbf{t}_k - \mathbf{z}_k)^2 = \frac{1}{2}||\mathbf{t} - \mathbf{z}||^2, \tag{4.16}$$

where $\mathbf{t}$ is the desired target and $\mathbf{z}$ is the actual output; $\mathbf{w}$ is the weights in c-layer network.

Commonly, gradient descent algorithm is used for error correction criterion. Using this algorithm, the weights are changed in a direction towards the error reduction. The

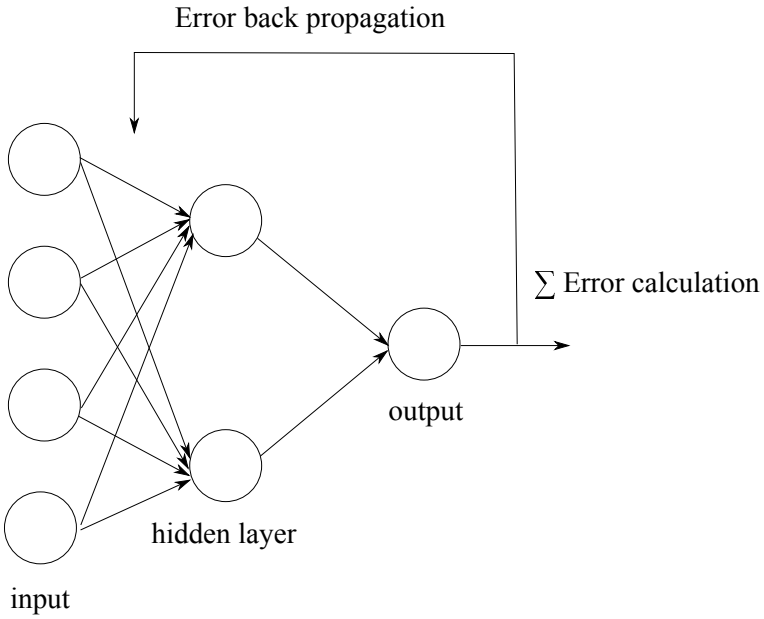


Figure 4.7: Three-layer neural network with back-propagation learning, an input is presented at the input layer, and pass through the hidden layer to determine the output. The error between the output and the target is calculated and is back-propagated to adjust the weight by minimizing the error.

change in weights is given by:

$$\Delta \mathbf{w} = -\eta \frac{\delta J}{\delta \mathbf{w}}, \tag{4.17}$$

where η is the learning rate.

The learning rate η must be carefully chosen so that the error continues to reduce. Moreover, since the error criterion in Equation 4.16 is never negative, the learning rate guarantees that the learning will stop when a pre-specified requirement is met. The learning is an iterative process by updating the weights at each loop in m iteration, expressed as:

$$\mathbf{w}(m+1) = \mathbf{w}(m) + \Delta \mathbf{w}(m). \tag{4.18}$$

However, the use of gradient descent algorithm in back-propagation has a limitation because it has a tendency to converge to local minima. This means that the result of the network depends on the initialization of the weights. Since the initial weights in neural network are set randomly, the gradient descent may get stuck at local minima

which could lead to a sub-optimal network performance. The training using back-propagation is also computationally expensive as the convergence process in the network is slow and is often sensitive to the choice of learning rate.

4.4 Summary

This chapter discusses the two most common classifiers used for classifications of microcalcification in mammograms; ie. SVM and ANN. Both classifiers are parametric, and use supervised learning algorithm. SVM is chosen to be the main classifier in the thesis because of its powerful properties that able to adapt itself to suit most of the problems of practical interest.

On the other hand, ANN has been used in a wide range of applications. However, the implementation of ANN architecture is a heuristic process. Their parameters such as number of hidden layers, number of nodes in each layers and initial weights have to be properly chosen and adjusted to ensure optimal result. This requires experience and skills to predict an appropriate ANN architecture that is applicable to solve the problem of the practical interests. However, relatively new approach called deep belief networks has been discovered. This approach offers property to overcome some shortcomings in conventional ANNs and will be discussed in Chapter 5.

This page is blank.

Chapter 5

Deep Belief Networks

THIS chapter discusses one of the relatively recent developments in pattern recognition, deep learning networks. This approach is attractive due to its deep architecture that is capable of extracting high-level representations of complex structures. Furthermore, most conventional machine learning techniques such as multi-layer perceptron (MLP) contain only single or few layers of non-linear features transformation. However, for example, in human sense processing mechanism such as vision and speech, multiple layers transformation information are required to extract the complex structures for cognitive purposes. In this context, a deep learning is acknowledged as the state of the art for machine learning. Thus, this motivates us to introduce the new method in the CADx system in mammography diagnosis specifically for classification of microcalcification clusters.

The rest of this chapter is organised as follows: Section 5.1 describes an overview of learning a deep belief network (DBN), followed by the main building block of DBN, called the restricted Boltzmann machine (RBM). The learning structures of DBN and RBM are discussed in Section 5.2. In Section 5.3, different types of DBN architecture and their functions are described. Finally, in Section 5.4, several applications of DBN in pattern recognition field are discussed.

5.1 Deep Network

The deep learning network concept emerged from the research work on the conventional artificial neural network, ANN. An example of architecture for the deep learning network is the multi-layer perceptron (MLP) with many hidden layers for the transformation of input layer to produce a result in the output layer. A more natural example with similar mechanism is the human visual system. For an example, to classify a physical object into classes visually, the underlying causes of the image, such as surface depth, orientation, boundary and reflectance must be identified first (Hinton 2010). The visual classification can be achieved since the cortex of the eyes contain an extremely complicated non-linear system. Hence the idea is to replicate such physical object using a computational network with non-linear transformations requires a hierarchical structure. At each hidden layer in the hierarchical structure, the activations of the lower layer become the input for the next sequence higher layer. The nodes in every layer also act as features detectors that capture features from lower level and pass them onto the higher level.

However, a concern over deep learning networks quickly arose in the way they learned the networks for solving pattern recognition task. A technique of back-propagation was introduced in 1980s. The technique of back-propagating error derivatives was introduced and became popular to solve the problem of learning multi-layers network. MLP is an example of this technique. In the back-propagating error derivatives learning, the weights between the layers of nodes are optimised by minimizing the error between the output of the network with the desired output. The error derivatives are then propagated backwards using the computed weights between the hidden layers. The weights are then updated to reduce the error. This process is performed recursively. However, the learning algorithm generally failed for networks with more than three layers, ie. network with more than one hidden layer (Hinton 2007, Bengio 2009). When the number of hidden layers increases, the network would often get stuck at local minima that decreases its performance. This may possibly be contributed from

the randomly initialisation weights in the gradient descent method in the recursive network.

An efficient learning algorithm for deep network was introduced in 2006 by Hinton *et al*. The network is named as deep belief network (DBN). This network is a probabilistic generative model that composed of multiple layers of stochastic binary nodes. The generative model network of DBN overcomes the limitations that occurred in the conventional discriminative neural network (eg. feed-forward neural network, FF-NN). In FF-NN, the weights are trained so that the network directly models the probability of the output and estimates $p(y|x)$ right away.

On the other hand, the weights in generative model network of DBN are trained so that the networks model the joint probability distribution between the input and the output. Hence, for a given input x that is used to predict an output y, the generative model network learns $p(x, y)$, from which it can further be used to estimate $p(y|x)$ and $p(x|y)$. With these properties, the efficiency and simplicity of gradient method in back-propagation technique can be retained for the use of modelling the joint relationship between the input and output (Hinton 2007).

In DBN, the top two layers are connected in an undirected and symmetric way. Whilst, the lower layers are connected in a directed sigmoid. The lowest layer represents an input data vector. In DBN learning, the training begins with the layer that is directly connected to the input vector. The activation nodes or the output of this layer are then act as the input to the next sequence layer. An illustrative of the connections in DBN is shown in Figure 5.1.

The learning algorithm of the network is based on greedy layer-by-layer training, which means the layers are learned one at the time. Using this approach, the weights in DBN are optimised with a time complexity that varies linearly with the depth of the network (Yu and Deng 2011). A DBN architecture is achieved by stacking up its constituents, called restricted Boltzmann machines (RBM) which will be discussed in Section 5.2.

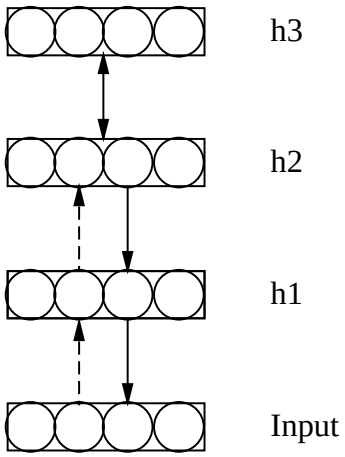


Figure 5.1: Deep belief network.
DBN with top two layers that has undirected connections, whilst lower layers are connected in a directed sigmoid that are represented by the solid top-down. The dashed bottom-up is not part of generative model in training DBN. The dashed bottom-up are used for inference.

5.2 Restricted Boltzmann Machine (RBM)

The main building block of DBN is a two-layer architecture called the restricted Boltzmann machine (RBM). The RBM is one of the Markov random field (MRF) networks. The RBM network is a structure of bipartite graph that consists of 2 layers of variables, the visible stochastic units, $V = \{v_i\}$, which are connected to the hidden stochastic units, $H = \{h_j\}$.

The connection of the network is illustrated in Figure 5.2. Normally, all visible units are connected to all hidden units. However, intra-layer connections are forbidden in this structure, ie. no connection between visible-visible units and hidden-hidden units, and hence the structure is named restricted Boltzmann machine. The visible units are observable and commonly represent the pixels of an image, which may be possibly have binary (Bernoulli) or real-valued (Gaussian) data.

The joint connections between each visible and hidden units are based on energy-based undirected generative model. The energy is represented by a probability distribution of the visible units and hidden units. The probability distribution over the visible units

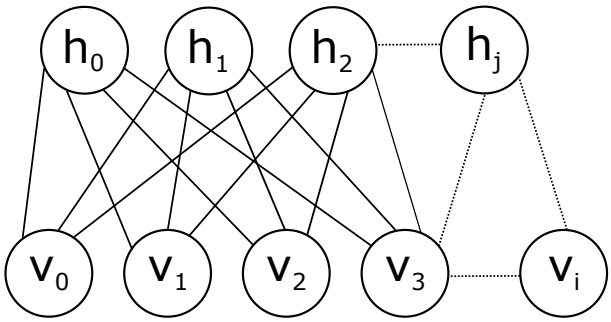


Figure 5.2: Structure of RBM, top layer of RBM consists of hidden variables $H = \{h_j\}$ and the bottom layer consists of visible variables $V = \{v_i\}$

and the hidden units are defined by the weights and biases via energy function given by (Hinton *et al.* 2006):

$$E(v,h) = -\sum_{i \in \text{visible}} a_i v_i - \sum_{j \in \text{hidden}} b_j h_j - \sum_{i,j} v_i h_j w_{ij}, \tag{5.1}$$

where v_i, h_j are binary state of visible unit i and hidden unit j, and w_{ij} is the weight between them. a_i and b_j are their biases.

In the case of real-value data, the binary visible unit can be replaced with linear Gaussian unit. The energy is now defined as (Hinton 2012):

$$E(v,h) = -\sum_{i \in \text{visible}} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{j \in \text{hidden}} b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij}, \tag{5.2}$$

where v_i, h_j are linear Gaussian state of visible unit i, and binary state of hidden unit j, respectively. w_{ij} is the weight between them, and a_i, b_j are their biases. σ_i is the standard deviation of Gaussian for visible unit i.

The joint distribution that the model assigns over visible vector $\mathbf{v}$ and hidden vector $\mathbf{h}$ with a given parameter θ is:

$$p(\mathbf{v}, \mathbf{h}; \theta) = \frac{\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h};\theta)}}{\sum_{\mathbf{u}} \sum_{\mathbf{h}} e^{-E(\mathbf{u},\mathbf{h};\theta)}}. \tag{5.3}$$

The conditional probabilities over visible units can be directly calculated as follow:

$$p(v_i = 1|\mathbf{h}) = \sigma\left(\sum_i w_{ij}h_j + b_j\right) \tag{5.4}$$

$$p(h_j = 1|\mathbf{v}) = \sigma\left(\sum_j w_{ij}v_i + a_i\right) \tag{5.5}$$

where $\sigma(x) = 1/(1 + \exp(x))$.

The probability expression in Equation 5.4 shows that to activate one visible unit given the states of hidden units is independent of other visible units. Likewise, the probability expression in Equation 5.4 shows that to activate one hidden unit given some states of visible units is independent of other hidden units. With these properties, the sampling in RBM is efficient as all hidden units can be sampled simultaneously followed by sampling all visible units simultaneously. This is known as an alternating Gibbs sampling, where the state of the units in one layer is updated in parallel with a given state of the units in another layer (Hinton *et al.* 2006). A one full step of Gibbs sampling is equivalent to an updating a hidden layer, $\mathbf{h}$, with a given input layer $\mathbf{v}$, then by similar steps, updating $\mathbf{v}$ with given $\mathbf{h}$.

To achieve maximum likelihood learning in RBM, the correlation $\langle v_i^0 h_j^0 \rangle$ is calculated for each weight w_{ij}. Here the data vector is clamped on visible units and the hidden states are sampled from their conditional distribution. By using Gibbs sampling and Markov chain, the correlation $\langle v_i^\infty h_j^\infty \rangle$ can be determined once an equilibrium is reached. This is the correlation defined by the model. By performing this learning, the update rule for weight can be obtained from the derivative of the log-likelihood, ie. the difference between two correlations obtained from the training data and defined by the model:

$$\frac{\partial \log P(\mathbf{v};\theta)}{\partial W} = \langle v_i h_j \rangle^0 - \langle v_i h_j \rangle^1 \tag{5.6}$$

where $\langle v_i h_j \rangle^0$ is the expectation distribution in training data and $\langle v_i h_j \rangle^1$ is the expectation distribution defined by the model.

During the learning, the expectation distribution can be calculated by clamping the visible units at the data vector, then sampling the hidden units, and then sampling the visible units again. This procedure is repeated many times by running a Markov chain. This is illustrated as in Figure 5.3. This procedure is known as contrastive divergence (CD).

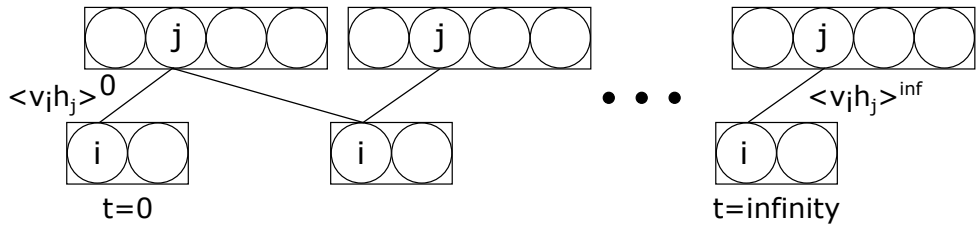


Figure 5.3: Alternating Gibbs sampling in Markov chain for maximum log-likelihood learning in RBM. During the learning, all the hidden units are updated simultaneously given the input vector in the lower layer. Then all the visible units (directly connected to the input vector) are updated simultaneously given the current hidden units. This is equivalent to one full step of Gibbs sampling. By run the Markov chain, the sampling can be performed until the equilibrium condition is achieved. The difference of the correlations at the beginning and the end of the chain is measured. This is the learning signal for weights updating in the network.

The use of CD learning in RBM has shown to be practically efficient. Other researchers have reported successful results using CD in RBM for applications such as modelling formation of topographic maps by Welling *et al.* in 2003, de-noising natural images by Roth and Black in 2005, and also rapid document retrieval by Welling *et al.* in 2005 (Hinton *et al.* 2006).

5.3 DBN Architectures

By stacking the layer-by-layer learning RBMs from bottom-up, a deep network is built. The learning procedure in DBN started with the learning in the first layer that is directly connected to the vector input, eg. pixels of an image. After the first layer is learned, the activation probabilities of its hidden layer will be the visible layer for the

second RBM layer. With these visible layers, the RBM is learned using the previously described CD procedure and the activation probabilities at its hidden layer will be the visible layer for the third RBM layer, and so on. This is as shown in Figure 5.4.

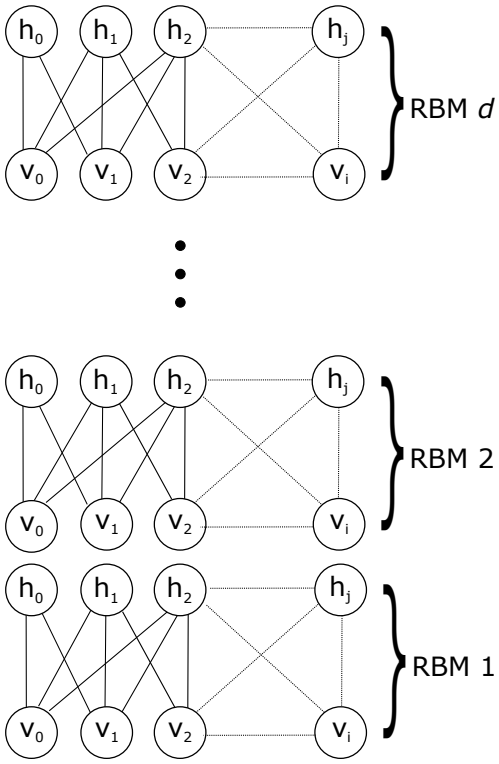


Figure 5.4: A *d*-layer DBN, made up from stacking *d* number of RBMs, with the first visible layer directly connected to the input vectors. The activation of hidden units in hidden layer of RBM 1 will be the visible layer for RBM 2 and so on.

A DBN can act as a generative learning model or discriminative learning model or combination of both models. The lattermost is commonly known as hybrid training. Generally, the type of DBN is defined by the top layer of the DBN. To act as a discriminative learning model, a DBN is learned to predict an output given the output. Thus, the top layer of the model is the output. For a supervised learning, the output of the DBN is compared with the desired output, and perform a back-propagation to update the weights and minimise the error. This is similar with the conventional neural network as supervised classifier except for the weights being obtained from CD procedure in DBN. Thus, the limitation in conventional neural networks where the network is always stuck at poor local minima can be prevented, especially for deeper network. This

is because the initial weights in DBN network are not randomly set as in conventional neural networks. Using gradient descent to update weights in the network will work well if the initial weights are close to good solution. Hence, the generative weights in DBN that are produced by greedy layer by layer training are able to overcome such limitation. This architecture of DBN is known as back-propagation DBN (BP-DBN) which as illustrated in Figure 5.5.

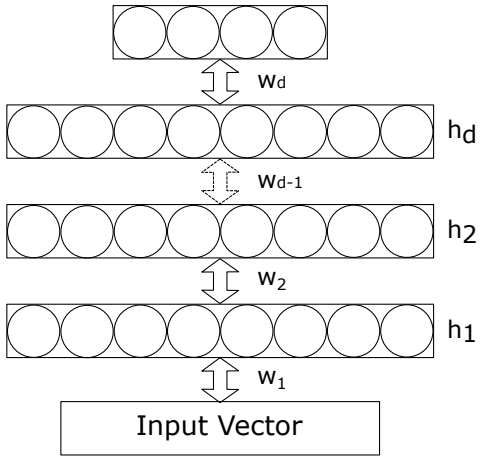


Figure 5.5: A *d*-layer back-propagation DBN. RBMs are stacked to build a DBN and learned layer-by-layer. The top layer is the output layer, which is then compared with the desired output. The error derivative is calculated, and weights are updated to minimise the error. The mechanism is similar with the conventional neural network, except for the weights are obtained from CD procedure instead of randomly initialised.

A hybrid training is a combination of generative and discriminative DBN. One of the example is called associative memory DBN (AM-DBN). This type of DBN has top RBM layers that defines a joint distribution of between hidden units and visible and class labels. In addition to that, in hybrid training, the weights are also fine-tuned in the network by back-propagation algorithm. The architecture of AM-DBN is illustrated as in Figure 5.6.

5.3.1 DBN Parameters Tuning

In training model using RBM, one of the parameters that needs to be carefully chosen is the number of nodes in the hidden layer. Hinton has written extensive guidelines

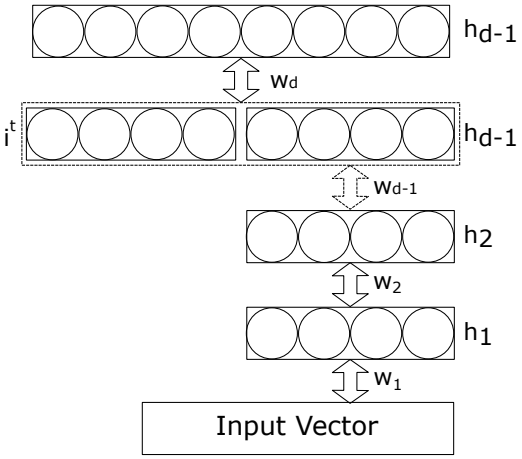


Figure 5.6: A d-layer associative memory DBN. The RBMs are connected and learned greedily. The first layer of the first RBM is directly connected to the input vectors. The activation of hidden layer of the first RBM is connected and act as visible layer to the next sequence RBM. In addition, the top two RBM will perform associative-memory DBN. The layer of class labels, i^t is cascaded with the first RBM of the top two layer of DBN. The top two layer of DBN models the joint probability of the activation of hidden units of previous RBM, which are their visible units together with class labels.

for choosing the number of hidden nodes based on a variety of factors, including the number of bits of the data vector of a grayscale image, which may be estimated using entropy measurements,

$$H = -\sum_{i=0}^{2^n-1} P(a_i)I(a_i) = -\sum_{i=0}^{2^n-1} P(a_i)\log_2 P(a_i), \tag{5.7}$$

where $P(a_i)$ is the probabilities of pixel a_i occurs in the image. In the guidelines, the estimation number of hidden nodes is a smaller order of number of bits multiply by number of training cases (Hinton 2012).

With these provided guidelines, the number of hidden nodes that will used for micro-calcification classification can be estimated. However, other factor such as computer's memory limits the algorithm to have huge number of hidden nodes. Hence, the number of hidden nodes at the first layer is first estimated using the equation but subject to the capability provided by the computer machine.

5.4 Previous Work in DBN

A deep network has attracted many researcher in various applications in pattern recognition area such as classification task in vision, audio and language. One of the successes of DBN is in handwritten character recognition (Hinton *et al.* 2006). In this work, Hinton used 44,000 training MNIST (Mixed National Institute of Standards and Technology) (Lee *et al.* 1998) images and divided them into 440 batches. Each batch contains ten examples of each digit, hence each batch has 100 training images. Associative-memory DBN is used for the training, where the class of each image is attached as part of the input at the top layer of the DBN. The validation test on the trained network had an error rate of 1.39%. The work proves that a deep, densely connected belief network can be trained one layer at a time.

Shusen *et al.* used a semi-supervised learning algorithm called discriminative DBN (DDBN) in image classification problems. The database used in this paper is a set of handwritten digits images obtained from MNIST dataset. However in this paper only subset of 10,000 random images are used. Similarly, layers of RBM were used as building blocks to construct DDBN. At the top layer, the weights parameter is fine tuned to maximise separability among labelled data using an exponential loss function. The results for the work show that DDBN obtained the lowest classification error-rate when compared to other classifiers (Shusen *et al.* 2010), where the lowest error rate is 14.13% compared to conventional neural network of 22.68% and SVM of 23% error rate.

In another application, DBN is used for dimensionality reduction. In this work, an encoder is created to transform a high dimensional data into a low dimensional code. And in similar way, a decoder to encode and recover the low dimensional code into the full data was implemented (Hinton and Salakhutdinov 2006). The whole system consisting of an encoder and a decoder is called an autoencoder. In this work, two types of input, binary and gray scales were used. For the binary images, MNIST handwritten images were used. The images were encode by five layers DBN encoder with

layer size of 784-1000-500-250-30 network. For grayscale images, Olivetti face data set were used and were encoded with a 625-2000-1000-500-30 network. The results from both datasets showed better reconstruction in terms of visualisation compared to when reducing the dimension using principal component analysis (PCA) method.

These previous work shown promising results in using DBN for either classification or dimensionality reduction. Thus, in this thesis, the use of DBN in classifying microcalcification clusters in mammogram will be further explored. Since the microcalcification clusters images are complex structures, the ability of DBN to train deep networks to extract more features might be useful. In addition, by using DBN, the dimension of the microcalcification clusters images can be reduced to obtain a low dimensional code, that presumed to contain the compressed information of the clusters for classification.

5.4.1 Autoencoder

An autoencoder consists of two processes, encode and decode. From the original image, an encoder will encode the image into a set of code. This code can later be decoded by a decoder to retrieve the original image. An autoencoder is an interesting technique as it can compress data, which are useful for dimensionality reduction purposes. Work by Hinton *et al.* has shown that DBN can build an autoencoder. In the next section, the experiment by Hinton *et al.* to autoencode binary images of MNIST handwritten digits is recreated. The purpose is to illustrate how this approach can be modified to the classification of microcalcification clusters.

The set of database is hand-written digits obtained from MNIST database. This database is also used in the Hinton's original experiment. The database contains 6000 training samples and 1000 testing samples of each 10 digits (0–9). The image is binary and centered, with a size of 28×28 pixels. Figure 5.7 displays few handwritten digit samples randomly selected from the MNIST database.

The autoencoder used three layers to encode the image, one code layer and three layers to decode with the transpose weights. Since the MNIST images are binary, the visible

Figure 5.7: Random sample of MNIST handwritten digits. The images were generated from MNIST database (Lee *et al.* 1998).

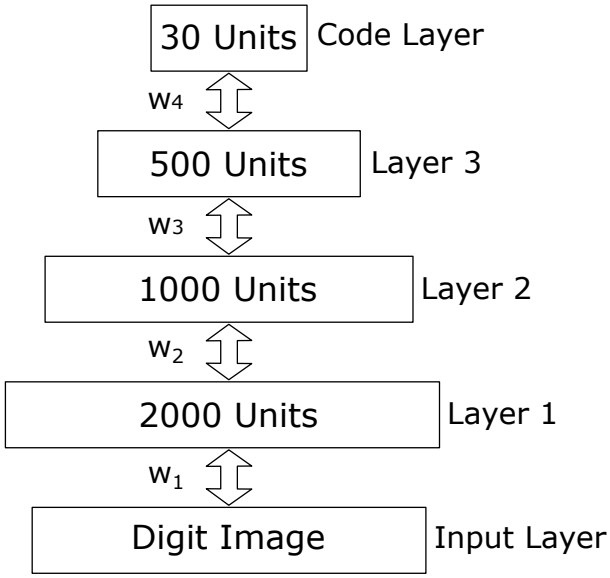


Figure 5.8: Autoencoder DBN for MNIST digits reconstruction procedure. The DBN contains 3 RBM layers and 1 code layer. All 3 layers are Bernoulli-Bernoulli RBM and a code layer is Bernoulli-Gaussian RBM. The procedure involves pre-training each RBM layer separately before stacking them up to build a DBN. The fine tuning using back-propagation algorithm is to optimise the weights between the layers.

units of the layer are also binary. The binary-visible and binary-hidden RBM is also known as Bernoulli-Bernoulli RBM. Each and every image is aligned to be in one vector with dimension of 784×1. Each visible unit is connected to every binary hidden unit.

Imitating Hinton's work (Hinton *et al.* 2006), this first hidden layer contains 2000 units. The second and third RBM layers are both Bernoulli-Bernoulli RBM with visible units obtain the input for hidden units from the immediate previous layer. The second and third layer has 1000 and 500 units, respectively. The fourth layer of DBN is called code layer. This layer is a linear RBM, which is different from the previous layers. The hidden layer of this RBM is not in binary state, it is linear which has 30 units of real value. After pre-training each RBM layer separately for 50 epochs each, these RBMs are stacked to build up a DBN. To get the reconstructed images, the network is then fine-tuned using back-propagation algorithm for 200 epochs. The flow diagram of the procedure is illustrated in Figure 5.8. Due to the large database involved in the procedure, the time taken to complete the procedure is long ie. approximately 70 hours.

The sum squared errors in each pre-training layer in the decoder process is observed and plotted as in Figure 5.9 (a)–(d). The sum squared errors in all pre-training layer showed a decay after 50 epochs. This means the CD learning in RBM has successfully modeled the distribution from the training data and achieved equilibrium condition in each layer.

To reconstruct the original image, a decoder process is carried out. The decoder process includes decoding the code layer and using the transpose weights between each layer in the encode process. The image is successfully decoded from the code layer and is shown in Figure 5.10.

The results obtained from Hinton's work shows than DBN is capable to learn and decode features from digit images, and store them in one code layer. The features in this code layer can then be encoded to retrieve the original digit images. This experiment

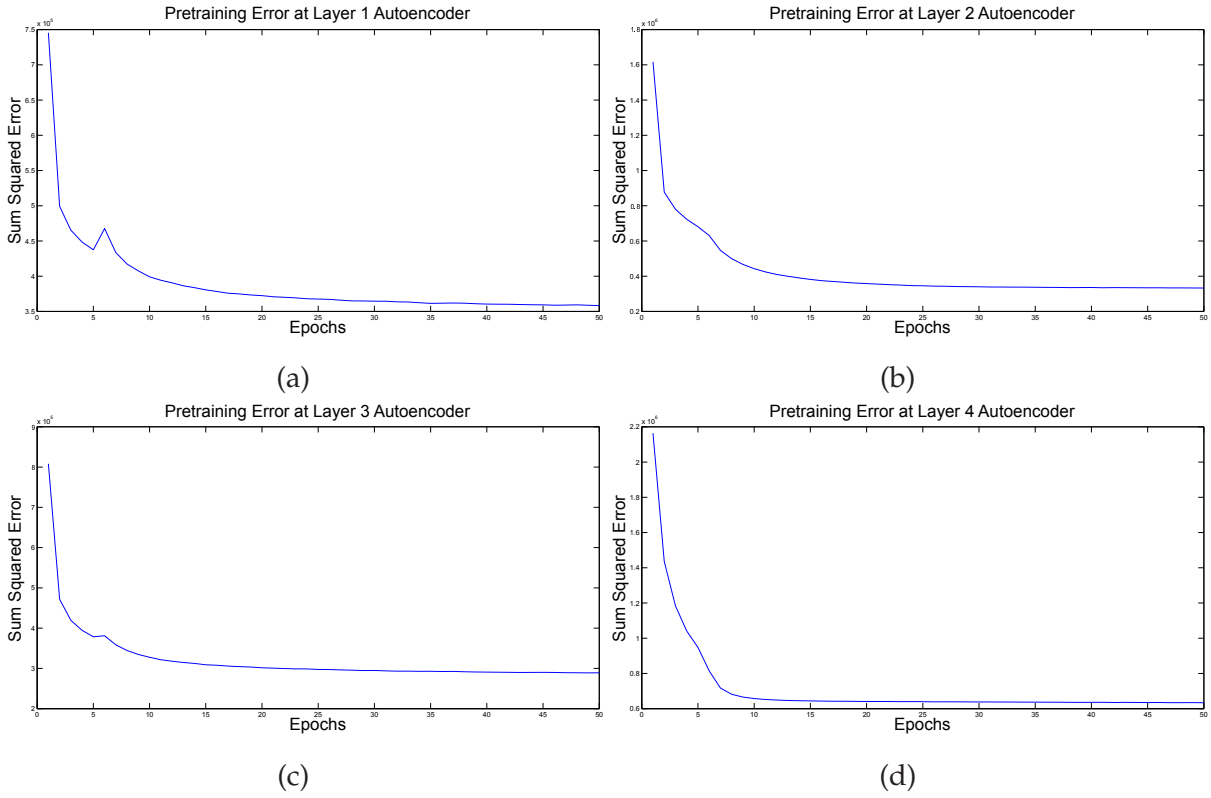


Figure 5.9: The sum squared errors of pre-training in the autoencoder's layers after 50 epochs in (a) first layer, (b) second layer, (c) third layer, and (d) code layer.

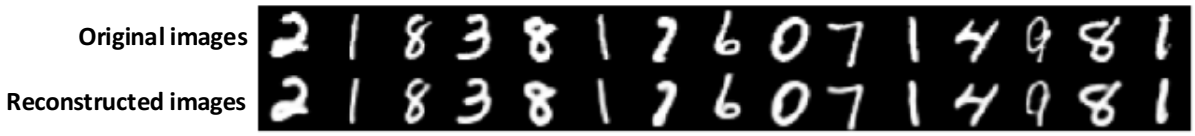


Figure 5.10: The first row represents the original MNIST digits images. The bottom row are the reconstructed images from the DBN autoencoder. The images were obtained from MNIST database (Lee *et al.* 1998).

motivates us to use deep network to extract features that were decoded from the micro-calcification images, and then to use them for malignancy analysis. It is assumed that the features extracted contained most of useful information for the analysis. In addition, it also can be used to compress the images into one code layer and thus reducing the number of features for classification.

5.5 Summary

It is known that DBNs have proved to be successful in a number of pattern recognition tasks that have stubbornly challenged older techniques. With the ability to suitably and effectively train very deep networks, the DBN has attracted many researchers in solving particular classification tasks. The inherent limitations in conventional neural network training schemes, researchers have been unable to get networks with sufficient depth to take on challenges such as classifying microcalcification clusters in mammograms. Greedily training layer-by-layer in DBN has shown to has an advantage to be able to train deeper network, and thus provide more non-linear transformation, which is able to analyse more complex patterns. Thus, in this thesis, the application of DBN for improving classification of microcalcification in mammogram is experimented with and analysed through extensive experiments presented in Chapter 6.

Chapter 6

Feature Extraction Experiments

THIS chapter discusses the experiments undertaken in this thesis. The experiments were performed in two parts. First part is the novel feature extraction for microcalcification clusters diagnosis. The experiments include the features that are extracted from the decomposed images and then are classified using a SVM classifier.

Second part of experiment is the novel automatic feature extraction and feature selection using deep belief network (DBN). Different DBN topologies were experimented with, to investigate their effects on overall classification performance. The motivation behind these experiments are based on the successful applications of DBN in other areas that required judicious selection of parameters, and there are no prior work using DBN for breast cancer diagnosis.

6.1 Pre-processing and Data Acquisition

Prior to the running of the experiments, a collection of images is gathered and pre-processed to ensure the format compatibility with the software used. Thus, in this section, the steps involved in data acquisition are briefly described as preparation for use in training and testing of the method implemented. The overall process in data acquisition and ROI segmentation is shown in Figure 6.1.

Figure 6.1: Flow chart of data acquisition and segmentation process.

6.1.1 Data Acquisition

The mammograms images in this research are acquired from the digitally stored database named Digital Database for Screening Mammography (DDSM) which can be obtained from the website http://marathon.csee.usf.edu/Mammography/Database.html. The database is a collaboration project by several institutes mainly Massachusetts General Hospital, the University of South Florida and Sandia National Laboratories, and is maintained by University of Florida. There are several research work in this area that use the images from DDSM database, hence the images are reliable to use (Dehghan *et al.* 2008, Ren 2012, Andreadis *et al.* 2015).

DDSM database

The film-screen mammograms in the DDSM database were digitised at 42, 43.5 and 50 micron resolution, using 4 different types of digitisers; DBA M2100 ImageClear, Howtek 960, Lumisys 200 Laser and Howtek MultiRad850 (Heath *et al.* 1998,

Number of Cases			Total
Normal	Benign	Malignant	
695	1011	914	**2620**

Table 6.1: Number of cases for each type normal, benign and malignant mammogram in DDSM database.

Heath *et al.* 2001). The DBA M2100 ImageClear produces digitised mammogram with 16-bit grayscale whilst the other 3 digitisers produce digitised mammogram with 12-bits grayscale. Although the number of bits per sample varies from 12 to 16 bits, the images used have all been reduced to 8-bits per sample. This implies that the images have 256 levels of gray. This number is convenient for computation programming specifically for image processing since each pixel occupies a single byte.

The database contains a total of 2620 normal, benign and malignant cases. Table 6.1 summarises the number of cases per type; normal, benign and malignant in DDSM database. The classes of each were determined according to the results obtained from the screening examinations. Thus, the database also provides ground truth for each mammogram in the database. A mammogram is categorised as normal if the screening examination result does not show any abnormality. Cases are considered benign if the results showed suspicious appearances but was decided that it is not malignant by the pathologists. Meanwhile, cases are considered malignant if the suspicious appearances on the results are decided to be cancerous by at least one pathology (Heath *et al.* 1998, Heath *et al.* 2001).

In the DDSM database, each case consists of a minimum of 6 files and maximum of 10 files. In a normal case, it consists of 1 header file, 2 images from views MLO and CC (see Figure 1.2(a) and Figure 1.2(b)) of both left and right breasts, and 1 small thumbnail image file. These gave a total of 6 files with 4 images per case. A header file contains the patient's information such as the patient's age, as well as the study information such as date of the study, date of film digitation and the type of digitiser used. It also contains the list of images in the case directory tagged with its resolution and size of the image pixels. For benign and malignant cases, an extra file called *overlay* is provided. This

Chain code value	0	1	2	3	4	5	6	7
x-coordinate	0	1	1	1	0	-1	-1	-1
y-coordinate	-1	-1	0	1	1	1	0	-1

Table 6.2: Chain code values and their direction in x and y coordinate

file contains specific information of lesion type, the assessment, the boundary outline of the lesion, the abnormalities' subtlety and also pathology. With these information, the ground truth of each case containing abnormalities can be retrieved.

The outline for the boundary of the suspected lesion is given as a chain code in its overlay file. The chain code values and their corresponding directions in x and y coordinates are shown as in Table 6.2. An example of an overlay file is illustrated in Figure 6.2. The steps for using these information to segment the region of interest will be further discussed in Section 6.1.3.

```
TOTAL_ABNORMALITIES 1
ABNORMALITY 1
LESION_TYPE CALCIFICATION TYPE PLEOMORPHIC-FINE_LINEAR_BRANCHING DISTRIBUTION REGIONAL
ASSESSMENT 5
SUBTLETY 4
PATHOLOGY MALIGNANT
TOTAL_OUTLINES 4
BOUNDARY
8 1368 4 4 4 4 4 4 4 4 2 2 2 2 2 2 2 2  ...  0 0 0 0 0 0 0 0 0 1 #
        CORE
168 1824 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  ...  1 0 1 1 0 1 1 0 1 1 #
        CORE
384 1848 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1  ...  0 0 0 0 0 0 0 0 0 0 #
        CORE
368 2192 6 6 6 6 6 6 6 6 0 0 0 0 0 0 0 0  ...  0 0 0 0 0 0 0 0 0 0 #
```

Figure 6.2: An example of the contents in the *overlay* file. The file is from case B_3024_1.RIGHT_-CC.OVERLAY adapted from the DDSM website.

The images in the DDSM database are stored using lossless JPEG encoding. Each image has a total size of roughly 10 MB, which makes a total of 265 GB for storing all images on the database.

In this thesis, the aim is to classify the microcalcification clusters detected on the mammograms into benign and malignant cases. Thus, the images in the DDSM database that contain only the clusters of microcalcification of both benign and malignant cases are selected. The images are chosen from the first three volumes containing benign

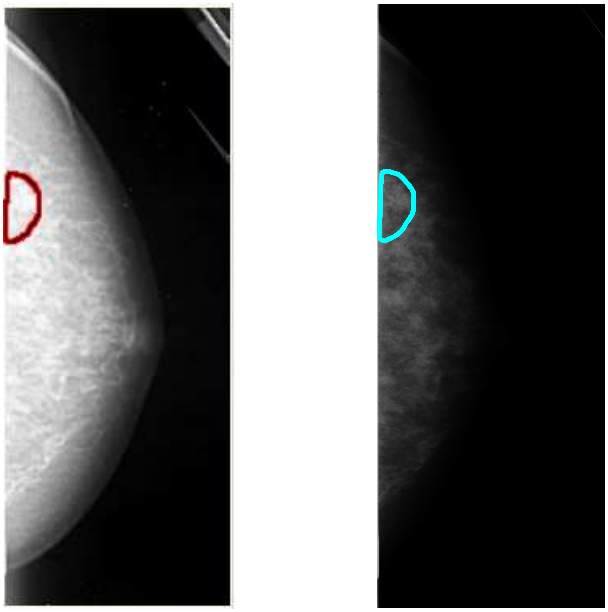
cases and the first three volumes containing cancer cases, ie. volume `benign_01` (80 cases), `benign_02` (69 cases), `benign_03` (64 cases), `cancer_01` (69 cases), `cancer_02` (88 cases) and `cancer_03` (66 cases). However, due to space limitation on computer disk, a subset of 118 from 436 images, 66 benign and 52 malignant cases are used for the study. The detail procedure of how these 118 images were chosen is described in Section 6.1.3. A complete list of names and details for all the 118 cases used is shown in Appendix A.

6.1.2 Image Format Conversion

In this research, Matlab is the main platform for building algorithm and running the code. Matlab is built for numerical computing and thus suitable for the image processing purposes. The lossless JPEG (.LJPEG) image format of DDSM database is not supported by Matlab, so it is necessary to convert a .LJPEG mammogram image to a lossless portable graymap format (.PGM) image. The process consists of two steps, as required by the limitation of the software tools available:

1. uncompress the compressed .LJPEG image to give the raw image in .LJPEG.1 format.
2. convert the uncompressed raw .LJPEG.1 image to portable graymap (.PGM) format.

The first process, which is to produce a raw image is performed by using existing program code in C-language. The source code is provided in the DDSM database. The simulation is carried out in Linux environment platform. After a raw image is obtained, the image is converted to .PGM format in the Matlab.



(a) (b)

Figure 6.3: (a) Thumbnail image of mammogram for case 1002_RIGHT_CC from volume CANCER_3 together with marked ground truth obtained from DDSM website. (b) The processed image of the same mammogram case. The blue line marks the boundary obtained from the Matlab simulation using chaincode provided by the overlay file.

6.1.3 Segmentation of Region of Interest (ROI)

As mentioned in Section 6.1.1, each case in the DDSM database provides an overlay file that describes the area that contains abnormal tissues. These areas were marked by radiologists and were used as the ground truth of each case. The perimeters of the areas are stored as a chain code. Thus, the segmentation of the ROI can be carried out based on the chain code. The chain code is shown in Table 6.2.

The purpose of segmenting the ROI is to reduce the size of the input and thus to reduce the computational time. It is also feasible to use smaller size of breast region compared with using the whole image, as the main purpose of the technique is to classify the class of the detected microcalcification into benign or malignant cases. Based on the extensive preview on previous work, it is assumed that the problem in the detection of microcalcification technique has been solved and is outside the scope of this thesis. In this thesis, the focus is on the diagnosis of the detected microcalcification clusters.

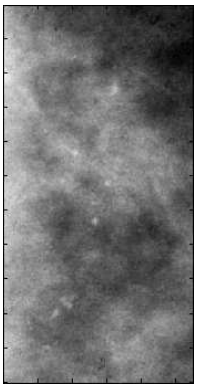


Figure 6.4: Segmented ROI from the same mammogram case in 6.3(b) using implemented Matlab routine.

Hence, to use an area of ROI that contains only the detected microcalcification clusters instead of the whole mammogram is justified in this study.

In this thesis, three volumes of each case of benign and malignant from the DDSM database are chosen. The benign images are from volume `benign_01`, `benign_02` and `benign_03`. Whilst the malignant images are from volume `cancer_01`, `cancer_02` and `cancer_03`. From these six volumes, only cases that contain abnormalities of cluster of microcalcification are chosen. This selection produces 403 mammograms with each contains at least one cluster of microcalcification. The segmentation process is performed on all selected mammograms. Since the clusters are vary in sizes, the sizes of the segmented ROIs obtained from the process are also varies. Figure 6.5 shows the histogram of the sizes (in pixels) of the segmented ROIs.

However, for consistency, segmented ROIs are required to be of the same size. The boxplot of the segmented ROIs sizes is illustrated in Figure 6.6. The median of the ROI sizes is approximately 371×371 pixels, which is in the range of $2^8 \times 2^8$ and $2^9 \times 2^9$ pixel size. Hence, in this thesis, the ROI with sizes within $2^8 \times 2^8$ or 256×256 pixels are selected. After the ROIs selection, a total of 118 images are obtained and used for all subsequent experiments.

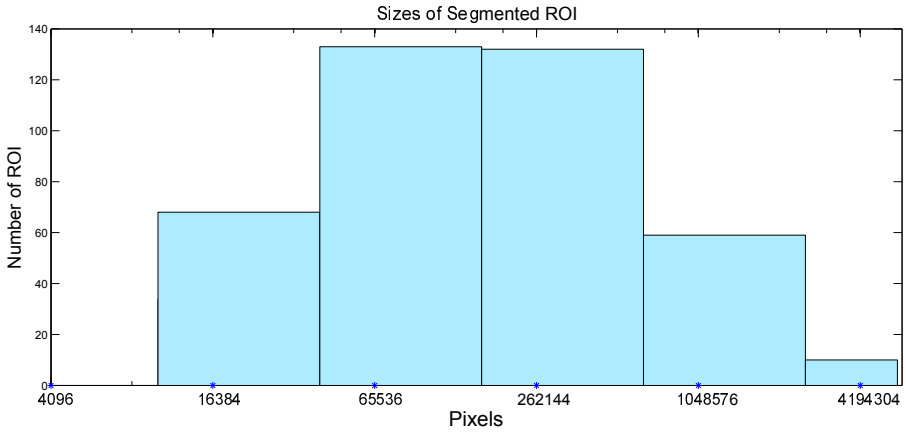


Figure 6.5: Histogram of ROI sizes from 6 volumes in DDSM; for benign cases: benign_01, benign_02 and benign_03, and for malignant cases: cancer_01, cancer_02 and cancer_03.

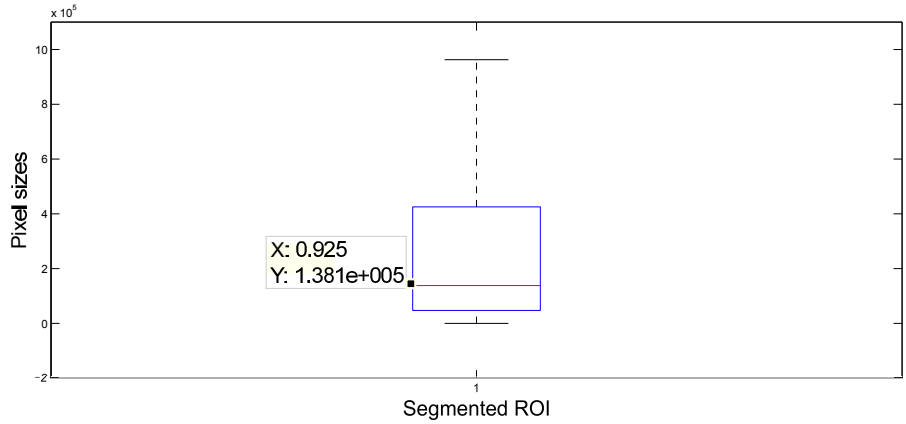


Figure 6.6: Boxplot of ROIs' sizes showing the median of 138112 pixels, or approximately 371 × 371 pixels. This range is between $2^8 \times 2^8$ and $2^9 \times 2^9$ pixel size. Thus, the ROI with sizes within $2^8 \times 2^8$ or 256 × 256 pixels are selected.

6.1.4 Data Scaling

Data scaling is necessary before the classification is carried out. This process overcomes the inconsistency in the data values. In this process, the data value, x_p is scaled by changing the range of the data intensity, $[t_{\min}, t_{\max}]$, so that its dynamic range is condensed to new range $[t_{\text{newmin}}, t_{\text{newmax}}]$, ie. in this work, the new dynamic range is chosen to be in $[0, 1]$. The data scaling process follows the equation:

$$x_{p(\text{new})} = \frac{(x_p - \min(x_p))}{x_{\text{range}}} \times x_{\text{new range}}, \tag{6.1}$$

where $\min(x_p)$ is the minimum value of the original data, x_{range} is the range of the original data:

$$x_{\text{range}} = t_{\max} - t_{\min}, \tag{6.2}$$

and $x_{\text{new range}}$ is the new range of $[0, 1]$

$$x_{\text{newrange}} = t_{\text{newmax}} - t_{\text{newmin}}. \tag{6.3}$$

The new scaled data value $x_p(\text{scaled})$ can be obtained by adding up the calculated $x_p(\text{new})$ with the new minimum, t_{newmin}, or mathematically expressed as:

$$x_p(\text{scaled}) = x_p(\text{new}) + t_{\text{newmin}}. \tag{6.4}$$

6.1.5 Training and Testing Datasets

For all experiments conducted in this thesis, the 118 ROI images are randomly divided into training and testing datasets. Very often, datasets are divided into training and testing of 70:30 ratio. Therefore, 80 images are randomly chosen as training set with 40 images from each case. The remaining 38 images are the testing dataset that will be used to validate the accuracy of the proposed system.

The next step before the proposed method for feature extraction can be evaluated is the integration with the classification step. The classifier chosen in this thesis is SVM with radial basis function, RBF. Parameters that require tuning are the penalty parameter C and sigma, σ in the Equation 6.5

$$K(x, y) = \exp(-\frac{||x - y||^2}{2\sigma^2}), \tag{6.5}$$

where $\sigma > 0$ defines the kernel width.

These two parameters are tuned by a grid-search with 5-fold cross validation. In 5-fold cross validation method, the dataset is divided into five equal subsets. One of the subsets is used as testing set on the remaining subsets as training set. A pair of parameters C and σ are tuned by grid-search in the classifier to find the average of the percentage that accurately classifies the class of microcalcification clusters. The pair of parameters that achieved the highest accuracy rates is then used for classification using SVM throughout this thesis. This is valid for the experiments where SVM is used as a classifier.

6.1.6 System Evaluation

The efficiency of the system can be evaluated based on the accuracy of the system in classifying the clusters. Receiver operating characteristic curve is commonly used to visualize the classifier performance. There are four possible outcomes of the classifier and these are defined as follows. These outcomes also can be summarised in the confusion matrix shown in Table 6.3.

- True positive (TP), when the classifier makes a correct hit, the patient has malignant tumour.
- True negative (TN), when the classifier makes a correct rejection, the patient has benign tumour.
- False positive (FP), when the classifier makes an Error Type I, the patient with benign tumour is diagnosed to have malignant tumour.
- False negative (FN), when the classifier makes an Error Type II, the patient with malignant tumour is diagnosed to have benign tumour.

The sensitivity or true positive rate (TPR) indicates the number of accurately identify a positive condition. It measures the rate of correctly diagnosing a condition. In another words, it measures the probability of correctly diagnosing positive (malignant)

		True Class: P	True Class: N
Predicted Class	P	True Positive	False Positive
	N	False Negative	True Negative

Table 6.3: Confusion matrix illustrated the four possible outcomes of the diagnose using a classifier, TP, TN, FP and FN, given the True Class condition.

conditions, and is complementary to the falsely diagnosing them. Mathematically,

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{6.6}$$

On the other hand, the specificity or true negative rate (TNR) indicates the number of accurately identify negative condition. Thus, in contrast to TPR, it measures the probability of correctly diagnosing negative (benign) conditions, complement to falsely identifying them as positive or malignant.

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}. \tag{6.7}$$

Both sensitivity and specificity measures are significant and commonly used in the field to indicate the performance of the diagnosis system. These outcomes can then be used as an input to plot a receiving operating characteristic, ROC curve. This curve represents the relationship between the TPR and the false positive rate (FPR). It is desirable to have a system with high accuracy or high TPR with an acceptable of FPR.

6.2 Feature Extraction using Steerable Pyramid Filtering

A novel technique for feature extraction using steerable pyramid is proposed to improve the classification rate of microcalcification diagnosis in mammograms. The orientation of clusters of microcalcification in mammograms may also contain information for malignancy analysis. Hence, the aim is to extract more features at multiple orientation and multiple resolution. Steerable pyramid filtering technique offers multiple

resolution and orientation decomposition. This property makes this filtering technique to be selected for microcalcification feature extraction.

The experiments are organised as follows:

1. In Section 6.2.1, experiments on several basic parameters are conducted. The parameters that need to be determined are: a) basis filters sets and b) number of levels and orientations.

2. In Section 6.2.2, the steerable pyramid filtering is used to decompose microcalcification cluster images into multiple orientation and multiple resolution, and extracting features from the images for malignancy analysis.

3. In Section 6.2.3, the principle component analysis (PCA) is used to reduce the number of features. The aim is to have same number of features from different techniques for fair comparison.

4. In Section 6.2.4, different sets of features that have been reduced by PCA are used for microcalcification classification.

6.2.1 Experiment 1: Steerable Pyramid Topology Selection

The purpose of this experiment is to choose a set of basis filters for image decomposition and examine the features that has been extracted. To reduce the computational complexity, three sets of basis filters proposed by Simoncelli were adapted for this experiment. Plots of these filters' responses are shown in Figures 3.13(a)–3.15(b) (Simoncelli and Freeman 1995).

It is also aimed to find the appropriate level of resolution to decompose mammogram images for feature extraction purposes. The images are decomposed using three different sets of basis filters ie. `sp1Filters`, `sp3Filters` and `sp5Filters`. A `sp1Filters` consists of two oriented basis band-pass that can be linearly combined together with its interpolation functions to produce filters at arbitrary orientations. A `sp3Filters` has

four oriented basis filters, and `sp5Filters` has six oriented basis filters. These filters were described in Section 3.4.

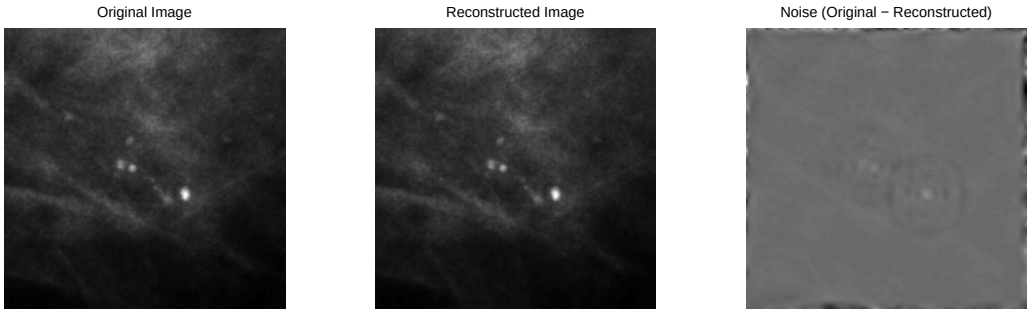
In multiple resolution analysis, the frequency contained in the signal can be analysed at different scales. The number of resolutions is chosen based on some initial explorations that show the common number of resolution levels for decomposing DDSM mammograms is at three-level. Here, only the maximally decimated filterbanks for steerable pyramid filtering is considered. Hence, the images are decimated by two to produce next scale of resolution. The largest possible decomposition level is 4 since the pixel size of the experimental images is 256×256. Thus, the experiment is performed when the number of resolution is at resolution levels 3 and 4.

For each experiment, images were decomposed into eight orientations at angles of; $0°, 11.25°, 22.5°, 33.75°, 45°, 56.25°, 67.5°, 78.75°$, and $90°$ using different sets of oriented basis filters. In each resolution, the approximated image is down-sampled by a scale of two, producing the approximation images at a coarser resolution, with half size of the original images. The down-sampled approximate image is further decomposed into eight orientations using the same set of basis filters. This iteration process is repeated to produce multiple resolutions. With these, a direct comparison of the output with the conventional dyadic wavelet transform decomposition can be performed.

The decomposition produces 24 and 32 images composing of eight different orientations at three and four different resolutions, respectively. After obtaining these decomposed images, the decomposed images are reconstructed to obtain the original image. The difference between the original and the reconstructed images can be measured to assess their performances. This is to measure any information loss occurred during the decomposition process. One of the measurements is by calculating the signal to noise ratio (SNR) of the noise in the reconstructed images. SNR is defined as power, P ratio of the signal, P_{signal} and its noise, P_{noise}. The calculation for SNR is as follows:

$$\text{SNR} = \frac{P_{\text{signal}}}{P_{\text{noise}}}. \tag{6.8}$$

Figure 6.7(a) shows one of the ROI image and its reconstructed image in Figure 6.7(b). In this example, the ROI image is decomposed by `sp1Filters` into eight orientations and three resolutions. After the decomposition process, the filtered images were then reconstructed. Figure 6.7(c) shows the difference between the original and the reconstructed image.



(a) (b) (c)

Figure 6.7: (a) Original image of segmented ROI, (b) its reconstructed images obtained after filtering the original image using set of bandpass filters, and (c) the loss measured by calculating the difference between the original and the reconstructed image.

The proposed method for calculating the SNR is illustrated in Figure 6.8.

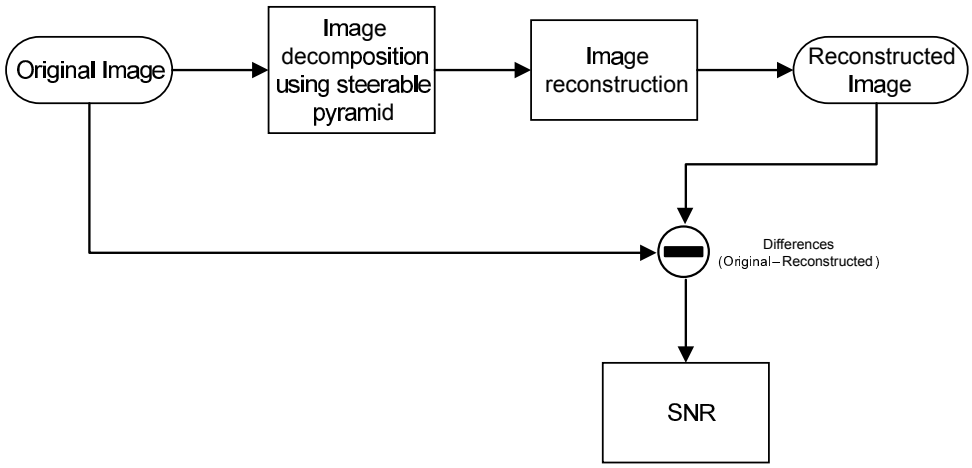
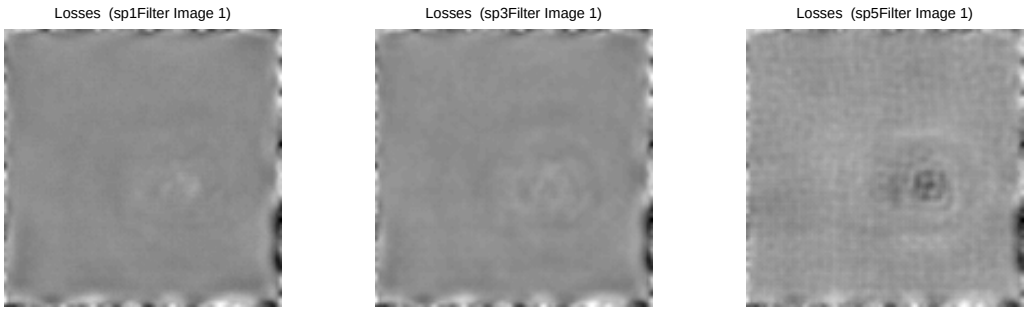


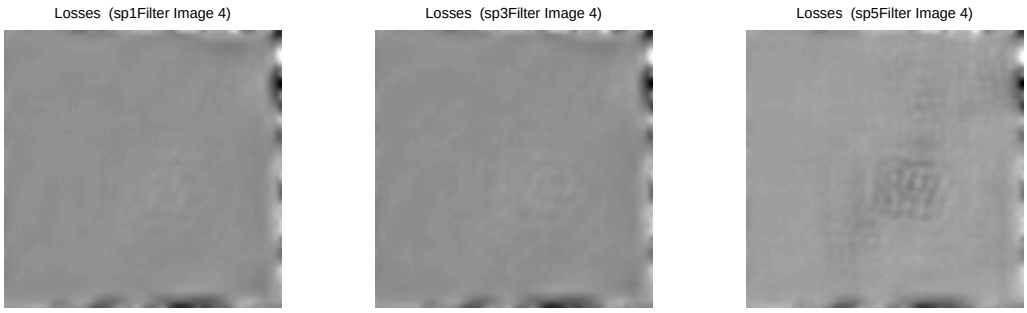
Figure 6.8: To calculate the SNR, the difference between the reconstructed and the original image is measured. This difference is use to find the SNR by finding the ratio of the original image to the difference.

From Equation 6.8, when the image difference between the original and the reconstructed is small, the reconstructed image is closely similar to the original image.

Hence, the smaller the image difference, the larger the SNR of the difference and the original. Figure 6.9 shows the difference images and their SNR, respectively. From the figure, it is observed that when the SNR is larger, the difference image has less noise compared to the difference image with lower SNR.



(a) SNR: 37.01 dB (b) SNR: 37.26 dB (c) SNR: 35.79 dB



(d) SNR: 39.04 dB (e) SNR: 39.24 dB (f) SNR: 37.69 dB

Figure 6.9: Losses images from 2 randomly chosen ROIs. The losses were obtained by measured the difference between the original image and its reconstructed image. From left to right: losses from reconstructing images after decomposing them using 3 different set of basis filters, sp1Filters, sp3Filters and sp5Filters.

The set of basis filters that produces higher SNR between the original image and its difference with reconstructed image shows that the image is reconstructed to higher fidelity than the others. The reason for this measurement is to choose a basis filters set and suitable level of decomposition that can decompose the images in the datasets without losing much information by measuring the reconstruction error. This is to preserve as much information as possible in each of the decomposed images. Thus, the result from this experiment will be used as an indication of choice of steerable pyramid parameters used in the entire experiments.

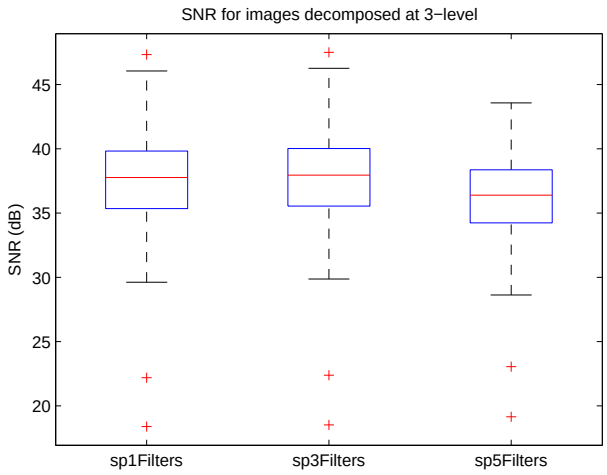


Figure 6.10: Boxplot of data from SNR calculation between the original image and its reconstruction image using 3 different set of bandpass filters at 3-level of decomposition. Each of the image has pixels size of 256×256.

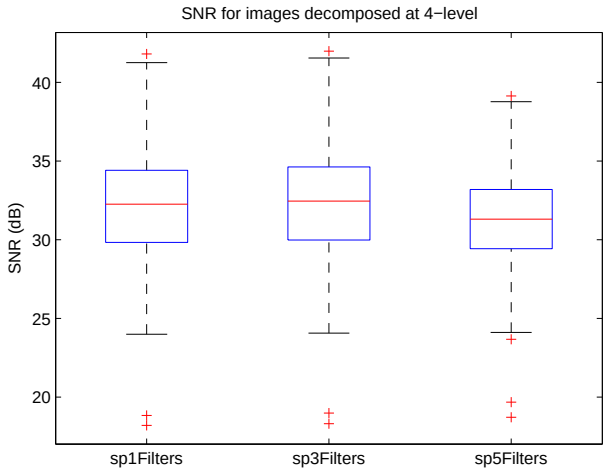


Figure 6.11: Boxplot of data from SNR calculation between the original image and its reconstruction image using 3 different set of bandpass filters at 4-level of decomposition. Each of the image has pixels size of 256×256.

From the boxplot in Figure 6.11, it shows that the median SNR value for `sp3Filters` is higher than `sp1Filters` and `sp5Filters`. Hence, the reconstructed images using this filter has lower losses compared with those decomposed using `sp1Filters` and `sp5Filters`.

The average and standard deviation results are also recorded in Table 6.4. From these results, `sp3Filters` has higher median SNR compared to other two basis filters sets, when the images were decomposed and reconstructed using 3-level decomposition. Hence, `sp3Filters` basis filters are said to be better at decomposing the ROI images with lower losses incurred during the process. This is also true when images decomposed at 4-level. However, the combination of using `sp3Filters` basis filters and at 3-level decomposition yields the highest SNR. Thus these parameters are chosen to be used for decomposing ROI containing microcalcification clusters image using steerable pyramid for the related experiment in this thesis.

	3-level	4-level
Basis filter	Mean SNR (dB)	Mean SNR (dB)
sp1Filters	37.61	32.06
sp3Filters	37.80	32.20
sp5Filters	36.21	31.25

Table 6.4: Mean SNR for three different sets of basis filters at 3 and 4 levels decomposition.

An experiment to calculate SNR of the reconstructed images when the input images' sizes are reduced to half; (128×128 pixels), is also conducted. The reason to reduce the size of the input image is to lessen the computational complexity, and also to provide consistency in later experiments. The boxplot is illustrated in Figure 6.12.

From the result, the highest mean SNR obtained when the basis set `sp3Filters` is used to decompose the images at 3 resolution levels. However, the mean SNR is reduced from 37.80dB to 32.83dB, when the sizes of the input images are reduced. This implies greater information loss has occurred during the reduction in the number of pixels.

Summary

From the experiments performed, the parameters for steerable pyramid filtering were decided. The four basis filters set gave a slightly better in terms of reconstruction error when compared to the basis filters of two and six. Thus, the basis filters `sp3Filters` will be used set to produce multiple orientation of the microcalcification cluster images. Moreover, three levels of pyramid steerable pyramid filtering showed higher

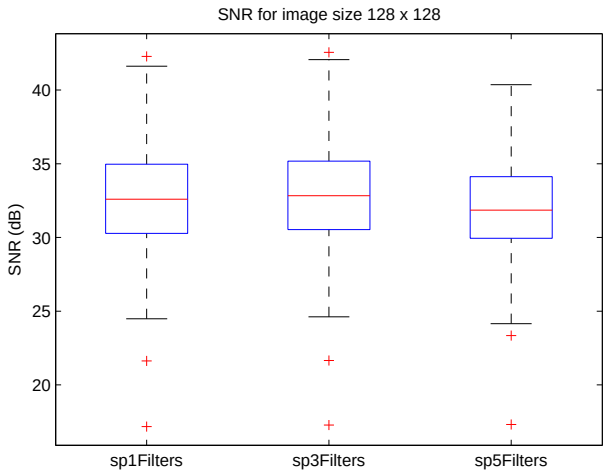


Figure 6.12: Boxplot of data from SNR calculation between the original image and its reconstruction image using set of sp3Filters bandpass filters at 3-level of decomposition, with reduced size of input images.

SNR compared to four levels of resolutions. With these results, three levels of decomposition are chosen for the subsequent classification experiments. Figure 6.13 illustrates the framework of the proposed feature extraction for microcalcification classification in this thesis.

6.2.2 Experiment 2: Steerable Pyramid Feature Extraction with SVM Classifier for Microcalcification Classification

In Section 6.2.1 the framework that will be used for image decomposition has been identified. The basis filters that will be used is a set of `sp3Filters` and the pyramid consists of three levels of resolutions at eight different angles of orientation as illustrated in Figure 6.13. In this section, the steerable pyramid filtering is proposed for feature extraction to classify the microcalcification clusters. Earlier in Section 6.1.3, a subset of 118 from over 2000 mammograms from the DDSM that contain benign and malignant microcalcification has been identified. Figure 6.14 is a sample of one of benign cases of microcalcification that has been segmented.

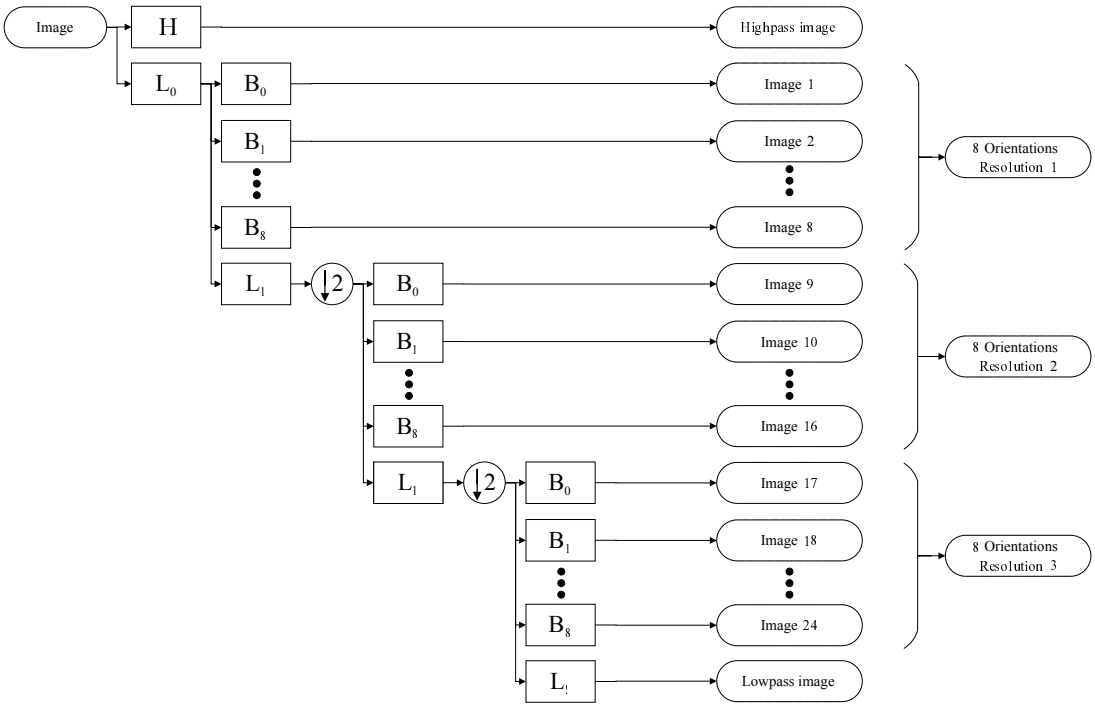


Figure 6.13: The topology of steerable pyramid consists of three level of resolutions and eight orientations at each resolution will be used for the rest of the experiment for steerable pyramid feature extraction.

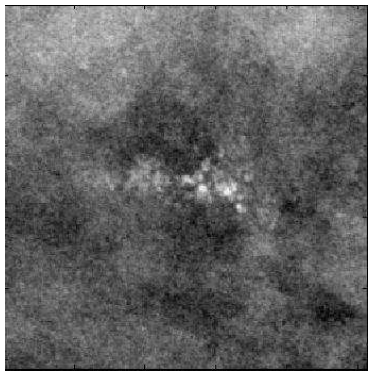


Figure 6.14: Cluster of microcalcification from Digital Database of Screening Mammography.

The proposed method is applied to the images. The resultant subband images are illustrated in Figure 6.15. In the figure, it is observed that 24 subband images were produced from the steerable pyramid decomposition method. The orientation of the filters is varied from 0 to 90 degrees of angle producing eight filtered images at different orientations.

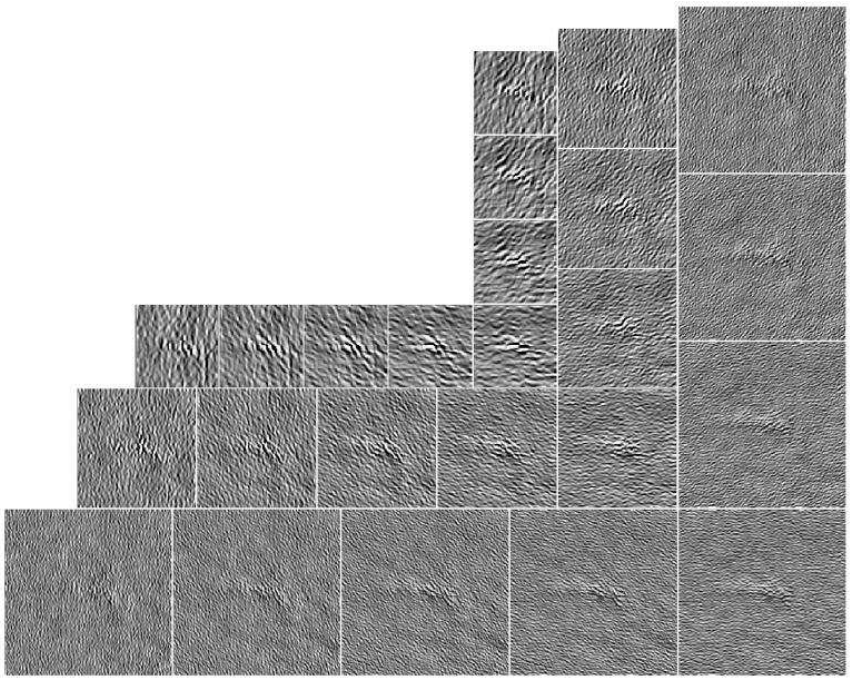


Figure 6.15: Detailed subimages of microcalcification cluster in Figure 6.14 using steerable pyramid filters for image decomposition at 8 orientations and 3 levels of resolutions.

On the other hand, the same image is decomposed with conventional wavelet transform for comparison. The resultant subband images produced from wavelet decomposition is shown in Figure 6.16. Using the conventional 2-D wavelet decomposition, the filtered images were produced at three fixed angles; 0 degree (horizontal), 45 degree (diagonal) and 90 degree (vertical).

By observing filtered images obtained from both methods of decomposition, a qualitative evidence showing clustering in the subband images is produced by the steerable filtering pyramid is obtained. In contrast, the filtered images from wavelet filtering do not exhibit the same degree of clustering.

For quantitative comparison of the two methods, their performances are evaluated using SVM classifier. Two statistical measures were calculated and used as features for classification measures.

$$\text{Energy} = \frac{\sum_{i,j} x^2}{N^2} \tag{6.9}$$

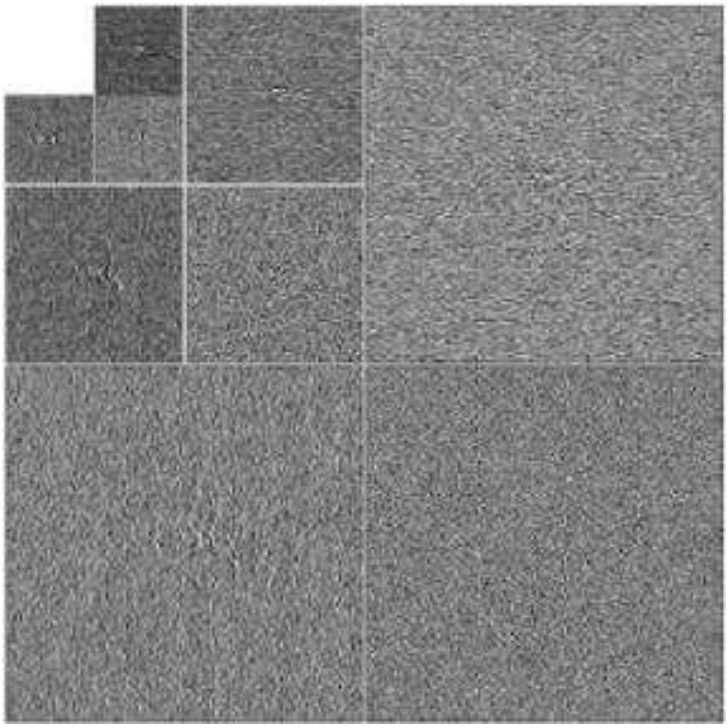


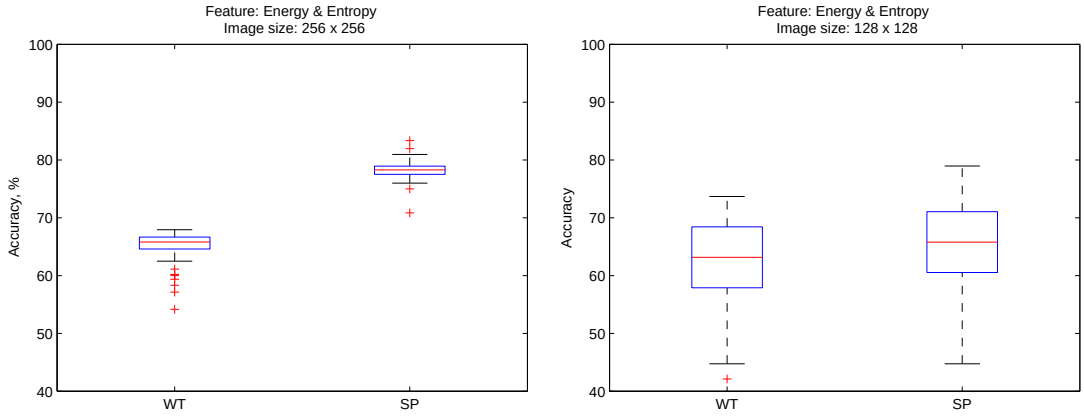
Figure 6.16: Detail subimages of microcalcification cluster image in Figure 6.14 using traditional daubechies wavelet transform for image decomposition at 3 levels of resolutions.

$$\text{Entropy} = -\frac{1}{\log_2 N^2}\sum_{i,j}\left[\frac{x_{ij}^2}{\text{norm}^2}\right]\log_2\left[\frac{x_{ij}^2}{\text{norm}^2}\right] \tag{6.10}$$

where x_{ij} is the ijth pixel value of detail images with $N \times N$ image.

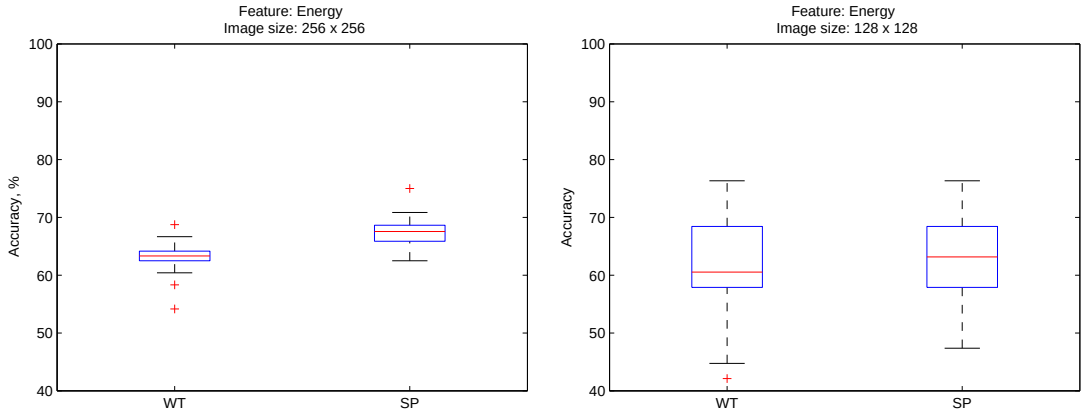
In pyramid filtering, energy and entropy of each sub image decomposed from three levels of resolution at eight orientations were calculated. This gave a total of 48 features extracted from steerable pyramid filtering (24 energy and 24 entropy features). While in wavelet transform, the images is decomposed into three levels of resolution at three orientations, ie. horizontal, vertical and diagonal. These gave a total of nine extracted features from wavelet filtering. All features are scaled to have a range between $[0, 1]$, which has been described in earlier Section 6.1.4. These scaled features are then used as input data for classification using SVM classifier. The input images are in 256×256 pixels. In addition, the features of the reduced pixels of 128×128 are also been extracted and classified.

For the classification step, a SVM with radial-basis kernel is used. The reason to use this kernel is because the feature space is highly non-linear. The parameters were set, using grid-search and pair of parameter C and σ of 4 and 4.0313, respectively.



(a) (b)

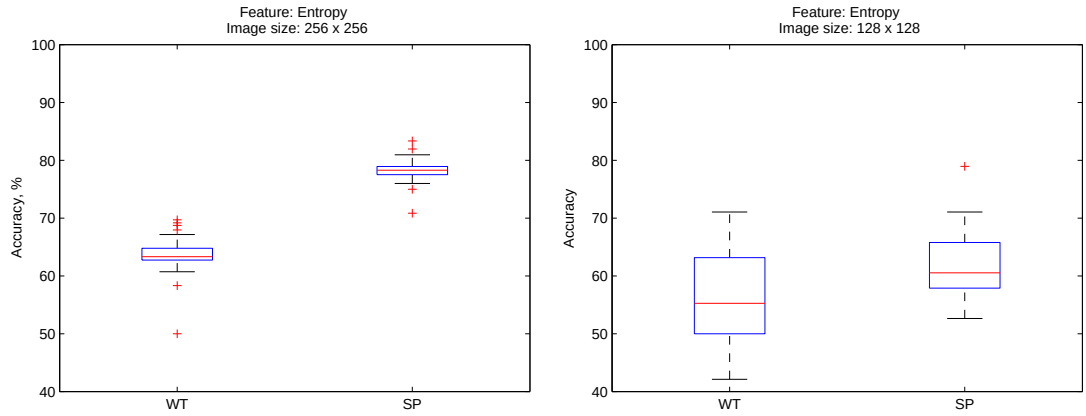
Figure 6.17: Boxplot of accuracy for both wavelet and steerable pyramid with combination of energy and entropy measures as the features for classification for image size of (a) 256×256 and (b) 128×128 pixels. Both classifications were run for 50 times with randomly chosen different set of training and testing.



(a) (b)

Figure 6.18: Boxplot of accuracy for both wavelet and steerable pyramid with energy measures as single features for classification for image size of (a) 256×256 and (b) 128×128 pixels. Both classifications were run for 50 times with randomly chosen different set of training and testing.

Figure 6.17, Figure 6.18 and Figure 6.19 show the accuracy of the SVM classifier when using three different sets of input features; ie. combination of energy and entropy, energy alone and entropy alone respectively. These boxplots compare the accuracy of the



(a) (b)

Figure 6.19: Boxplot of accuracy for both wavelet and steerable pyramid with entropy measures as single features for classification for image size of (a) 256 × 256 and (b) 128 × 128 pixels. Both classifications were run for 50 times with randomly chosen different set of training and testing.

system when the features were extracted from proposed steerable pyramid filtering with wavelet decomposition. The results are presented in Table 6.5.

Decomposition method	Pixels size		256 × 256		128 × 128	
	Features		Accuracy, %		Accuracy, %	
	Energy	Entropy	Mean	Median	Mean	Median
Steerable pyramid	√	√	78.26	78.28	65.00	65.79
		√	78.26	78.28	61.90	60.53
	√		67.34	67.56	62.05	63.16
Wavelet transform	√	√	64.91	65.78	61.39	63.16
		√	63.26	63.33	56.79	55.26
	√		63.22	63.27	61.42	60.53

Table 6.5: Comparison of mean and median accuracy for microcalcification diagnosis between features measured from steerable pyramid filtering and wavelet transform.

Summary

The mean accuracies achieved by the two methods were compared. The proposed steerable pyramid with combination of energy and entropy features achieved mean accuracies of 78.26% for the 256 × 256 pixels and 65.00% for the 128 × 128 pixels. Whilst the wavelet transform method with the same combination (energy and entropy

features), the mean accuracies achieved were 64.91% for the original image size and 61.39% for the reduced size.

Similar trend was observed when the two methods used entropy features as a single input. The proposed steerable pyramid achieved mean accuracies of 78.26% for the 256×256 pixels and 61.90% for the 128×128 pixels. Whilst the wavelet transform method achieved the mean accuracies of 63.26% for the original image size and 56.79% for the reduced size.

With energy features as a single input, the proposed steerable pyramid achieved mean accuracies of 67.34% for the 256×256 pixels and 63.22 for the 128×128 pixels. Whilst the wavelet transform method achieved the mean accuracies of 63.22% for the original image size and 61.42% for the reduced size.

These results clearly showed that the steerable pyramid filtering offers better accuracies in all the types of features arrangement. However, the mean accuracies degraded as the pixel size is reduced to half from the original. This observation is consistent with the result of the experiments discussed in Section 6.2.1.

6.2.3 Experiment 3: Data Dimensions Reduction using Principal Component Analysis (PCA)

In previous experiments in Section 6.2.2, the number of features input for each decomposition method is different. If energy or entropy is used as a single feature input, the total number of 24 and 9 features is obtained from steerable pyramid filtering and wavelet transform, respectively. When the combination of both features is used as input features for the classifications, the total number of 48 and 18 features from steerable pyramid filtering and wavelet transform, respectively, is obtained. The inconsistency in features number may lead to an unfair comparison. To have a fair comparison between two methods, the number of features input is reduced so that both will produce features with the same dimensionality for presentation to the classifier. In addition, the

dimensionality reduction of the features vector reduces the computational burden and of the classification process. This is practically important especially when dealing with large datasets.

One method to reduce the dimensionality of the features vector is by applying the principal component analysis method (PCA). The method was first introduced by Pearson in 1901 and has been commonly used especially to reduce the data dimension using linear technique (Van Der Maaten *et al.* 2008). In this linear technique, the data is transformed by calculating the covariance matrix to obtain an orthogonal data projection basis. With this technique, the principal directions of which the data varies can be identified. The PCA technique also has been applied in other fields such as face recognition (Venkatarajan and Braun 2004) and seismic series analysis (Raytchev *et al.* 2004).

The first step in PCA is normalizing the dataset, to produce zero mean and a unit standard deviation. The mean $\bar{X}$ can be calculated by adding up the dataset across dimension and divide by dataset size, n,

$$\text{mean}, \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}. \tag{6.11}$$

The standard deviation, σ measures the spread of the data,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}. \tag{6.12}$$

Using Equation 6.11 and Equation 6.12, the normalised data, X is obtained as:

$$X_{\text{normalized}} = \frac{X - \overline{X}}{\sigma}. \tag{6.13}$$

The next step is to calculate the covariance of the data. Covariance measures the relationship between the dimensions in the datasets. Similar to the standard deviation, variance measures the spread of data in the set. When the covariance between one dimension and itself is calculated, it is called variance. However, for higher dimension,

eg. n-dimensional data $(X_1, X_2, ..., X_n)$, the covariance is calculated between X_1 and X_2, X_2 and X_3 and so on until between X_{n-1} and X_n, where n is dimension of the data.

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}. \tag{6.14}$$

From Equation 6.14, the covariance matrix, **C** is obtained in form of:

$$\mathbf{C} = \begin{pmatrix} \text{cov}(X_1, X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\ \text{cov}(X_2, X_1) & \text{cov}(X_2, X_2) & \cdots & \text{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \cdots & \text{cov}(X_n, X_n) \end{pmatrix}. \tag{6.15}$$

Eigenvectors and eigenvalues were then calculated from the covariance matrix. The eigenvectors of the covariance matrix **C**, **v** can be found by finding the solution of:

$$\mathbf{C}\mathbf{v} = \lambda\mathbf{v} \tag{6.16}$$

where λ is the eigenvalue corresponds to eigenvector, **v**.

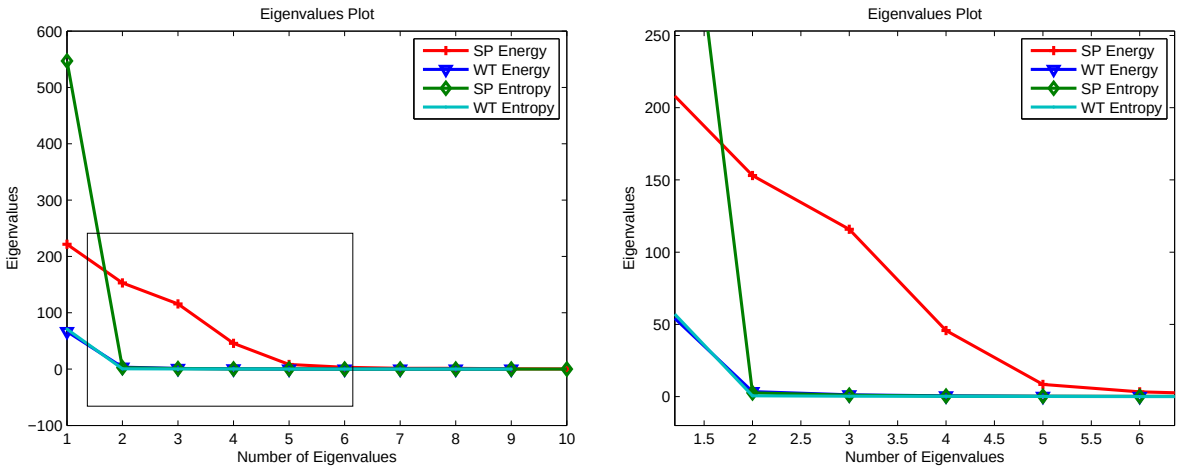
The PCA transformed values are the product of eigenvectors and the original features vector. By finding the eigenvectors and eigenvalues of the data matrix, the data can be mapped to the dimension that preserved the most significant directions,

$$X_{\text{PCA projected}} = \mathbf{v}^T X, \tag{6.17}$$

where $\mathbf{v}^T$ is the transpose of matrix **v**.

The eigenvalues for both features; energy and entropy are calculated for each decomposition technique. These values are then plotted to observe the number of most significant eigenvalue in both energy and entropy features. Figure 6.20 illustrates the

number of eigenvalues and their corresponding values for both energy and entropy statistical measures from each decomposition method.



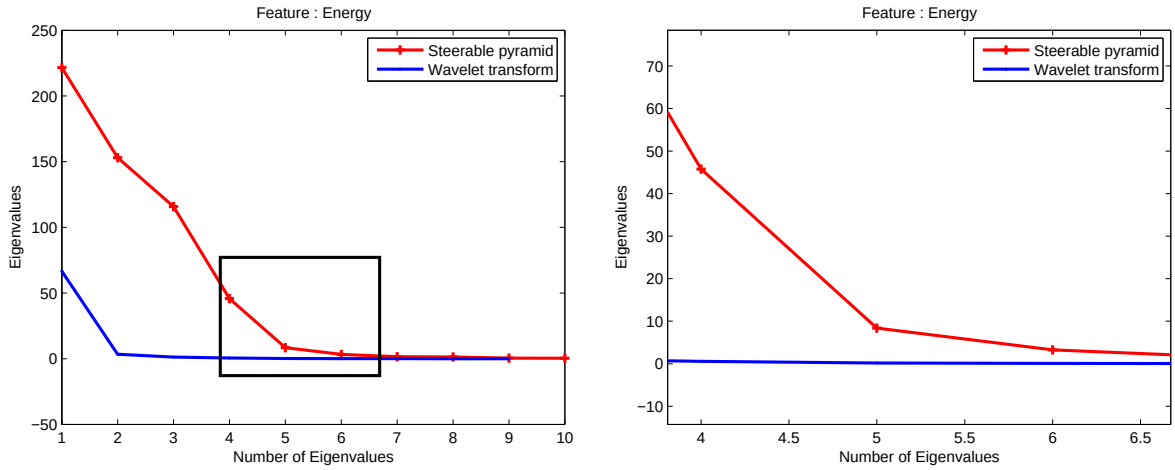
(a) (b)

Figure 6.20: (a) Eigenvalues plot obtained from PCA method for energy and entropy statistical features extracted from decomposition using steerable pyramid (SP) and wavelet transform (WT). (b) The close up graph of the rectangular region in (a).

The eigenvalues and its corresponding values for each energy and entropy are plotted separately for a better close up view. Figure 6.21 is the eigenvalue plot for energy features and its close up view in Figure 6.21(b). From both Figures 6.21(a) and 6.21(b), it is observed that a knee occurs at eigenvalues 2 for wavelet transform and 5 for steerable pyramid decomposition.

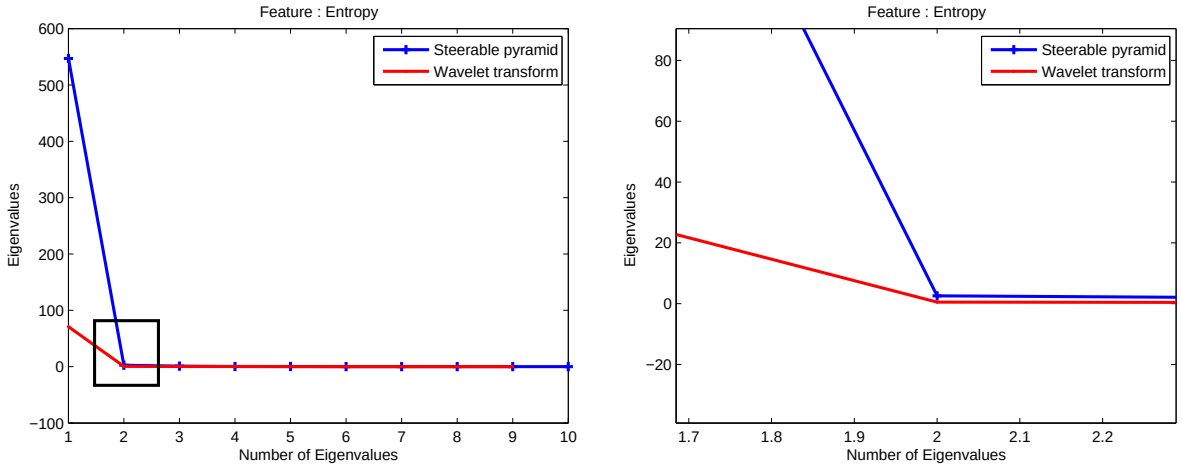
The eigenvalues plot for entropy features is plotted as in Figure 6.22. From the close up Figure 6.22(b), the knee occurs at eigenvalues 2 both wavelet transform and steerable pyramid decomposition.

From the obtained results, an extension to this experiment is performed next using the reduced number of features that varies from 2 to 5, for both energy and entropy features as well as the combination of both features for the presentation to the classifier.



(a) (b)

Figure 6.21: (a) Eigenvalues plot obtained from PCA method for energy extracted from decomposition using steerable pyramid (SP) and wavelet transform (WT). (b) The close up graph of the rectangular region in (a).



(a) (b)

Figure 6.22: (a) Eigenvalues plot obtained from PCA method for entropy extracted from decomposition using steerable pyramid (SP) and wavelet transform (WT). (b) The close up graph of the rectangular region in (a).

6.2.4 Experiment 4: PCA-SP Features with SVM Classifier

Based on the previous results in Section 6.2.3, experiments with several number of PCA transformed features from both energy and entropy features are performed. The number is chosen to be in the range of 2 to 5 of each statistical measure, since the eigenvalues obtained from previous experiment shows knee at eigenvalues 2 from entropy

feature and 5 from energy feature. The features that will be experimenting with are the PCA transformed features extracted from both 256×256 and 128×128 pixels images.

In the first experiment, the PCA transformed energy features are used as a single feature for microcalcification classification. The accuracy obtained are plotted as boxplot and performed the t-test for each pair of features. For an example, the accuracy of the classifier of using 2 PCA energy features from wavelet transform is compared with the accuracy of the classifier of using 2 PCA energy features from steerable pyramid. The reason for the use of the same number of features is to avoid an unfair comparison between both decomposition techniques.

Figure 6.23 and Figure 6.24 illustrate the boxplot obtained when using PCA transformed from energy as single feature with different number of these features as input for microcalcification classification for image pixels 256×256 and 128×128, respectively.

The statistical analysis of each experiment is summarised and presented as in Table 6.6 and Table 6.7. From the results of the analysis, most of the experiments with PCA energy feature as a single feature show wavelet transform decomposition leads to higher accuracy compared to steerable pyramid decomposition. This is mostly true when the features are extracted from input images of 256×256 pixels, with the p-value < 0.05. However, for the features that are extracted from input image of 128×128 pixels, the t-test for the number of 2, 3, 4 and 5 show the p-values of 0.0522, 0.5018, 0.8118 and 0.6189, respectively. The result of p-value > 0.05 shows that the improvement in accuracy achieved is not statistically significant.

The results of the classification rates in this experiment show that the energy of the images are giving less information for malignancy analysis. In the next part, the PCA transformed entropy features are used as a single input features. Similarly, different number of these features is used as input for classification. The number also range from 2 to 5. Figure 6.25 and Figure 6.26 are the boxplot obtained when using PCA transformed from entropy as a single feature with different number of these features as

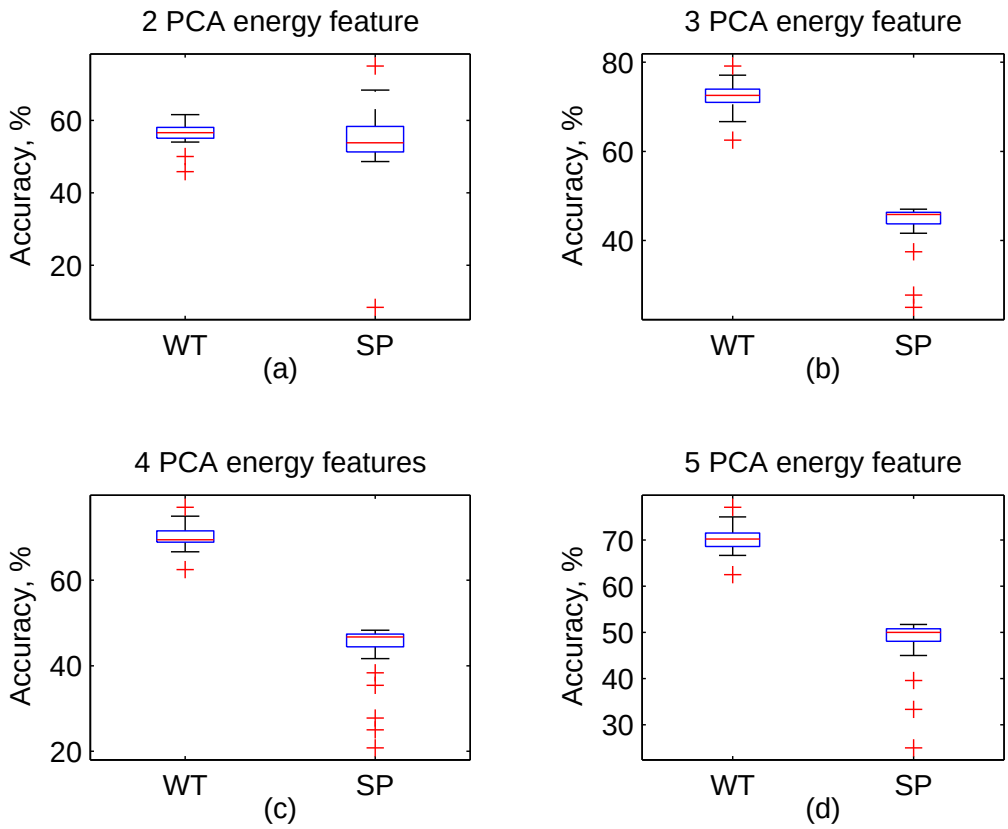


Figure 6.23: Boxplot of accuracy for both wavelet and steerable pyramid with PCA transformed from energy measures as the single features for classifications; with (a) two PCA energy features, (b) three PCA energy features, (c) four PCA energy features and (d) five PCA energy features, for image pixels 256×256.

input for microcalcification classification, for features that are extracted from 256×256 and 128×128, respectively.

The summary of the analysis obtained from the experiment is also presented in Table 6.8 and Table 6.9. The accuracies obtained show that steerable pyramid achieved better accuracy when the features are extracted from the original image size. Entropy features extracted from steerable pyramid achieved highest accuracy of 89.58% when 2 PCA transformed features were presented as input to the classifier. The overall result also shows that steerable pyramid filtering has outperformed features extracted from the wavelet transform. The p-values from the t-test also show that the observed differences are significant.

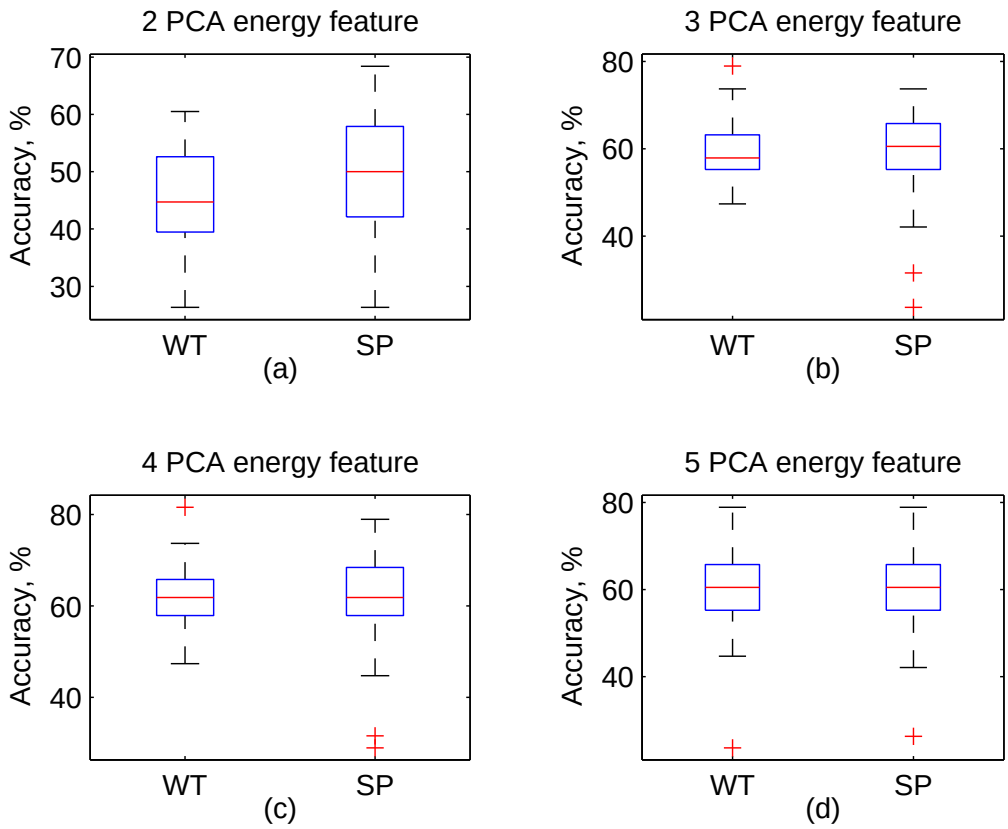


Figure 6.24: Boxplot of accuracy for both wavelet and steerable pyramid with PCA transformed from energy measures as the single features for classifications; with (a) two PCA energy features, (b) three PCA energy features, (c) four PCA energy features and (d) five PCA energy features, for image pixels 128×128.

When the images are reduced to half of the original size, the accuracies achieved are reduced when compared to the result obtain when features were extracted from the original pixels size. Comparing the performances with features that are extracted from wavelet transform, the highest accuracy is achieved when the number of PCA entropy features used are 2 and 3. However, from the t-test analysis, the p-values for both are 0.6539 and 0.3674, which are more than 0.05, hence the improvement shown are not significant. When using 4 PCA features, wavelet transform achieved a significant improvement compared to steerable pyramid as it obtained a p-value of 0.0048. The accuracy improvement of wavelet transform over steerable pyramid when 5 PCA features are used as classification input also not significant with a p-value of 0.0681.

	No of PCA features	Accuracy			t-test p-value
		Mean	Median	SD	
Steerable pyramid	2	54.52	56.61	2.95	0.08
Wavelet transform		56.55	56.61	2.78	
Steerable pyramid	3	43.98	45.83	5.00	0.00
Wavelet transform		72.46	72.58	2.66	
Steerable pyramid	4	44.54	46.74	5.72	0.00
Wavelet transform		70.23	69.49	2.25	
Steerable pyramid	5	48.07	50.00	5.68	0.00
Wavelet transform		70.21	70.20	2.39	

Table 6.6: Statistical analysis: mean, median and standard deviation of accuracies obtained when using PCA energy features as input for microcalcification classification, for image size 256×256. A t-test analysis is made between the two decomposition technique, steerable pyramid and wavelet transform.

	No of PCA features	Accuracy			t-test p-value
		Mean	Median	SD	
Steerable pyramid	2	49.47	50.00	8.93	0.05
Wavelet transform		45.74	44.74	8.44	
Steerable pyramid	3	59.21	60.53	9.27	0.50
Wavelet transform		60.16	57.89	6.40	
Steerable pyramid	4	61.05	61.84	9.41	0.81
Wavelet transform		61.42	61.84	6.92	
Steerable pyramid	5	59.53	60.53	9.05	0.62
Wavelet transform		60.37	60.53	9.30	

Table 6.7: Statistical analysis: mean, median and standard deviation of accuracies obtained when using PCA energy features as input for microcalcification classification, for image size 128×128. A t-test analysis is made between the two decomposition technique, steerable pyramid and wavelet transform.

The accuracies obtained when using single feature of energy or entropy as an input for microcalcification classification show that steerable pyramid features are unable to outperform wavelet transform features, specifically when the image sizes were reduced. In the next experiment, both features are combined as the input for microcalcification classification. A pair of the same number from energy and entropy PCA features will be used, ie. 2 energy and 2 entropy, 3 energy and 3 entropy, 4 energy and 4 entropy and finally 5 energy and 5 entropy PCA features.

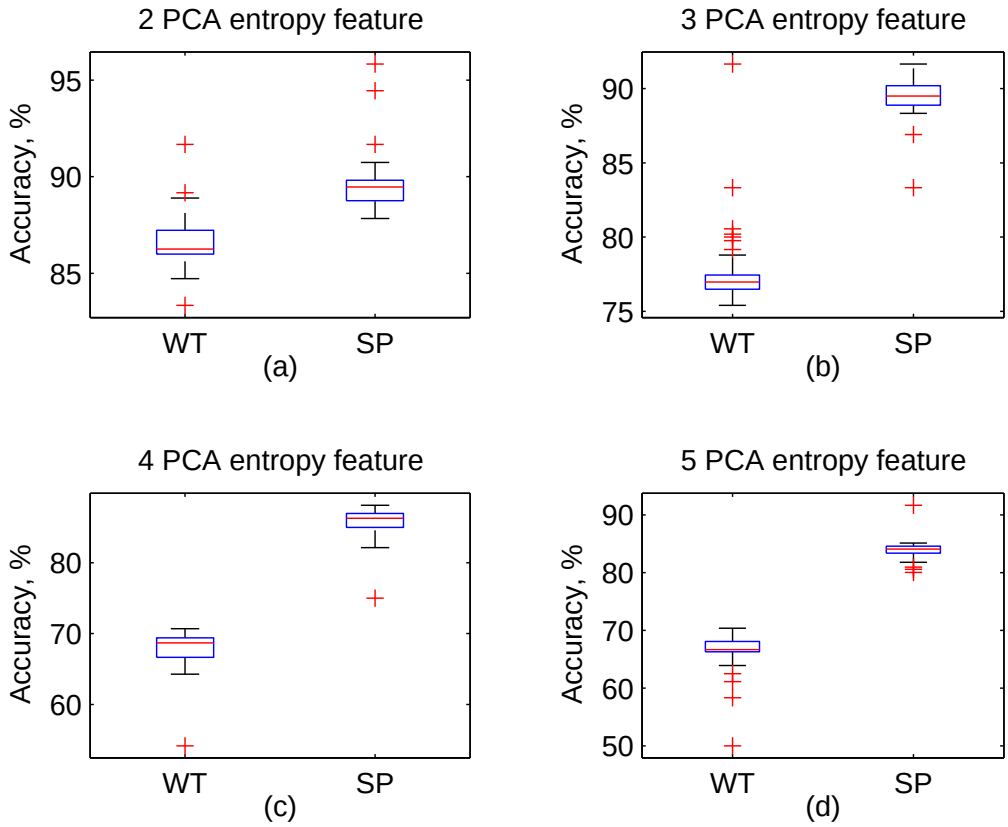


Figure 6.25: Boxplot of accuracy for both wavelet and steerable pyramid with PCA transformed from entropy measures as the single features for classification; with (a) two PCA entropy features, (b) three PCA entropy features, (c) four PCA entropy features and (d) five PCA entropy features, for image pixels 256 × 256.

	No of PCA features	Accuracy			t-test p-value
		Mean	Median	SD	
Steerable pyramid	2	89.58	89.46	1.46	0.00
Wavelet transform		86.60	86.26	1.31	
Steerable pyramid	3	89.46	89.50	1.28	0.00
Wavelet transform		77.79	76.98	2.74	
Steerable pyramid	4	85.85	86.28	2.05	0.00
Wavelet transform		68.03	68.70	2.57	
Steerable pyramid	5	83.92	84.06	2.03	0.00
Wavelet transform		66.65	66.67	3.24	

Table 6.8: Statistical analysis: mean, median and standard deviation of accuracies obtained when using PCA entropy features as input for microcalcification classification, for image size 256 × 256. A t-test analysis is made between the two decomposition technique, steerable pyramid and wavelet transform.

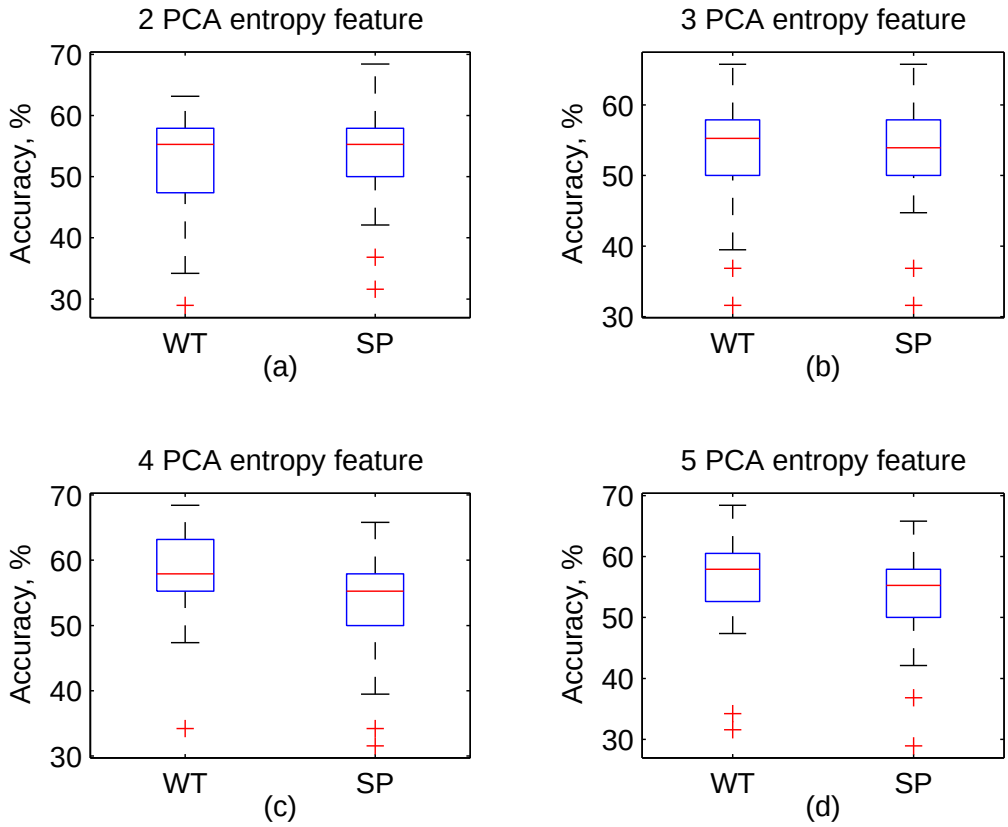


Figure 6.26: Boxplot of accuracy for both wavelet and steerable pyramid with PCA transformed from entropy measures as the single features for classification; with (a) two PCA entropy features, (b) three PCA entropy features, (c) four PCA entropy features and (d) five PCA entropy features, for image pixels 128×128.

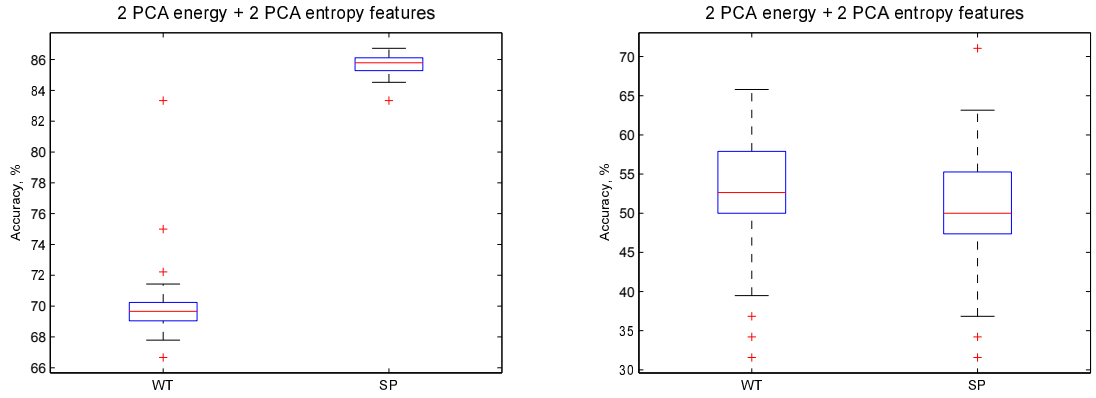
	No of PCA features	Accuracy			t-test p-value
		Mean	Median	SD	
Steerable pyramid	2	53.11	55.26	6.88	0.65
Wavelet transform		52.37	55.26	8.17	
Steerable pyramid	3	54.16	53.95	6.58	0.36
Wavelet transform		53.05	55.26	7.47	
Steerable pyramid	4	53.89	55.26	6.81	0.00
Wavelet transform		57.84	57.89	6.59	
Steerable pyramid	5	53.84	55.27	7.12	0.07
Wavelet transform		56.63	57.89	7.33	

Table 6.9: Statistical analysis: mean, median and standard deviation of accuracies obtained when using PCA entropy features as input for microcalcification classification, for image size 128×128. A t-test analysis is made between the two decomposition technique, steerable pyramid and wavelet transform.

Similar to the previous experiments of using single feature, the accuracies obtained when these 2 sets of input features are used for classification are analysed using boxplot and t-test analysis, to observe if the improvement made by another method is significant.

Combination of 2 energy and 2 entropy PCA features as input for classification

In this experiment, the combination of two PCA features from each decomposition technique; two energy and two entropy features are used for classifications. Thus, the total number of input features is four. This number is chosen from the result obtained in the previous experiment in Section 6.2.3. The accuracy obtained is plotted using boxplot as in Figure 6.27. A t-test analysis is summarised as in Table 6.10 and Table 6.11.



(a) (b)

Figure 6.27: Boxplot of accuracy for both wavelet and steerable pyramid with PCA transformed features for both energy and entropy measures as the features for classification; with two PCA entropy and two PCA energy features, for image size of (a) 256 × 256 and (b) 128 × 128 pixels.

Decomposition technique	N	Mean	SD	DF	p
Steerable pyramid	50	85.52	0.88	49	0.00
Wavelet transform	50	69.97	2.30		

Table 6.10: T-test analysis between steerable pyramid and wavelet transform, for image 256 × 256, with two PCA entropy and two PCA energy features.

The result obtained from this set of features shows that steerable pyramid filtering achieved higher accuracy compared to wavelet transform, with mean of accuracy of

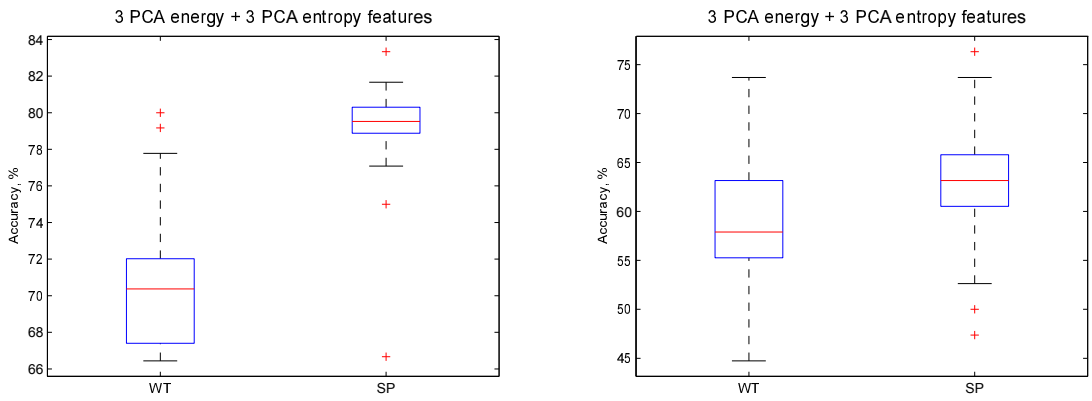
Decomposition technique	N	Mean	SD	DF	p
Steerable pyramid	50	50.63	7.2812	49	0.12
Wavelet transform	50	52.53	8.5225		

Table 6.11: T-test analysis between steerable pyramid and wavelet transform, for image 128 × 128, with two PCA entropy and two PCA energy features.

85.52% and 69.97% respectively. However, when the image size is halved, the mean accuracy of both methods reduced to 50.63% and 52.53%. Although the mean accuracy achieved by the wavelet transform is higher than steerable pyramid, this difference is not significant, as the p-value obtained from the t-test is greater than 0.05.

Combination of 3 energy and 3 entropy PCA features as input for classification

In this experiment, the number of features is increased to three energy and three entropy, hence the total features input is six features. Similarly as in previous experiment, the accuracy obtained is plotted using boxplot as in Figure 6.28 and a summary of the t-test is presented in Table 6.12 and Table 6.13.



(a) (b)

Figure 6.28: Boxplot of accuracy for both wavelet and steerable pyramid with PCA transformed features for both energy and entropy measures as the features for classification; with three PCA entropy and three PCA energy features, for image size of (a) 256 × 256 and (b) 128 × 128 pixels.

In Table 6.12, the results show an improvement when using six PCA features, three energy and three entropy measures for both wavelet and steerable pyramid decomposition technique; 70.54% and 79.25% respectively. Hence, it shows that steerable

pyramid achieved higher mean accuracy compared to wavelet transform when using six features as input for microcalcification classification. Statistically, the p-value obtained from the t-test is less than 0.05, therefore the difference achieved when using features extracted in steerable pyramid filtering is significant compared to wavelet transform. This result is also consistent when the features are extracted from half image size, where accuracy of steerable pyramid filtering achieved is 63.00% and wavelet transform is 58.74% as shown in Table 6.13.

Decomposition technique	N	Mean	SD	DF	p
Steerable pyramid	50	79.25	2.28	49	0.00
Wavelet transform	50	70.54	3.67		

Table 6.12: T-test analysis between steerable pyramid and wavelet transform, for image 256 × 256, with three PCA entropy and three PCA energy features.

Decomposition technique	N	Mean	SD	DF	p
Steerable pyramid	50	63.00	5.8340	49	0.00
Wavelet transform	50	58.74	6.3006		

Table 6.13: T-test analysis between steerable pyramid and wavelet transform, for image 128 × 128, with three PCA entropy and three PCA energy features.

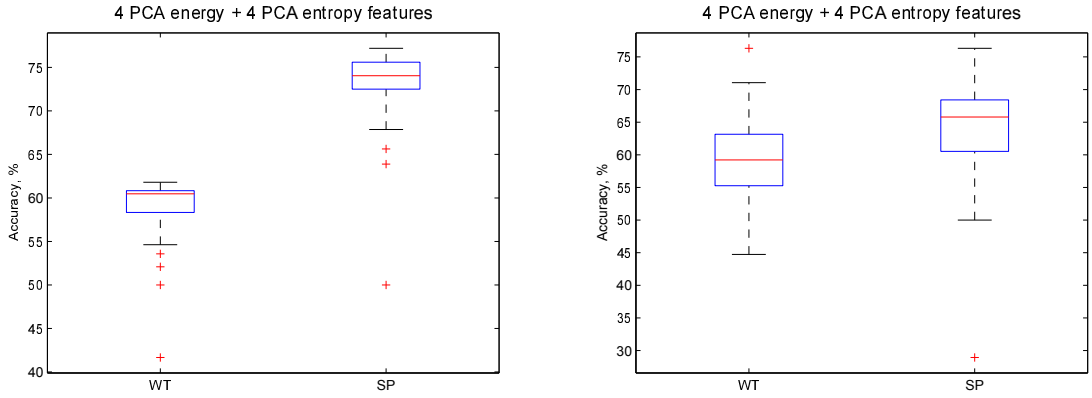
Combination of 4 energy and 4 entropy PCA features as input for classification

Similar method as in previous experiment is used to compare the overall performance of classification when using eight PCA features in total, with four features from energy and four from entropy. The results is presented as in boxplot in Figure 6.29.

The statistical analysis is carried out by performing the t-test analysis and is presented as in Tables 6.14 and 6.15.

Decomposition technique	N	Mean	SD	DF	p
Steerable pyramid	50	74.05	5.46	49	0.00
Wavelet transform	50	60.48	3.62		

Table 6.14: T-test analysis between steerable pyramid and wavelet transform, for image 256 × 256, with four PCA entropy and four PCA energy features.



(a) (b)

Figure 6.29: Boxplot of accuracy for both wavelet and steerable pyramid with PCA transformed features for both energy and entropy measures as the features for classification; with four PCA entropy and four PCA energy features, for image size of (a) 256 × 256 and (b) 128 × 128 pixels.

Decomposition technique	N	Mean	SD	DF	p
Steerable pyramid	50	63.95	8.19	49	0.00
Wavelet transform	50	59.58	6.89		

Table 6.15: T-test analysis between steerable pyramid and wavelet transform, for image 128 × 128, with four PCA entropy and four PCA energy features.

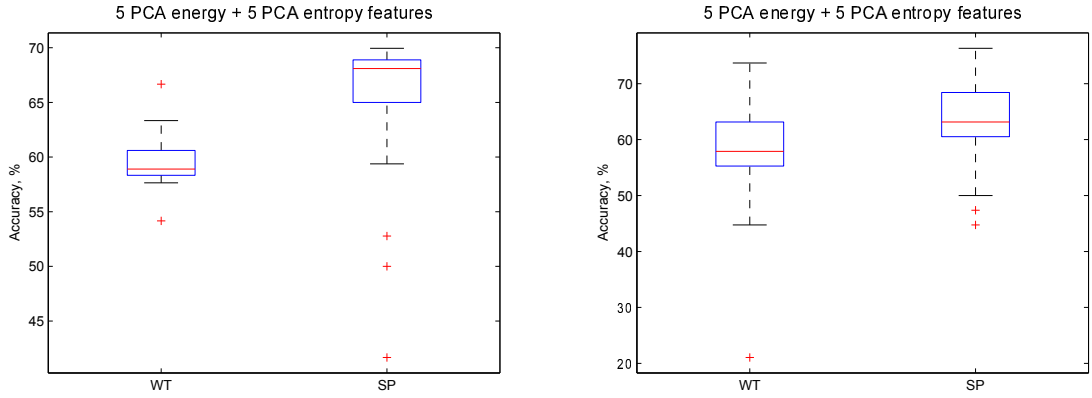
From the result presented, it is shown that the accuracy for steerable pyramid is 74.05%, compared to wavelet transform of 60.48%. It is also shows that steerable pyramid achieved higher accuracy compared to wavelet transform and the difference is significant since the p-value obtained from the t-test analysis is less than 0.05.

Combination of 5 energy and 5 entropy PCA features as input for classification

For the experiment with five PCA energy and five PCA entropy features used for classification, the results obtained are recorded as in boxplotin Figure 6.30. Similarly, a t-test analysis is performed to measure the overall performance statistically.

Decomposition technique	N	Mean	SD	t	DF	p
Steerable pyramid	50	66.00	5.33	3.7491	49	0.00
Wavelet transform	50	59.49	1.81			

Table 6.16: T-test analysis between steerable pyramid and wavelet transform, for image 256 × 256, with five PCA entropy and five PCA energy features.



(a) (b)

Figure 6.30: Boxplot of accuracy for both wavelet and steerable pyramid with PCA transformed features for both energy and entropy measures as the features for classification; with five PCA entropy and five PCA energy features, for image size of (a) 256 × 256 and (b) 128 × 128 pixels.

Decomposition technique	N	Mean	SD	t	DF	p
Steerable pyramid	50	63.11	6.76	3.7491	49	0.00
Wavelet transform	50	58.32	9.15			

Table 6.17: T-test analysis between steerable pyramid and wavelet transform, for image 128 × 128, with five PCA entropy and five PCA energy features.

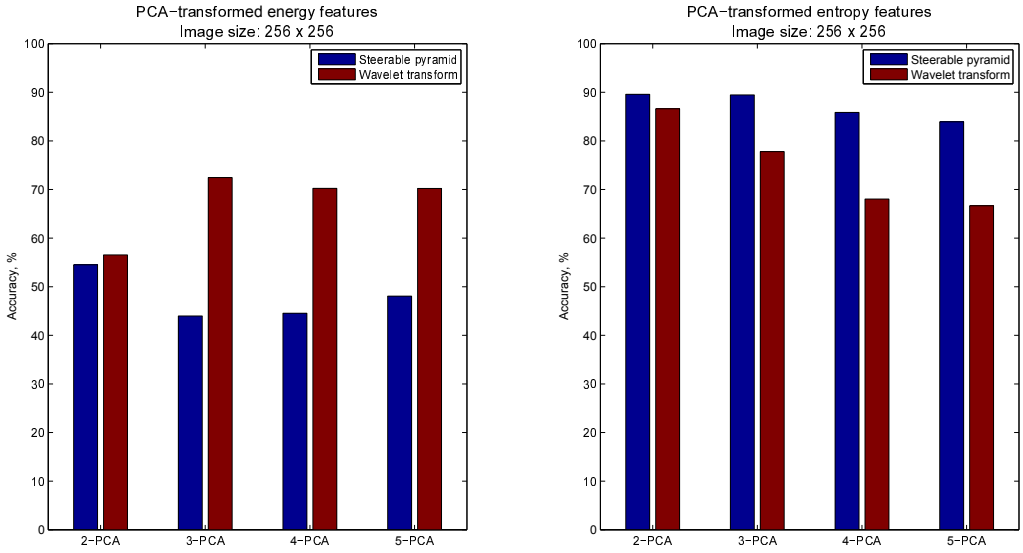
The results obtained as in Table 6.16 and Table 6.17, it is shown that steerable pyramid achieved higher mean accuracy of 66.00% compared to wavelet transform which achieved mean accuracy of 59.49%. This difference is also significant with p-value from the t-test is less than 0.05. The accuracy performances are decreased for both methods when compared to the accuracy achieved by 4 PCA transformed from each energy and entropy features. This shows that the overall performance does not continually improve as the number of features increases.

Summary

The experiments conducted that used PCA transformed energy or entropy as single features for microcalcification classification show that the accuracies improvement is not significant. Hence, the combination of energy and entropy as features is recommended for microcalcification classification.

The results obtained by using reduced number of features found that the best mean accuracy achieved so far is when four features; two energy and two entropy which is 85.52% for steerable pyramid filtering and 69.97% for wavelet transform. The performance achieved when all the 48 features are used as input for steerable pyramid, the mean accuracy obtained is 78.26%. Hence, it is observed that the data dimension reduction using PCA is able to reduce the number of features as input for microcalcification classification. The question of finding the optimal features combination will possibly be an area of investigation in the future work.

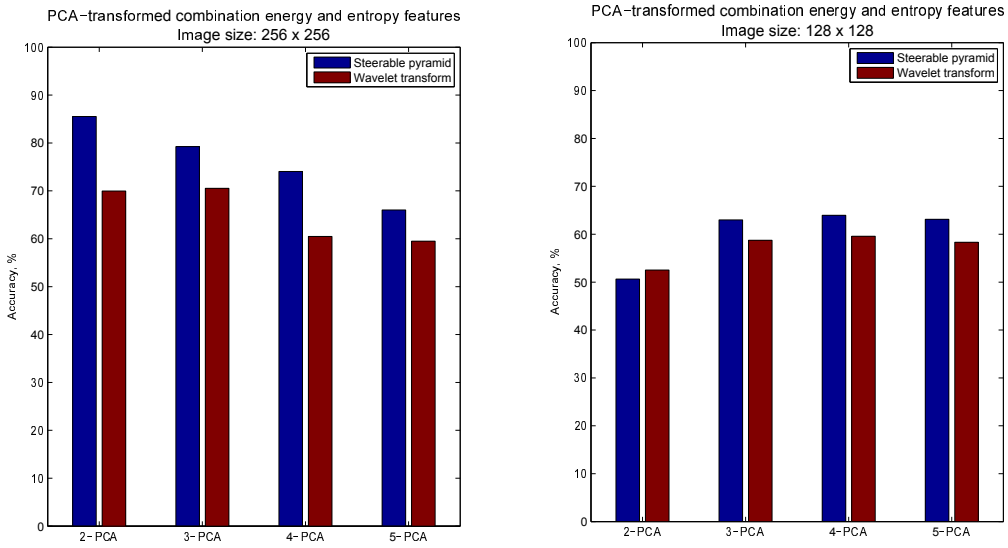
Figure compares the performance of the accuracies obtained when the PCA transformed energy or entropy is used as a single input features. The results show that when the entropy features extracted from the steerable pyramid filtering outperform the classifier's performance when compare to the wavelet transform extraction technique. However, the performance when the energy features are used demonstrates that the energy features contain less malignancy information when the features are extracted from the steerable pyramid filtering.



(a) (b)

Figure 6.31: Comparison chart of accuracy when single features, either (a) energy or (b) entropy were used for microcalcification classification for image pixels size 256 × 256.

In Figure 6.32 shows that the combination of energy and entropy features used to classify the microcalcification cluster achieve a highest accuracy of 85.2% when the 2-PCA transformed features are used. Whereas, the wavelet transform extraction of the same number of features achieve an accuracy of 69.97%. However, when the image pixels size is reduced to 128 $\times$ 128, ie. half of the original image size, the accuracies degraded. This is also consistent with the earlier work where the SNR of the reduced images also decreases as the pixels decreases. Information loss during the resizing process may contribute to the degradation. The reason to use the reduced images is to provide consistency in input size for the following experiments.



(a) (b)

Figure 6.32: Comparison chart of accuracy when energy and entropy features were used for microcalcification classification for image pixels size (a) 256 $\times$ 256 and (b) 128 $\times$ 128.

A work by Dhawan *et al.* (1991) extracted features from wavelet decomposed images and yield ROC performance of 0.74. The 191 images were used in the experiment, and the source of the database was not given. Their work is almost similar to the conducted experiments. However, instead of using SVM, they used a neural network as their classifier. A paper in 2010 (Tirtajaya and Santika 2010) reported the dual-tree complex wavelet technique was used to decompose and extract features from mammograms images, with 57 images from MIAS database were chosen. The SVM classifier

was adapted and they achieved an accuracy of 88.64%. Although the dataset used are different, comparing with these two almost similar works, the result obtained from the conducted experiment has achieved comparable levels of accuracy. In addition, it is also showed that multiple orientation features have contributed useful information for microcalcification classification.

The experiments conducted so far are manually extracted the features; ie. manually computes and limits the features to be either energy, entropy or a combination of both. This limits the potential of other features contain in the ROIs that may contribute to the malignancy analysis. In the next section, a method that be able to extract and reduced the data dimension automatically, using deep belief network, DBN is conducted.

6.3 Automatic Feature Extraction using DBN

In the second part of the chapter, a novel method using deep belief network (DBN) as an automatic feature extractor and data reduction for microcalcification in mammograms diagnosis is investigated. In principle, DBNs enable the automatic extraction of useful features for classifications instead of heuristically chosen one. The use of DBN will possibly reduce the need to select the features from the data, and instead learns from the presented data. As reviewed in Chapter 5, the ability of DBN to map the relationship between the input and output in a highly complex non-linear relationship makes it interesting for classification of microcalfication in mammogram, since there are a variety of shapes and sizes of cluster of microcalcification, which make it difficult to linearly classified them. Therefore, the aim of this section is to seek an automatic model to extract the useful high-level features; such as orientations of the cluster of microcalcification for the classification purposes. It is hoped that these extracted features will possibly be more useful to the classifier compared to the use of raw pixels in the malignancy analysis.

As explained in Chapter 5, the mammogram images have real-valued pixels, thus, the DBN network with Gaussian-Bernoulli RBM at the first layer was implemented. In

this experiment, the DBN is used as feature extractor to determine a suitable features dimension. The DBN can be used as a fully automated classifier by implementing a discriminative DBN, ie. the microcalcification features are extracted by the network and at the final layer of the network is a single node that represents the class of the features. For a DBN to act only as feature extractor, the top layer of RBM is connected to the supervised classifier, chosen to be the SVM classifier. This choice is made to enable comparison with previous technique of feature extraction using steerable pyramid filtering. Moreover, the choice of the network shape is based on the decision to have DBN act as data dimension reducer. Hence, the width of the layer decreases as the layer level increases.

For the work in this thesis, four layers DBN network is chosen to be used for further investigation of extracting features in the microcalcification images. As reviewed in Chapter 2, the use of conventional neural network with at most three layers has gained most interest of researchers in the field. Here, deeper networks for classification of the microcalcification in mammogram are explored. In deep network, it is possible to have multiple hidden layers. In each hidden layer, it will transforms its inputs; that come from previous layer's outputs, by a non-linear transformation. Hence, this allows the deep network to compute much more complex features of the input as the number of the layers in the network increases. As for a start, a network that is slightly deeper that the conventional network, ie. a 4-layer network is developed.

For complex images such as clusters of microcalcification, deep network allows the decomposition of the images to be learn at each layer one at a time. For example, the first layer of the network might learns the image to detect edges. The second layer will learns more complex structure, such as part of the objects. As the network gets deeper, the higher layers will learn much more structures of increasing complexity. It is interesting to explore this property to extract meaningful features from the mammograms, with the eventual goal is to improve the classification accuracy.

The experiments are organised as follows:

1. In Section 6.3.1, experiment with the topology of the network by choosing different number of hidden nodes in each layer of the network is performed. The aim is to build the network by using the sum squared errors in the CD learning as a selection criterion.

2. In Section 6.3.2, the deep network with the chosen topology obtained from experiment in Section 6.3.1 is used to extract features from the microcalcification images and used them as the input for malignancy analysis.

3. In Section 6.3.3, a novel approach for the microcalcification classification using multiple orientations and multiple resolutions DBN is performed.

6.3.1 Experiment 5: DBN Architecture Selection

Similar to experiment in the Section 6.2.1, the initial step is to experiment with few parameters to find the most feasible architecture for the system, particularly for classification of the microcalcification in mammogram. One of the parameters is the number of hidden nodes in each layer of the network. In addition, since the aim of this experiment is to use deep network for microcalcification feature extraction and dimensionality reduction, it is expected that the number of hidden nodes in the layer decreases as the number of layer in the network increases. It is observed the sum squared errors between the original data with the reconstructed data in each network and the time taken to complete the CD learning procedure. Network with the smallest error and converged quickly implies that the network has learnt the features better in a shorter time. The time taken is also observed to ensure the feasibility of the system to complete the task on the machine with the 64-bit core i5 processor, at 1.60 GHz. The reduced size of images that have pixels size of 128×128 are used throughout the DBN experiments, due to the high computational and memory costs on the machine to compute a larger data size.

Number of hidden nodes per layer

In this subsection, the experiment was aimed at choosing the optimal number of hidden nodes at each layer in the DBN autoencoder. The number of hidden nodes at each layer is chosen heuristically by running few experiments and observing the sum squared error. The sum squared error observed is the difference between the original data with the reconstructed data after the CD procedure learning.

Consistent with greedy training in DBN, Consistent with greedy training in DBN, an empirical comparison between several different numbers of hidden nodes at each layer is conducted. The experiments will be carried out individually per layer. The reason for running each layer individually is that the computational cost to run the whole pre-training in DBN is expensive, which would render the approach infeasible to run the whole DBN. This training scheme is also faster. The number of hidden nodes that produces the smallest error and converged quickly, will be chosen as the number of nodes in the visible layer for RBM for the next consequence layer.

Layer 1 The first layer of DBN is directly connected to the image pixels. In this experiment, the size of image is 128×128 pixels. The image is flattened into a horizontal vector, thus, the size of the vector is 16 384, which is also the same size as the visible units in the visible layer. Several number of experiments are performed individually with different number of hidden nodes in the hidden layer of the first layer DBN. There are four different hidden nodes numbers, ie. 5000, 2500, 1500 and 200 connected to all the visible nodes. Since the aim is to extract useful features and at the same time to reduce the data dimensions for classification, the number of hidden nodes must not be higher that then number of the nodes in input layer, and hence the choice of the number of nodes to be experimented. Furthermore, hidden nodes with more than 5000 are computationally expensive.

Hinton *et al.* have provided guidelines in choosing this parameters of DBN. In the guidelines, Hinton mentioned that the number of hidden nodes can be estimated by

finding number of bits that would take to describe the data in the best model. The bits is then multiply with the number of training nodes (Hinton 2012). However, the decision of the parameters choice must be based on the empirical evidence. Hence, the next conducted experiment is to decide the number of hidden nodes that would be suitable in representing the data; ie. microcalcification cluster data images.

In each experiment, the sum squared errors of the original and the reconstructed data are observed. This is to ensure the learning is successfully carried out by observing the errors to decrease monotonically as the epochs progress.

Since the input or the visible layer is not binary, ie. the real value of the pixel images, the energy function used for the first layer of DBN is defined by the Gaussian-Bernoulli connection.

$$E(v,h) = -\sum_{i \in \text{visible}} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{j \in \text{hidden}} b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij}, \tag{6.18}$$

where v_i, h_j are linear Gaussian state of visible unit i and binary state of hidden unit j respectively, and w_{ij} is the weight between them; a_i, b_j are their biases and σ_i is the standard deviation of Gaussian for visible unit i.

The layer is pre-trained for 200 epochs using the 80 training sets of images as described in Section 6.1.5. This number of training sets is also consistent with the previous experiments in Section 6.2. The sum squared error, SSE for each epoch is calculated using Equation 6.19:

$$\text{SSE} = \sum \langle \text{data} \rangle^2 - \langle \text{model} \rangle^2. \tag{6.19}$$

The errors are plotted in Figure 6.33. In the figure, the errors are plotted for the first 100 epochs. It is observable that the errors continue to decrease until the final 200 epochs. The errors are presented in Table 6.18, in steps of five epochs since there are no changes in the trend of error. For reference, the complete table of each error is presented in Appendix B.

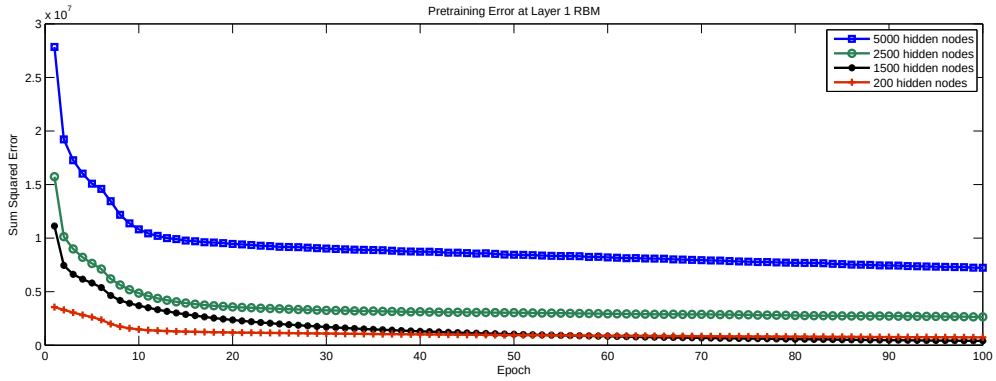


Figure 6.33: Sum squared errors in pre-training the first layer RBM after 200 epochs with 5000, 2500, 1500 and 200 hidden nodes. The visible layer of the RBM is connected to the pixels of the microcalcification images.

It is observed that from the result as shown in Figure 6.33, the sum squared errors of all experiments showed a monotonic decay after at least 30 epochs. Hence, it is sensible to stop the training at 200 epochs since the error will not rise up again. From the result, it is also observed that when 1500 hidden nodes are used in the first layer, the sum squared errors showed the lowest error after 70 epochs.

Hidden nodes	5000	2500	1500	200
Epoch	1.0e+ 06 *	1.0e+ 06 *	1.0e+ 06 *	1.0e+ 06 *
5	16.2534	10.3748	7.6277	2.4275
10	13.7639	6.9028	4.9115	2.2852
15	11.0566	6.7831	4.2703	2.0345
20	10.1166	5.0290	3.9672	1.7644
25	8.8350	4.5512	3.0274	1.5475
30	8.2715	4.1921	2.6889	1.4036
35	7.6439	3.8324	2.5912	1.3180
40	7.1991	3.7066	2.3868	1.2588
45	6.8503	3.4727	2.2724	1.2120
50	6.5677	3.3347	2.2075	1.1681

Hidden nodes	5000	2500	1500	200
Epoch	1.0e+ 06 *	1.0e+ 06 *	1.0e+ 06 *	1.0e+ 06 *
55	6.3546	3.1859	2.1193	1.1251
60	6.1546	3.0617	2.0406	1.0868
65	6.0081	2.9381	1.9788	1.0521
70	5.8816	2.8407	1.9140	1.0218
75	5.7782	2.7490	1.8535	0.9957
80	5.7182	2.6572	1.7939	0.9725
85	5.6740	2.5990	1.7419	0.9527
90	5.6099	2.5216	1.6900	0.9349
95	5.5886	2.4676	1.6464	0.9197
100	5.5594	2.4079	1.6046	0.9053

Table 6.18: Pre-training error in the first layer up to epoch 100 in step of 5 epochs.

The execution times for each network are also been recorded. From Table 6.19, it is observed that as the number of hidden nodes increases, the time taken for the RBM learning also increases. 5000 hidden nodes shows the most expensive computational time as it took nearly 8 hours to complete the pre-training procedure. The 200 hidden nodes shows the least time taken to complete the procedure. From Table 6.19, it is observed that the time taken to complete the procedure is approximately linearly proportional with number of hidden nodes.

Number of hidden nodes	5000	2500	1500	200
Time taken (s)	31351.45	11841.92	5122.68	486.70

Table 6.19: Execution time for pre-training Layer 1 RBM after 200 epochs

From this experiment, the first layer is chosen to have 1500 nodes. This is based on the lowest error achieved compared to the other 3 numbers of hidden nodes. Furthermore, the time taken to complete the procedure is acceptable, ie. 85 minutes. Although 200 hidden nodes shows the least time taken, ie. 8 minutes to complete the procedure, the number of hidden nodes from visible nodes is drastically reduced. This can potentially lead to a lot of information lost from that layer. In addition, the training time taken to complete 200 epochs does not affect the overall performance of the DBN during the testing phase.

Layer 2 Next, the output from these 1500 nodes is used as an input to the next RBM layer. At this stage, the architecture of the DBN is 16 384 (input) - 1500 (layer 1) - layer 2. Again, four different numbers of hidden nodes were experimented. The number of hidden nodes are 1500, 1000, 500 and 200. These numbers are chosen to be no greater than the number of hidden nodes in the previous layer, ie. 1500 nodes. This is because the aim of the network is to reduce the dimension of the data, which the nodes at the top of network will be presented as the input features for the classification. Since the output from the previous layer is in a binary state, hence the energy connection of the layers now is Bernoulli-Bernoulli. Therefore, the Equation 6.20 is used to define the weights between the two layers:

$$E(v,h) = -\sum_{i \in \text{visible}} a_i v_i - \sum_{j \in \text{hidden}} b_j h_j - \sum_{i,j} v_i h_j w_{ij}, \tag{6.20}$$

where v_i, h_j are binary state of visible unit i and hidden unit j, and w_{ij} is the weight between them. a_i, b_j are their biases.

Similarly, the sum squared errors are observed and the execution time for each run of experiment is recorded. Figure 6.34 shows that the sum squared errors for these four

experiments converged towards zero at around 40 epochs. The result shows there is not much difference in the total error for four different set of RBM. The detailed results of the pre-training errors are also provided in Appendix B.

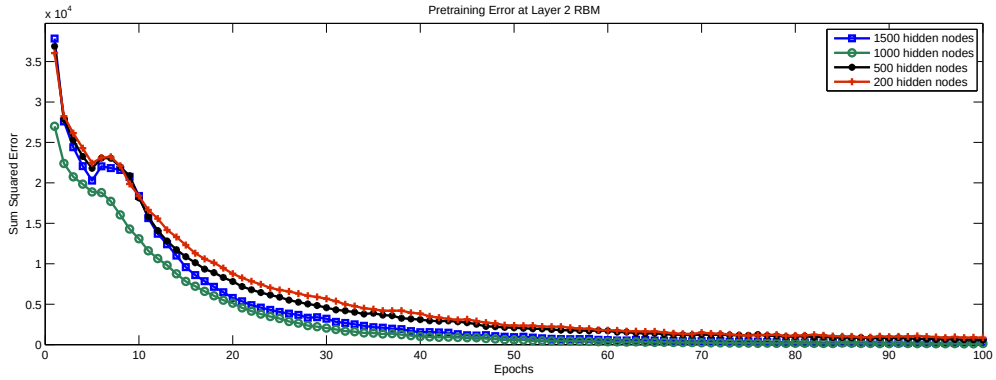


Figure 6.34: Sum squared errors in pre-training the second layer RBM after 200 epochs. The visible layer of the second RBM is the output of the 1500 hidden nodes from previous layer. The number of hidden nodes experimented are 1500, 1000, 500 and 200.

Number of hidden nodes	1500	1000	500	200
Time taken (s)	208.12	142.18	77.28	43.65

Table 6.20: Execution time for pre-training Layer 2 RBM after 200 epochs

The execution times for each network are also recorded. From Table 6.20, it is observed that the results are similar to the experiment conducted in the first layer. It is also observable that as the number of hidden nodes increases, the time taken for the RBM learning also increases. 1500 hidden nodes shows the most expensive computational time as it took the longest to complete the pre-training layer 2 procedure, ie. 3.46 minutes. Likewise, the layer with 200 hidden nodes shows the least time taken to complete the procedure, ie. 43.65 seconds. However, the time taken for all different experiments is not as crucial as in pre-training the first layer, which is directly connected to the visible input.

Hence, unlike in the first layer, the error may no longer be useful as a criterion in choosing the number of hidden nodes in layer two, since the error for all four different RBMs converged to zero, at least after 40 epochs. To avoid a drastic drop in number of

hidden nodes between layer, which may lead to some information loss, the number of hidden node chosen for layer two is 1000. Thus, up to this point, the second layer of the DBN is chosen to have an 16 384 (input) - 1500 (layer 1) - 1000 (layer 2) architecture. Since the aim is to have a deeper network, the similar procedures for the following layer, ie. third layer is performed.

Layer 3 For the third layer in DBN, the pre-training procedure is similar to the procedure in the second layer. The visible and hidden nodes of the RBM is in binary state, thus the energy function as in Equation 6.20 is used to define the weights connection. The result for the sum squared errors are recorded as well as the execution time for each experiment with different number of hidden nodes in the third layer ie. 1000, 500 and 200. Again, the numbers are chosen to be no greater than the number of nodes in the previous layer, ie. second layer. As in Figure 6.35, it is observed that the total error converged quickly to zero at around 20 epochs.

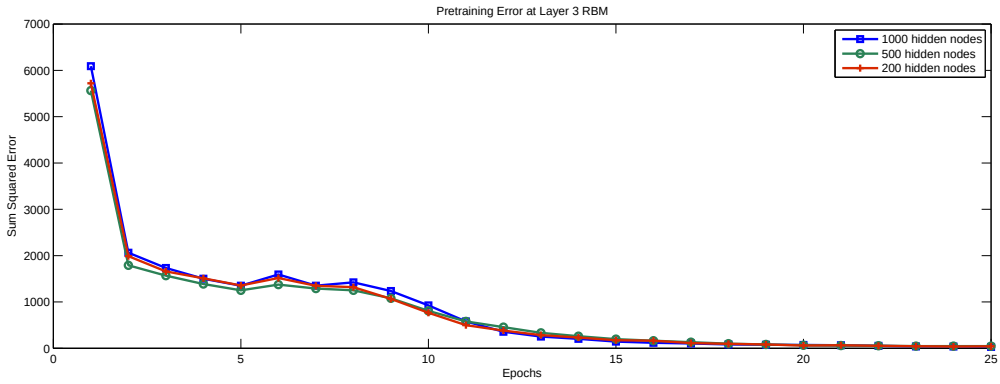


Figure 6.35: Sum squared errors in pre-training the third layer RBM after 200 epochs. The visible layer of the second RBM is the output of the 1000 hidden nodes from previous layer. The number of hidden nodes experimented are 1000, 500 and 200.

Number of hidden nodes	1000	500	200
Time taken (s)	95.94	69.22	51.57

Table 6.21: Execution time for pre-training Layer 3 RBM after 200 epochs

The result obtained from this experiment shows that 500 hidden nodes has the lowest error after 5 epochs. However, after 10 epochs, the pre-training error from all three

hidden nodes shows almost similar result, ie. 0.5×10^3. For the third layer analysis, the time taken to complete the layer pre-training procedure is also proportional to the number of hidden nodes in the layer. 1000 hidden nodes took the longest amongst them, 95.94 seconds. 500 and 200 hidden nodes took 69.22 and 51.57 seconds to complete the procedure respectively, which are not as crucial as time taken for pre-training the first layer of DBN. The time taken is summarised in Table 6.21.

The errors obtained were similar to the errors in the second layer experiment, hence, may no longer be useful in choosing the number of hidden nodes. Furthermore, the errors converged so quickly may suggest that the three layers architecture may not be very important to the overall DBN routine. However, to explore a deep network, the number of hidden nodes is chosen to be 500 nodes, so that the difference between this third layer and the second layer is consistent with the difference between the second layer and the first layer.

Final DBN architecture There are many possible of combinations of number of nodes per hidden layer in DBN as well as number of hidden layer in the DBN. The experiments conducted earlier is one of several sensible methods to choose the appropriate number of hidden nodes in each layer in DBN. Previous works as reviewed in Chapter 2 have shown that conventional neural network with one hidden layer is amongst the common feature extractor and classifier in the classification of microcalcification in mammogram. Hence, a slightly deeper network with three hidden layers DBN will be used as the feature extraction technique.

Summary

From these experiments, the final architecture of the DBN with three hidden layers is chosen to be 1500-1000-500, as illustrated in Figure 6.36. The code layer contains the features that are extracted and is used as an input for the classification. The number of code layer will be experimenting in the following Section 6.3.2.

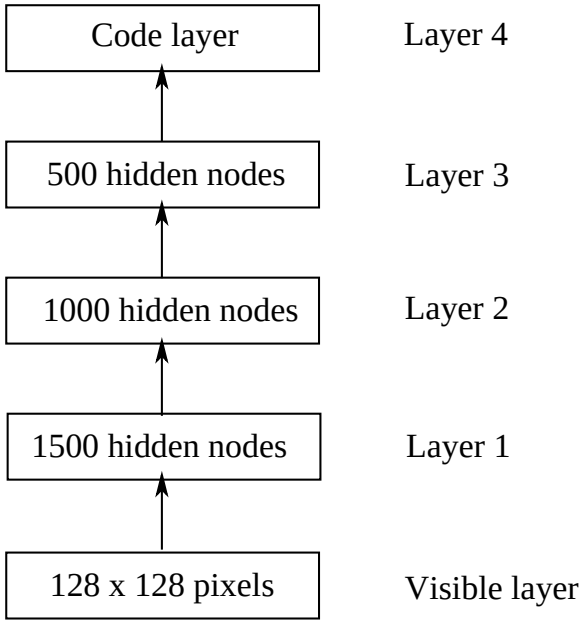


Figure 6.36: The first 3 hidden layers of DBN architecture that will be used in the thesis. The code layer contains the features that are extracted and is used as an input for the classification of microcalcification clusters in mammogram.

6.3.2 Experiment 6: Feed-Forward DBN Feature Extraction for Microcalcification Classification

Following the previous work in Section 6.3.1 that establishes a DBN architecture, which can learn patterns through greedy pre-training of the layers, this section of the thesis discusses attempts to fully utilize the DBN as an automatic feature extractor. To classify the extracted features, two techniques are proposed. One technique is to stack one layer containing single node at the top of the features or code layer. This single node will activate and predict the class of the microcalcification cluster without the information of their class. Hence, the DBN will act as an unsupervised DBN. Another technique is to use SVM as a classifier, and classifies the features that had been extracted by the code layer. This will produce a supervised DBN. It is assumed that the code layer contains the compressed information of the features. The compression will reduce the gigantic size of the data to be presented as the input to the SVM classifier.

Feed Forward DBN as automatic features extractor and unsupervised classifier

In this experiment, DBN is used to extract features directly from the images of the ROI. Each layer in the DBN extracts features that will be the input to the next layer. The method to extract the features is part of autoencoder DBN. The ROI images were decoded at the first three layers. The subsequent layer contains the features extracted or decoded from these layers, hence the layer is named *code layer*. For a full autoencoder, these features were then encoded at the next following three layers. The weights in the encoder part for these layers are the transpose of the weights in the decoder part. To complete the process, the top layer of DBN reconstructs the original images at the input.

However, in this work, the interest is to use the features that had been extracted in the code layer for the classification. The code layer contains the compressed information obtained from the ROI images at the input. This reduces the data dimensions from raw pixel images, which is also act as to reduce the dimension of the data. Thus, in this experiment, only the decoder part is performed and the features extracted from the decoder part is used to classify the images.

The aim in this experiment is to investigate the ability of DBN to act as an automatic feature extractor and unsupervised classifier. In the first layer, the features are extracted directly by the DBN from the raw image pixels. In the following layers in DBN, the number of nodes is reduced as passing through to the next stacked layer in DBN. The features are said to be automatically extracted. Moreover, since the class information is not included during the pre-training, the classification made by DBN is performed in an unsupervised manner. The advantage of using this method is that the whole image pixel is used for feature extraction purposes. In contrast, conventional feature extraction methods require the measures to be predetermined, instead of data driven.

Several different number of code layer nodes are investigated in the experiment. The numbers of nodes used in this experiment are 10, 30, 50, 100, 300 and 500. The aim is to

Code layer nodes	TP	TN	FP	FN	Execution time (s)	TPR	TNR
10	23	45	29	21	3777.34	0.4423	0.6818
30	34	34	32	18	3551.31	0.6538	0.5151
50	14	33	38	33	4625.46	0.2692	0.5000
100	20	35	32	31	7209.79	0.3846	0.5303
300	34	32	18	34	16619.62	0.6538	0.4848
500	32	29	20	37	23898.76	0.6154	0.4394

Table 6.22: Summarised results using different number of code layer nodes. The number of code layer nodes represent the number of extracted features for classification process using unsupervised feed forward DBN.

reduce the dimension of the features, hence the number of the nodes in the code layer is smaller than the number of nodes in the previous immediate layer, ie. 500 nodes in layer 3 of the DBN. The summary of total condition output from the DBN; true positive (TP), true negative (TN), false positive (FP) and false negative (FN) is presented in the Table 6.22. Using Equation 6.6 and Equation 6.7, the TPR and TNR are calculated. The sensitivity (TPR) and the specificity (TNR) of the system for each experiment are compared.

The results are shown in Table 6.22. The highest TPR is achieved when the code layer is 300 nodes, ie. 0.6538, which is also the same when 30 nodes is used in the code layer. Other number of code layer that achieved TPR higher than 0.5 is 500. The curves of the TPR vs TNR for nodes achieved TPR more than 0.5 are plotted in Figure 6.37.

The results from the ROC curve shows that a DBN with 300 nodes in the code layer achieved the best performance compared to the 30 and 500 nodes in the code layer. However, the costs of computation time for completing the task using the larger number,ie. 300 and 500 nodes are greatly increased. Therefore, it is feasible and sensible to choose 30 nodes in the code layer.

The parameters for the first three hidden layers architecture are defined in Section 6.3.1. In addition to these parameters, another parameter at the top of the DBN that contains one node is attached. This node is activated by the features in the code layer and will giving the output of 0 or 1. The architecture of DBN as feature extractor and unsupervised classifier is illustrated in Figure 6.38.

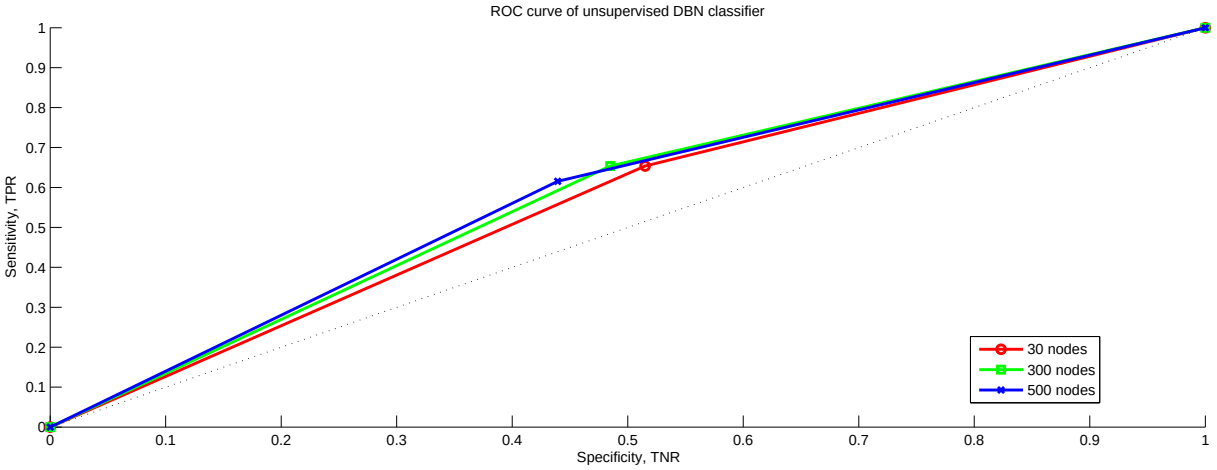


Figure 6.37: ROC curve of feed-forward unsupervised DBN with 30, 300 and 500 number of nodes in the code layer.

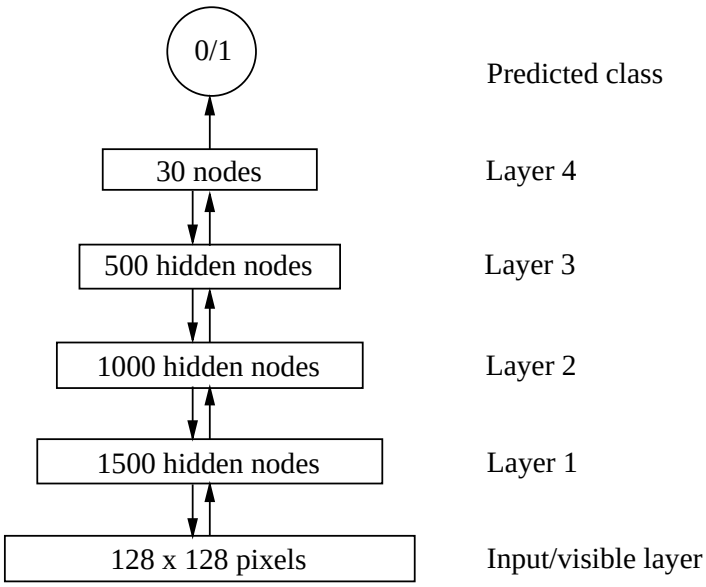


Figure 6.38: DBN as feature extractor and classifier. The code layer contains the extracted features and is used for classification of microcalcification in mammogram, and the final layer has one hidden node for class prediction.

The procedures in this experiment that uses DBN as feature extractor and as an unsupervised classifier are as follows: 1. The images of microcalcification clusters in mammogram are directly connected to nodes in the first layer. The nodes are the input of the DBN, without giving any information of the cluster class, either benign or malignant. 2. These features are passing through to the sequence of layers. At each layer,

more complex features are extracted. 3. At the final layer of the DBN, there is a single node that clusters and classifies the extracted features in the code layer to give the final class label. The result of the classifier above assign the value 0 for benign and 1 for malignant for sensitivity and specificity calculation.

From this experiment, it is found that the unsupervised classification is unable to classify the microcalcification clusters in mammograms. However, the result from this experiment can be used to determine the structure of the DBN that is suitable and feasible to use for future experiments. Continuing to the next experiment, a four layers DBN attached with a supervised classifier, SVM is proposed for microcalcification cluster classifications. This structure is illustrated as in Figure 6.39. The reason to choose SVM as the classifier is to provide a consistency for the classification process throughout the thesis.

Feed-forward DBN as features extractor with SVM classifier

In feed-forward DBN experiment, a supervised classifier, SVM is attached at the top of the DBN to learn the features for classification. The aim is to investigate whether the features that are extracted from the raw images using DBN contain useful information to differentiate between benign and malignant cases, by using a supervised classifier.

The architecture of the DBN is identical to the previous experiment in Section 6.3.2. The DBN has 1500, 1000 and 500 number of hidden nodes at the first, second and the third hidden layers, respectively. The result obtained from the previous experiment in Section 6.3.2 shows that there is no distinct trend for the relationship between the number of hidden nodes and the accuracy achieved. However, only 30 nodes are used for the code layer. The main reason is to provide an acceptable number of features as input to the classifier, SVM and to prevent the curse of dimensionality. Although this number of features is still large, it has been reduced from 16384 number of pixels, which made it reasonably acceptable for presentation to the classifier. The 4-layer DBN architecture is illustrated in Figure 6.39.

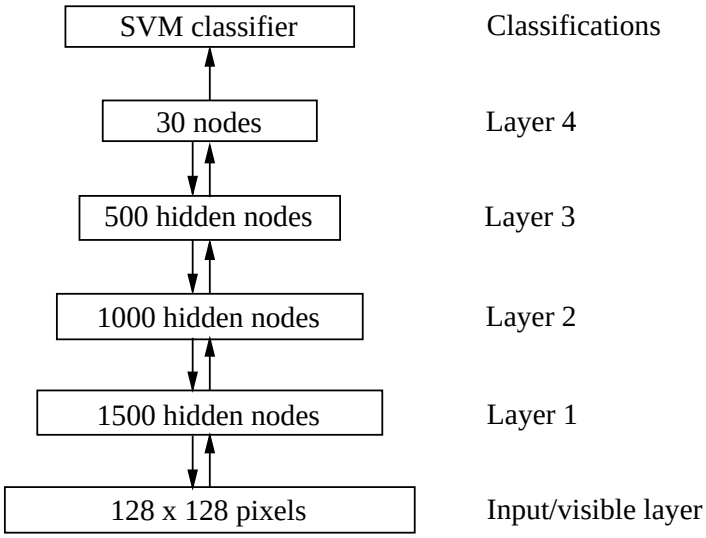


Figure 6.39: DBN as feature extractor and SVM for classifier. The code layer contains the extracted features and is used for classification of microcalcification in mammogram by a SVM classifier.

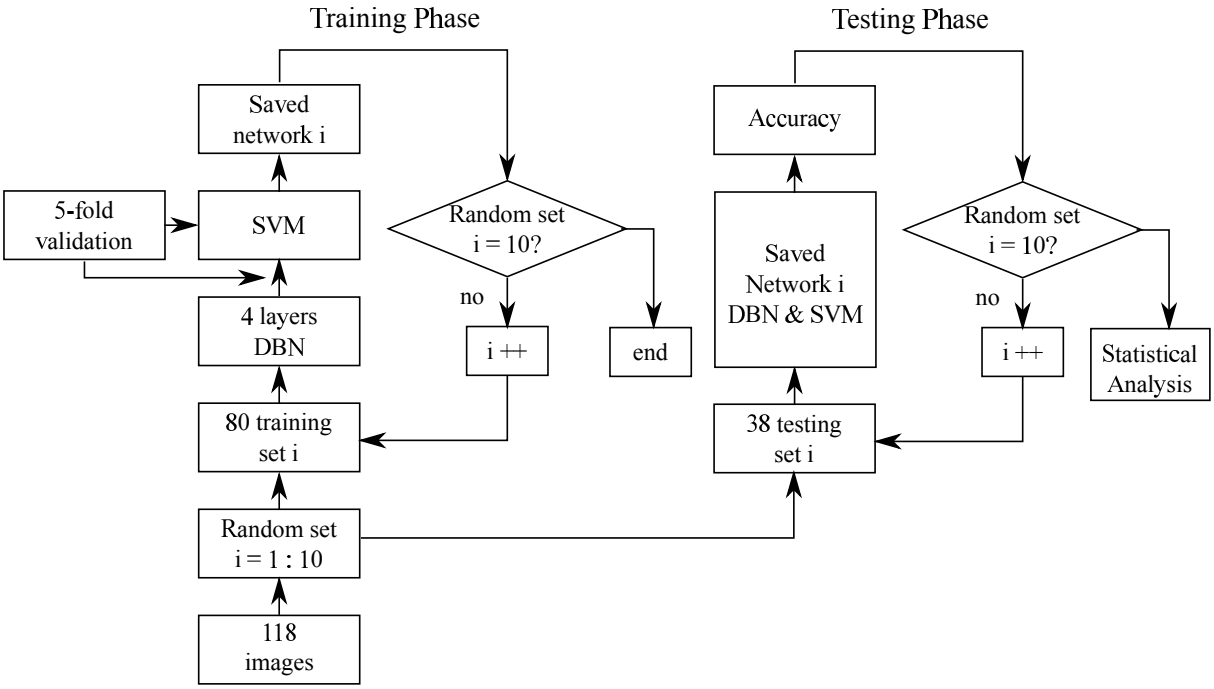


Figure 6.40: Flowchart of the DBN-SVM training and testing phase.

The flowchart in Figure 6.40 illustrates the overall process of this experiment. The 118 images in the database will be randomly divided into two sets, training and testing sets. Ten different sets are generated, hence this produces ten different sets of training

images, and ten different sets of testing images. In training phase, the first training will be pre-trained by the proposed four layer DBN, where its architecture is as follows: 1500 nodes in the first hidden layer, 1000 nodes in the second hidden layer, 500 nodes in the third hidden layer and 30 nodes in the code layer. The 30 nodes in the code layer are the features that are extracted and assumed to contain the compressed data of the original images. These are the features that will be presented as the input for the SVM classifier. Next, in SVM, these features will be trained, and using the 5-fold validations method to set the parameters in the SVM. In testing phase, the testing datasets are used, and these images will use the learned network in DBN and SVM to predict the output. The accuracy from the output is then calculated. These training and testing phases will be repeated for the next nine random set of training and testing datasets. At the end of the process, the performance of the system is then validated by the statistical analysis.

The summary of the process of the classification is as follow:

Input database The raw pixels of 118 ROIs that contain microcalcification clusters that have been segmented as described in Section 6.1.3 are used as the input for the system.

Training input 80 of 118 images are randomly chosen with 40 benign and 40 malignant cases for training sets. This is repeated 10 times, with different training sets of 80 generated each time.

Training phase 80 training images are attached to the lowest layer of DBN, ie. the visible layer. The CD learning is used to learn the weights between this layer and the next hidden layer, which has 1500 nodes. The output of this hidden layer will be the input to the next 1000 nodes hidden layer. Similar process for the third and fourth layers which has 500 and 30 nodes, respectively. The final output of the DBN, ie. 30 features will be the input to train the SVM classifier for tuning the parameters using 5-fold validation method.

Testing input The remaining 38 images for each run will be the testing set used to measure the sensitivity and the specificity of the proposed system.

Testing phase The learned network is tested with the 38 testing images, repeated 10 times with different testing sets.

The accuracy of correct classification from each network is recorded. To visualise the result obtained, a boxplot is generated and presented in Figure 6.41. From the analysis, it is observed that the variance in the result is high, where the standard deviation of the 10 results is 17.28. The median of the accuracy achieved is 51.32% and the mean accuracy is 47.90%.

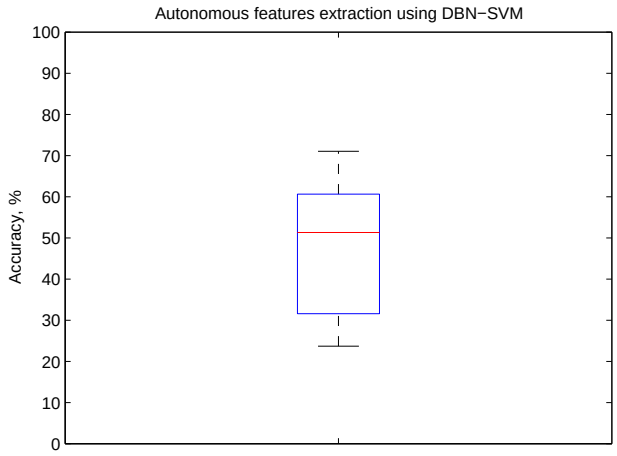


Figure 6.41: Boxplot of accuracy for feature extraction using DBN directly from the raw pixels of segmented ROI with SVM classfier.

Summary

From these experiments, it can be concluded that the extracted features from raw image pixels using DBN are not significant to differentiate between benign and malignant classes. Although in the second experiment, a supervised classifier, SVM is used to classify the DBN extracted features, the classification rate has not been significantly improved. The raw images of mammogram that are low contrast and the appearances of cluster of microcalcification in mammogram are vague and not well defined

may possibly contribute to this unsatisfactory result. Hence, to further investigate the use of DBN for microcalcification classification, a novel approach that combines multi-orientation and multi-resolution filtered images of cluster with a deep network for classifications is proposed and to be experimented next.

6.3.3 Experiment 7: Microcalcification Classification using Multiple Orientation and Multiple Resolution DBN

The results obtained from the experiments in Section 6.3.2 show that the features extracted directly from raw images using DBN are not significantly able to distinguish the microcalcification cluster features and classify them into benign or malignant classes. Hence, in this experiment, an image filtering process is performed before the DBN is used to extract the features. The proposed filtering method, ie. the steerable pyramid, is used to produce filtered images at different orientation and resolution. The idea is to extract more features at multiple resolution and multiple orientation that can potentially contribute towards the improvement of the system sensitivity. Although the filtered images are used instead of raw images, this experiment does not entirely defeat the initial goal of automatic features learning, as it still automatically extracted features from the filtered images.

The process in the proposed system follows the flowchart shown in Figure 6.40. However, a pre-processing step where the images are filtered using steerable pyramid filtering is added before the extraction using DBN is performed.

Using steerable pyramid method, the images were filtered at three different levels of resolution and eight orientations at each level. Thus, the total filtered images are 24 images. Since the filtered images produced are at different resolutions, the sizes of the images vary for each resolution. At resolution 1, the size of the filtered images are 128×128 pixels, whilst at resolution 2 and 3, the filtered images has size of 64×64 and 32×32 pixels, respectively. Thus, the number of hidden nodes is adjusted to be proportionally with the size of the input image at the first visible layer of DBN.

For images with size of 128×128, the same DBN architecture is adapted, ie. four-layers DBN with 1500-1000-500-30 nodes. As the size of the image is reduced, the number of hidden nodes is also reduced. Hence, for the 64×64 pixels images, the four-layers DBN has 750-500-250-15 nodes. To approximately preserve the layer size ratio, the 4-layer DBN has 375-250-125-8 hidden nodes for the 32×32 pixels images. The reason for the different number of hidden nodes to be scaled down accordingly with respect to the image sizes is to reduce the possibility of extracting redundant features if the number of nodes is much larger than the number of pixels. Furthermore, if the number of hidden nodes are too much smaller than the input layer, there will be information loss due to the abrupt changes.

For each resolution, the DBN network is trained separately because the number of the hidden nodes in the hidden layers are different. For the first resolution, eight filtered images are trained using 1500-1000-500-30 DBN. This produces 30 features to be input to the SVM classifier. The process is repeated 10 times with 10 different randomly selected sets of training and testing data. The accuracies obtained after 10 sets of classification process were recorded and presented as in boxplot shown in Figure 6.42. The mean accuracy achieved is 64.47% with median of 67.11%. The results have shown improvement in the accuracy when compared to the previous experiment in Section 6.3.2, when the input of the network is directly connected to the raw pixels of images.

The process is also repeated for images filtered at resolution two. Similarly, 10 different randomly chosen set of training and testing data are used to analyze and validate the results. However, the difference is the size of the DBN, where the nodes in this four-layer DBN has 750-500-250-15 number of nodes. Hence, the number of features that are extracted from resolution two of the images is 15. The boxplot is shown in Figure 6.43. The mean of accuracy achieved is 57.63% with the median of 56.58%.

For images filtered at resolution three, the size of network in DBN is reduced to 375-250-125-8 nodes, as reason mentioned before. Hence the number of features extracted is 8. Again, the process is repeated for ten times and the accuracy is plotted and shown

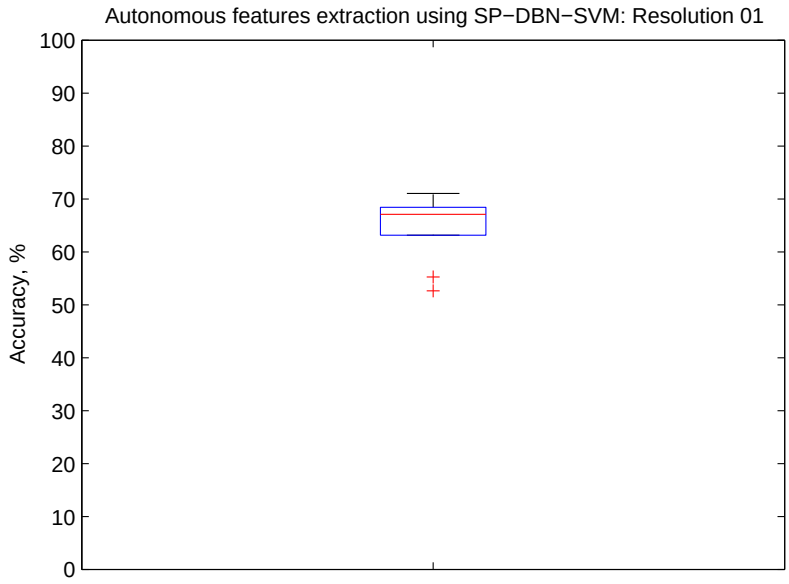


Figure 6.42: Boxplot of accuracy achieved in a hybrid system of steerable pyramid filtering, DBN and SVM for microcalcification classification. The input images are the filtered images at resolution 1.

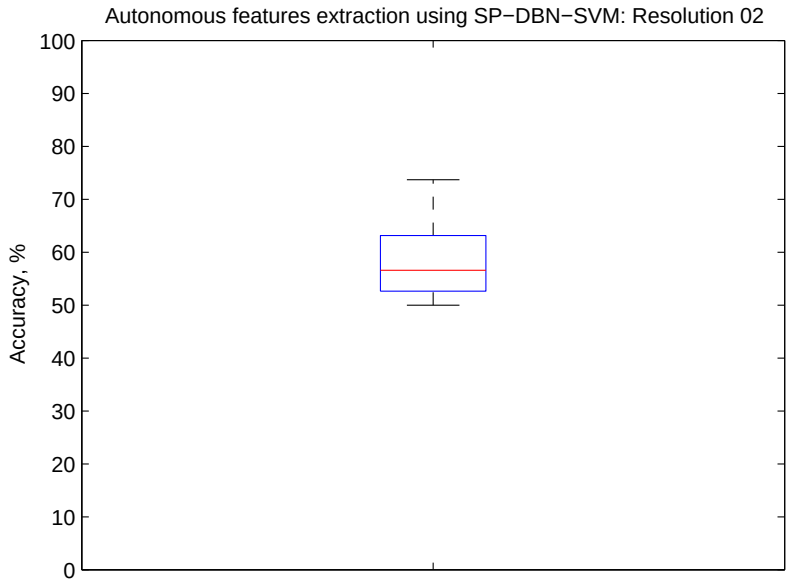


Figure 6.43: Boxplot of accuracy achieved in a hybrid system of steerable pyramid filtering, DBN and SVM for microcalcification classification. The input images are the filtered images at resolution 2.

in Figure 6.44. The mean and median of accuracy achieved are 54.33% and 55.63% respectively.

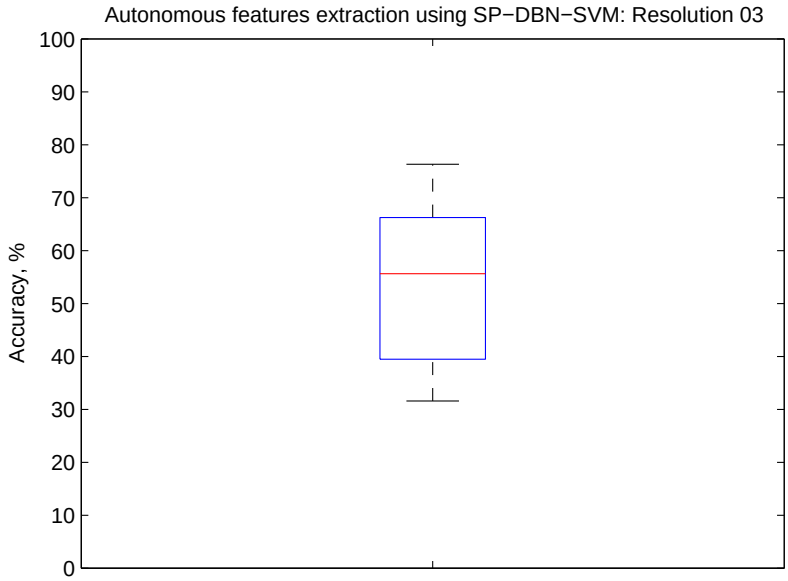


Figure 6.44: Boxplot of accuracy achieved in a hybrid system of steerable pyramid filtering, DBN and SVM for microcalcification classification. The input images are the filtered images at resolution 3.

The results are summarised as shown in Table 6.23. From the results, it is observed that higher accuracy is achieved when the input for classification are the filtered images using steerable pyramid filtering at first resolution. As the resolution increases, the accuracy achieved is slightly decreased. At the smaller resolutions, the images will be more compressed and contained less information which lead to the deteriorating in classifications accuracy.

Image input	Features nodes	Accuracy, %	
		Mean	Median
Raw pixels	30	51.32	47.90
Filtered images at resolution 1	30	64.47	67.11
Filtered images at resolution 2	15	57.63	56.58
Filtered images at resolution 3	8	54.33	55.63

Table 6.23: Summary results to compare the accuracy achieved with different input images to the DBN.

Next, all three resolutions are combined to obtain features to use for classification. However, the total features will then be 53 nodes for one orientation. Since there are eight different orientations, the total features would be 424 features. To reduce the number of features, one layer DBN will be use with hidden nodes of 30. The architecture of this DBN is illustrated in Figure 6.45.

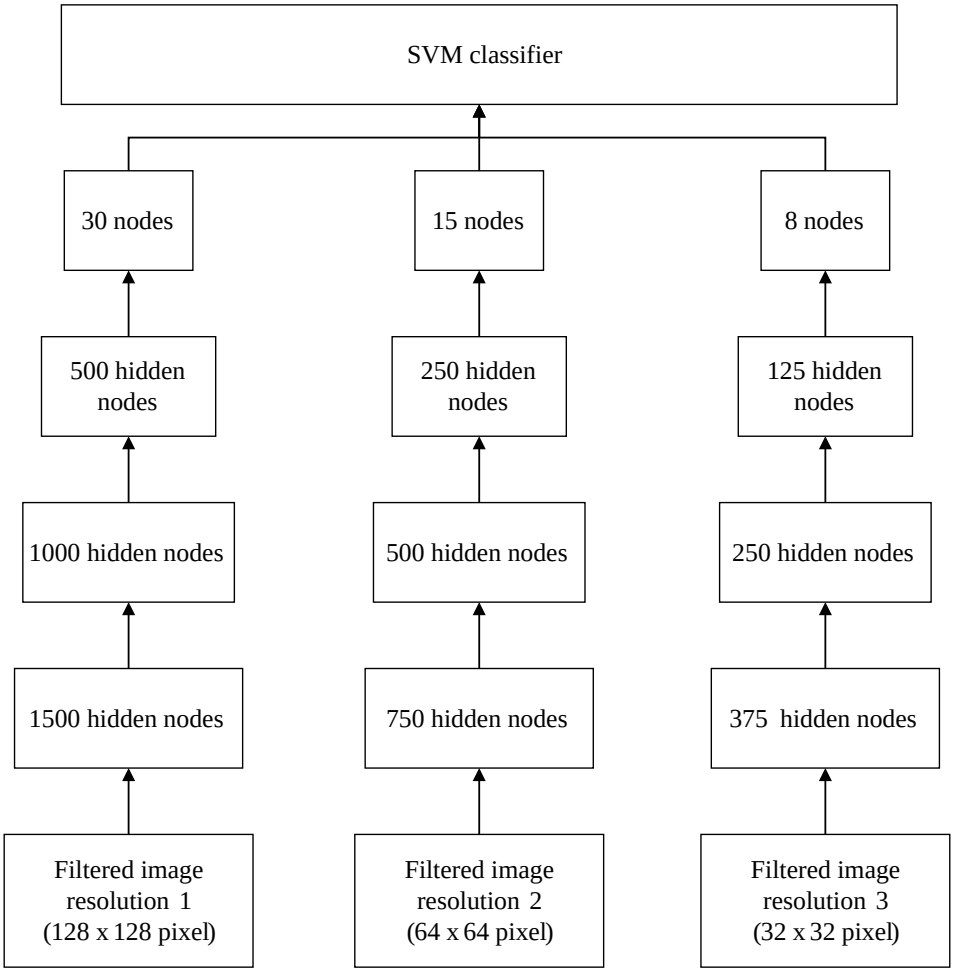


Figure 6.45: Multiple resolution and multiple orientation DBN with SVM classifier for microcalcification classification.

The summary of the process of the proposed classification system is as follow:

Input database The 118 images are filtered using set of steerable pyramid filters as in experiment in Section 6.2.2. However, instead of computing statistical measurements, ie. energy and entropy of the filtered images, the DBN is used to extract features from the filtered images.

Training input 80 of 118 images are randomly chosen with 40 benign and 40 malignant cases for training set. Since the images are filtered using steerable pyramid filtering, 24 filtered images are obtained from each image; ie. filtered at 3 levels of resolutions and 8 orientations.

Training phase The DBN pre-training will be carried out separately for each resolution and each orientation. This means that the pre-training process will use 24 different sets of input at the visible layer. The top layer of each process will then be combined before putting into the SVM for classification. The output of the final RBM will be the input features to the SVM.

Testing input The remaining 38 images will be the testing set used to validate the sensitivity and the specificity of the proposed system.

Testing phase The learned DBN and SVM network is validated with the 38 testing images. The process is repeated 10 times with different sets of training and testing and the statistical analysis is performed to validate the results.

The result is presented in a boxplot shown in Figure 6.46. The mean accuracy achieved is 60.79% and the median is 63.16% for the filtered images at all resolutions are combined at the input of DBN. This result is compared with the experiment using the raw image pixels as input of DBN as in experiments in Section 6.3.2. A t-test analysis is performed with p-value obtained is ≤ 0.05 which shows the difference in mean accuracies between both methods is significant. The t-test result is presented in Table 6.24.

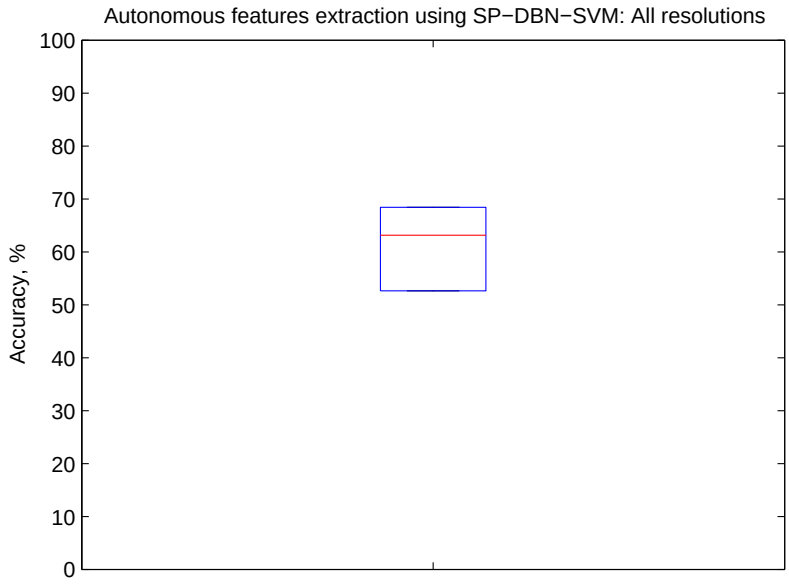


Figure 6.46: Boxplot analysis of accuracies achieved in a hybrid system of steerable pyramid filtering, DBN and SVM for microcalcification classification. The input images are the combination filtered images at resolution 1, 2 and 3.

Input image	N	Mean	SD	t	DF	p
Raw pixel images	10	47.91	17.2781	16.7345	19	7.9143e-13
Filtered images	10	60.79	7.3866			

Table 6.24: T-test analysis between different type of images at the input of the DBN features extractor.

Summary

The summary of the accuracy achieved by different method is presented in Figure 6.47. The graph shows that the feature extraction method (raw-DBN) as described in Section 6.3.2 has not performed well in automatically extracting features from raw images of microcalcification.

Therefore, in this experiment of Section 6.3.3, an extra process to produce multiple orientation and resolutions of the images is conducted. DBN is used to extract features from these filtered images and used them as input to the classifier. In Figure 6.46, the result has shown that the hybrid approach of multiple resolution and multiple orientation DBN (SP-DBN) for microcalcification diagnosis has contributed in improving

the performance of the classification. The experiments conducted so far are using low resolution images of 128 × 128 pixels due to the computational constraint. From the comparison graph as shown in Figure 6.47, it also shows that the result obtained by the method is comparable to the result obtained using other methods, as conducted in experiments in Section 6.2.4. This shows that DBN holds a promising result in extracting features for microcalcification classification. The accuracy may be further improved by using higher resolution images, suggested to be carried out in the future work related to this topic.

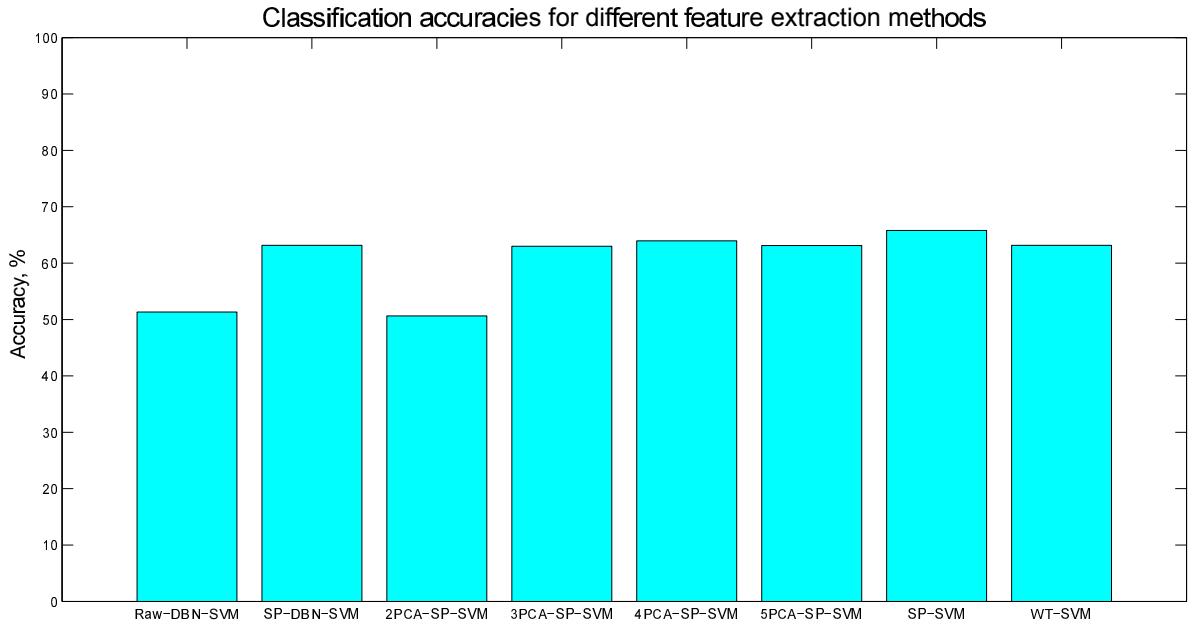


Figure 6.47: Comparison graph of mean accuracies achieved using different feature extraction approaches. The features used are either raw image pixels (raw), extracted from steerable pyramid filtering (SP) or extracted from wavelet transform (WT). The features are selected using principal component analysis (PCA) or automatically by DBN. These features are then classified using SVM classifier for malignancy analysis.

6.4 Summary

The first part of chapter summarised the potential application of the steerable pyramid filtering to extract more useful information features, specifically for microcalcification

diagnosis on mammograms. The features extracted were chosen to be the energy, entropy or the combination of both. However, with this approach, the features for the images were not optimally extracted. The successes of DBN in other applications, motivate us to investigate the ability of DBNs to automatically extract useful information from the raw images. Despite the promises of DBNs to have the ability in analysing the highly complex structure, the conducted experiments showed that DBNs are not sufficient for learning large size of data. Therefore, it is proposed to use DBN to extract features from filtered steerable pyramid images. With this guidance introduced at the input of DBN, it shows the use of hybrid method of steerable pyramid filtering and DBN possesses better potential for microcalcification classification on mammograms.

Chapter 7

Conclusion

THIS thesis addresses potential advantages of the feature extraction techniques for medical applications; mammograms, in which novel techniques for feature extraction for malignancy analysis of microcalcification clusters are proposed. In addition, this thesis promises to be useful by providing relevant supporting evidences of the work. Thus, the accuracy and the reliability of the techniques used to extract features are the main outcomes that were achieved.

Section 7.1 provides the summary of the thesis, along with the original findings and contributions of the thesis. In Section 7.2, few suggestions for improvement and further investigation will be provided for future explorations, and a concluding remark at the end of the chapter, in Section 7.3.

7.1 Summary of Findings

The practical application of CAD systems have been widely used to assist radiologists in making decisions; either for detection or for diagnosis of the abnormalities. Problems took place in interpreting mammogram images manually. Alternatively, the mammogram can be digitised so a computerised system can be used as an aid to radiologists. The advantages of using computerised systems to assist the radiologists to interpret mammograms have been extensively described in Chapter 1. However, computerised systems require the know-how skill to link the theory of computation with the practical interpretation of the mammogram. The advancement of this skill is new, hence this is the gap that this thesis seeks for solution.

Previous works in this field of study have been discussed and summarised in Chapter 2. The citations provide deeper and extensive background of the subject matter. The strong foundation of background knowledge enables a proposed method to be developed, which contribute solutions for the CAD system.

Many earlier researchers reported detection schemes of the CAD system for mammogram analysis which provided promising results. However, the result of mammogram analysis obtained from the diagnosis scheme of the CAD system is still somewhat below clinical requirements. There were suggestions proposed by earlier researchers to either improve the feature extraction or improve the technique for classification. These views informed the focus on the development of a novel method in CAD system for mammogram analysis, particularly the malignancy analysis of microcalcification clusters formed in breast tissues.

Previous studies also reported that multiple resolution analysis approaches for images provide useful information for classification of microcalcification in mammograms. The particular type of multi-resolution analysis that is offered by the wavelet transform has prompted us to use this technique for extracting features for the classification. Hence, the mathematical theory of the wavelet transforms is explored, starting from the motivation behind it, ie. Fourier analysis which has been introduced by a

mathematician, Joseph Fourier in 1800s, and the analysis is extensively described in Chapter 3.

Further exploration on the method that inherits the properties of the wavelet transform is performed. This method offers an additional property, which is able to analyse mammograms with multiple orientations. A method that offers the combination of multiple resolution and multiple orientation is the steerable pyramid filtering. The ability of steerable pyramid filtering technique to provide multiple of both resolutions and orientations has promoted interest to apply the technique to improve on the malignancy analysis of microcalcification clusters.

Hence, this has driven the thesis to proceed with the experiments using the steerable pyramid filtering technique in screening mammograms. Original experiments are conducted extensively to seek whether the features extracted from multiple orientation and multiple resolution analysis of microcalcification clusters are useful for classification.

The first part of Chapter 6 contains the detail results of these experiments. The experiments in the first part of Chapter 6 is divided into two main parts. The first experiment uses all features extracted from the steerable pyramid filtering as the input for classification. Whilst, the second experiment combines the feature extraction technique with a feature selection technique, the principal component analysis (PCA) method. The feature selection step is performed to reduce the dimension of the input vector to the classifier. A comparative study is made between the features extracted from the proposed technique, steerable pyramid with the commonly technique used for multiple resolution decomposition technique, wavelet transform. The results show that the extra features measured from multiple orientations obtained useful information for malignancy analysis. The comparison of the accuracy achieved between the two techniques is analysed by t-test, and is observed that the improvement is significant.

The effectiveness of the diagnosis system depends on the classifier used. Chapter 4 provides an overview of the commonly used classifiers in the area. Two types of most

commonly used classifiers are the support vector machine (SVM) and neural network (NN). The SVM has been used in many research problems due to its practicality in its implementation and they are able to adapt themselves to suit most of the problems of the practical interest. On the other hand, NN are intelligent machines that are able to map an input to an output through highly non-linear relationship. This makes NN to be one of the favored classification technique when dealing with input data that has complex structure, such as microcalcification cluster. However, due to the common practice of randomly initialising the weights, the trained networks have a tendency to be stuck at poor local minima and decreases the performance of the network, especially when the network is deeper than three layers.

Recent advances in deep network research has demonstrated an ability to extract high-level of representations of the input. This prompted the interest to further explore the use of deep networks in diagnosis of microcalcification clusters in mammograms. Recently new method named as deep belief network, DBN, has been proposed by Hinton *et al.*, that uses a greedy learning scheme to train the network. Chapter 5 of the thesis describes the details of the algorithm of the network, and demonstrates the use of DBN for MNIST handwritten digits classifications. The use of DBN to extract high level features has attracted many researchers to apply it in many fields of studies, such as audio classification (Ballan *et al.* 2009), medical application (Tamilselvan *et al.* 2012) and transportation (Huang *et al.* 2014).

The ability of greedy learning in deep network provides a highly non-linear mapping of the input and the output. Hence, this gives an advantage of DBN to analyse complex patterns; in this thesis DBN is applied to classify each microcalcification cluster into either benign or malignant case. In the second part of Chapter 6, extensive sets of original experiments were conducted where DBNs are used in extracting features for microcalcification classifications. The experiment using DBN solely as feature extractor and classifier of raw pixel microcalcification images showed no significant improvement. Therefore, a novel technique is proposed to use filtered images as the input to the DBN. The DBN will extract features from the filtered images instead of the raw pixels. The

result of analysis shows a significant improvement when the technique is applied, although the overall performance was still inferior to using steerable pyramids features. The initial findings in this thesis offer a platform to conduct further research on the use of DBNs for microcalcification clusters analysis.

7.2 Suggestions for Future Work

The research work reported in this thesis provides a basis for other emerging CADx system, specifically in feature extraction method for medical imaging applications. The findings presented in this thesis have demonstrated potential contributions in the field. However, much room remains for enhancing the capabilities of the CADx system in the proposed microcalcification diagnosis system.

Some of the suggested future work, which are in line with this thesis include:

1. Future work in the integration of the proposed texture features with other features, such as conventional shape features, edge features would be a promising task. A CADx system that can include as much informative features for malignancy analysis can be developed if more features to describe the microcalcification malignancy cases are extracted.

2. The work presented in this thesis has demonstrated the capability of DBN to extract useful microcalcification features for malignancy analysis. Despite the low resolution images used in the experiments, the results showed that DBN has potential future in this field. Therefore, the DBN is suggested for feature extraction from high resolution images, which may contribute in improving the classification rate of accuracy in the next investigation.

3. The optimal design of DBN architecture is another potential area of research in the future work. The number of hidden layers in the DBN contributes on the effectiveness of using the network for extracting useful features. It has demonstrated in other work that a deeper network is able to extract many features at

high level. In future work, deeper networks with more hidden layers can be considered for microcalcification malignancy analysis.

7.3 Concluding Remark

The motivation of thesis is to develop a robust feature extraction technique for the automatic classification for microcalcification clusters in mammograms. Steerable pyramid filtering technique combines the multiresolution property of wavelet transform with the crucial multi-orientation property, which makes it a useful technique for improving microcalcification malignancy analysis. This work also investigates a recently discovered technique in the field of artificial intelligence called deep belief network. Both approaches are evaluated through extensive computer experiments using a publicly available dataset. Knowledge contained in this thesis contribute to the greater pursuit of achieving rapid and reliable computer aided diagnosis of breast cancer.

Appendix A

Database

This chapter lists the name of the 118 images obtained from Digital Database of Screening Mammogram (DDSM) that were used as dataset in the experiments conducted in the thesis. For each case, there is also the descriptions of the abnormality type.

A.1 Benign Microcalcification Clusters

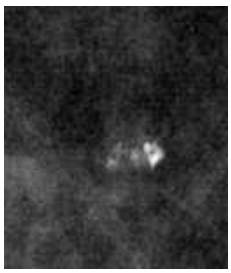
No.	Volume	Case	Left/Right	View	Abnormality type
1	benign_01	B_3104	Left	CC	Pleomorphic
2	benign_01	B_3104	Left	MLO	Pleomorphic
3	benign_01	B_3117	Right	CC	Pleomorphic
4	benign_01	B_3117	Right	MLO	Pleomorphic
5	benign_01	B_3127	Right	CC	Pleomorphic
6	benign_01	B_3127	Right	MLO	Pleomorphic
7	benign_01	B_3130	Right	CC	Pleomorphic
8	benign_01	B_3130	Right	MLO	Pleomorphic
9	benign_01	B_3148	Right	CC	Amorphous-Pleomorphic
10	benign_01	B_3148	Right	CC	Amorphous-Pleomorphic
11	benign_01	B_3148	Right	MLO	Amorphous-Pleomorphic
12	benign_01	B_3148	Right	MLO	Amorphous-Pleomorphic
13	benign_01	B_3159	Left	CC	Lucent-centered
14	benign_01	B_3159	Left	MLO	Lucent-centered
15	benign_01	B_3160	Left	CC	Lucent-centered
16	benign_01	B_3160	Left	MLO	Lucent-centered
17	benign_01	B_3162	Left	CC	Lucent-centered
18	benign_01	B_3162	Left	MLO	Lucent-centered
19	benign_01	B_3162	Right	CC	Round and regular lucent-centered
20	benign_01	B_3162	Right	CC	Round and regular lucent-centered
21	benign_01	B_3162	Right	MLO	Round and regular lucent-centered
22	benign_01	B_3162	Right	MLO	Round and regular lucent-centered
23	benign_01	B_3165	Left	CC	Lucent-centered
24	benign_01	B_3165	Left	CC	Lucent-centered
25	benign_01	B_3165	Left	MLO	Lucent-centered
26	benign_01	B_3165	Left	MLO	Lucent-centered
27	benign_01	B_3165	Right	MLO	Lucent-centered
28	benign_01	B_3166	Left	CC	Lucent-centered
29	benign_01	B_3166	Left	MLO	Lucent-centered
30	benign_01	B_3169	Right	CC	Lucent-centered
31	benign_01	B_3169	Right	MLO	Lucent-centered
32	benign_01	B_3175	Right	CC	Lucent-centered
33	benign_01	B_3175	Right	MLO	Lucent-centered
34	benign_01	B_3184	Right	CC	Lucent-centered
35	benign_01	B_3184	Right	MLO	Lucent-centered
36	benign_01	B_3185	Left	CC	Lucent-centered

No.	Volume	Case	Left/Right	View	Abnormality type
37	benign_01	B_3185	Left	MLO	Lucent-centered
38	benign_01	B_3186	Right	CC	Lucent-centered
39	benign_01	B_3186	Right	MLO	Lucent-centered
40	benign_02	A_1265	Right	CC	Pleomorphic
41	benign_02	A_1265	Right	MLO	Pleomorphic
42	benign_02	A_1268	Left	CC	Pleomorphic
43	benign_02	A_1268	Left	MLO	Pleomorphic
44	benign_02	A_1269	Right	CC	Pleomorphic
45	benign_02	A_1269	Right	MLO	Pleomorphic
46	benign_02	A_1270	Right	MLO	Pleomorphic
47	benign_02	A_1275	Left	CC	Pleomorphic
48	benign_02	A_1275	Left	MLO	Pleomorphic
49	benign_02	A_1278	Left	CC	Punctate
50	benign_02	A_1278	Left	MLO	Punctate
51	benign_02	A_1280	Left	CC	Pleomorphic
52	benign_02	A_1280	Left	MLO	Pleomorphic
53	benign_02	A_1281	Left	MLO	Pleomorphic
54	benign_02	A_1285	Left	CC	Pleomorphic
55	benign_02	A_1285	Left	MLO	Pleomorphic
56	benign_02	A_1310	Right	MLO	Pleomorphic
57	benign_02	A_1328	Right	CC	Pleomorphic
58	benign_02	A_1328	Right	MLO	Pleomorphic
59	benign_02	A_1331	Right	CC	Pleomorphic
60	benign_02	A_1331	Right	MLO	Pleomorphic
61	benign_02	A_1333	Right	CC	Pleomorphic
62	benign_02	A_1333	Right	MLO	Pleomorphic
63	benign_02	A_1349	Right	MLO	Pleomorphic
64	benign_03	A_1332	Left	CC	Pleomorphic
65	benign_03	A_1374	Right	CC	Pleomorphic
66	benign_03	A_1374	Right	MLO	Pleomorphic

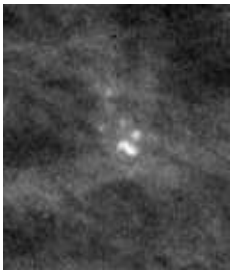
A.2 Malignant Microcalcification Clusters

No.	Volume	Case	Left/Right	View	Abnormality type
67	cancer_01	B_3005	Left	MLO	Pleomorphic
68	cancer_01	B_3005	Left	MLO	Pleomorphic
69	cancer_01	B_3005	Left	MLO	Punctate-Amorphous
70	cancer_01	B_3019	Left	CC	Pleomorphic
71	cancer_01	B_3019	Left	MLO	Pleomorphic
72	cancer_01	B_3025	Right	CC	Fine-linear branching
73	cancer_01	B_3025	Right	CC	Fine-linear branching
74	cancer_01	B_3025	Right	MLO	Fine-linear branching
75	cancer_01	B_3025	Right	MLO	Fine-linear branching
76	cancer_01	B_3025	Right	MLO	Fine-linear branching
77	cancer_01	B_3026	Right	CC	Pleomorphic
78	cancer_01	B_3026	Right	CC	Pleomorphic
79	cancer_01	B_3026	Right	CC	Pleomorphic
80	cancer_01	B_3026	Right	MLO	Pleomorphic
81	cancer_01	B_3026	Right	MLO	Pleomorphic
82	cancer_01	B_3026	Right	MLO	Pleomorphic
83	cancer_01	B_3037	Left	CC	Punctate-Pleomorphic
84	cancer_01	B_3037	Left	MLO	Punctate-Pleomorphic
85	cancer_01	B_3044	Left	MLO	Pleomorphic
86	cancer_01	B_3045	Left	CC	Fine-linear branching
87	cancer_01	B_3045	Left	CC	Fine-linear branching
88	cancer_01	B_3045	Left	CC	Fine-linear branching
89	cancer_01	B_3045	Left	MLO	Fine-linear branching
90	cancer_01	B_3045	Left	MLO	Fine-linear branching
91	cancer_01	B_3045	Left	MLO	Fine-linear branching
92	cancer_01	B_3049	Right	CC	Pleomorphic
93	cancer_01	B_3059	Right	CC	Fine-linear branching
94	cancer_01	B_3059	Right	MLO	Fine-linear branching
95	cancer_01	B_3059	Right	MLO	Fine-linear branching
96	cancer_01	B_3076	Left	CC	Amorphous round and regular
97	cancer_01	B_3076	Left	MLO	Amorphous round and regular
98	cancer_01	B_3076	Left	MLO	Amorphous round and regular
99	cancer_01	B_3079	Right	MLO	Pleomorphic
100	cancer_03	A_1000	Left	MLO	Pleomorphic
101	cancer_03	A_1003	Left	MLO	Pleomorphic
102	cancer_03	A_1005	Right	MLO	Punctate
103	cancer_03	A_1024	Right	MLO	Pleomorphic

No.	Volume	Case	Left/Right	View	Abnormality type
104	cancer_03	A_1031	Left	CC	Pleomorphic
105	cancer_03	A_1031	Left	MLO	Pleomorphic
106	cancer_03	A_1037	Right	CC	Pleomorphic
107	cancer_03	A_1037	Right	MLO	Pleomorphic
108	cancer_03	A_1041	Right	CC	Pleomorphic
109	cancer_03	A_1043	Left	CC	Amorphous
110	cancer_03	A_1043	Left	MLO	Amorphous
111	cancer_03	A_1047	Left	CC	Pleomorphic
112	cancer_03	A_1047	Left	MLO	Pleomorphic
113	cancer_03	A_1051	Right	MLO	Punctate
114	cancer_03	A_1052	Right	MLO	Fine-linear branching
115	cancer_03	A_1084	Right	CC	Pleomorphic
116	cancer_03	A_1084	Right	MLO	Pleomorphic
117	cancer_03	A_1085	Right	CC	Round and regular
118	cancer_03	A_1085	Right	MLO	Round and regular



(1)



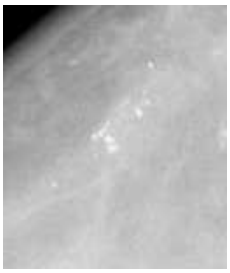
(2)

(3)

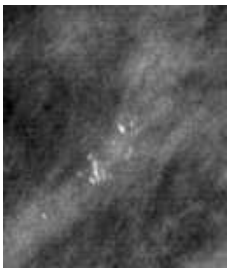
(4)

(5)

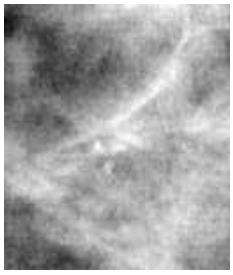
(6)



(7)

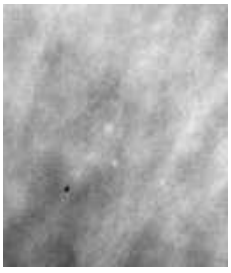


(8)

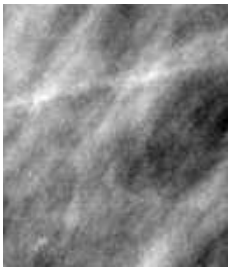


(9)

(10)

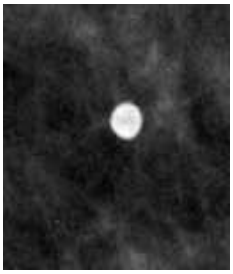


(11)

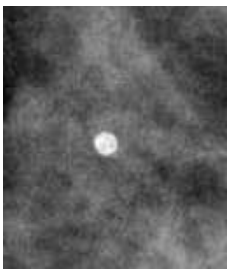


(12)

(13)



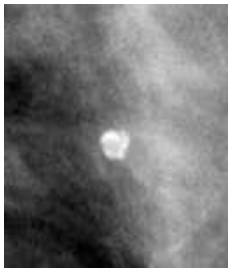
(14)



(15)

(16)

(17)

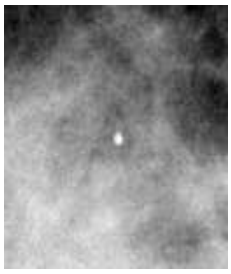


(18)

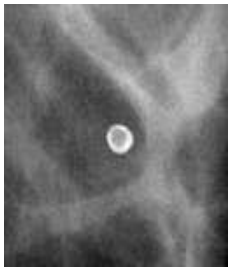
(19)

(20)

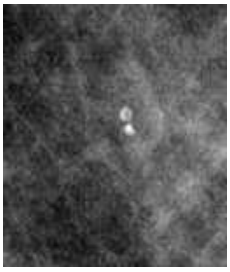
(21)



(22)

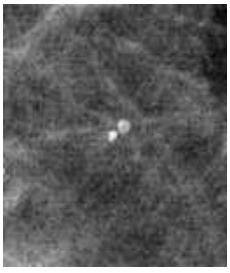


(23)



(24)

(25)

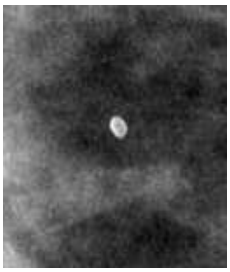


(26)

(27)

(28)

(29)



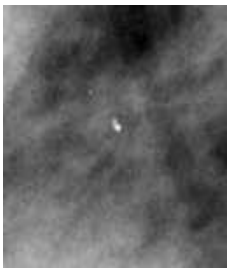
(30)

(31)

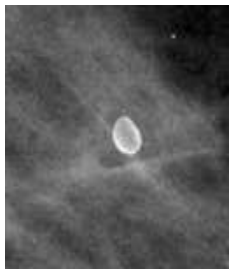
(32)

(33)

(34)

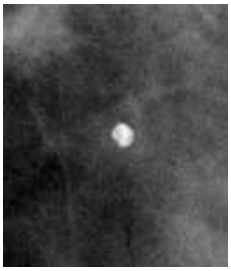


(35)



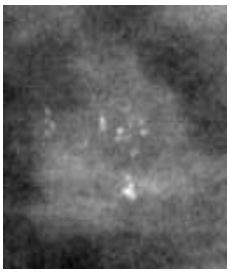
(36)

(37)

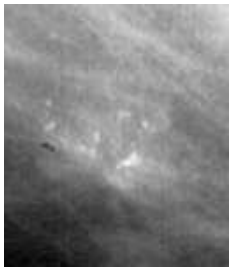


(38)

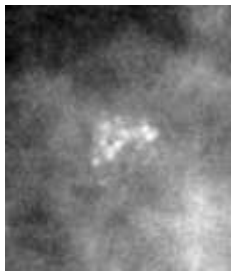
(39)



(40)



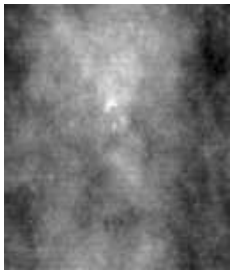
(41)



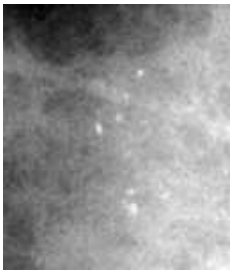
(42)

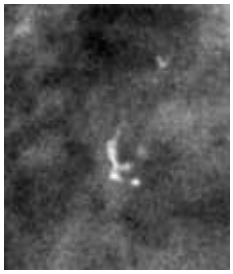
(43)

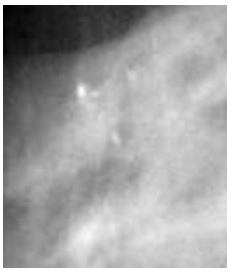
(44)

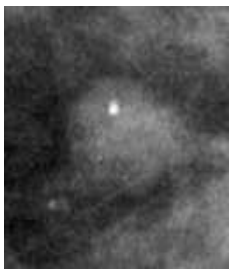


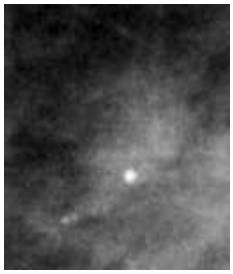
(45)

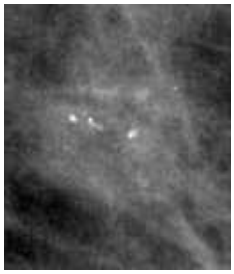

(46)


(47)

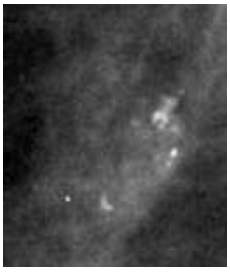

(48)

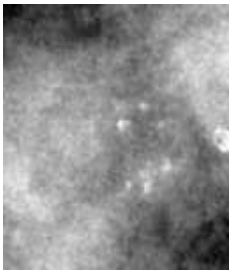

(49)


(50)

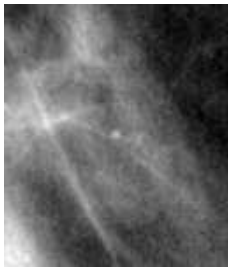

(51)


(52)

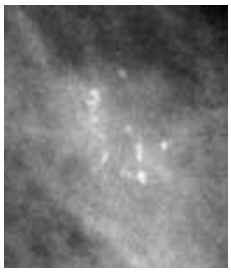

(53)

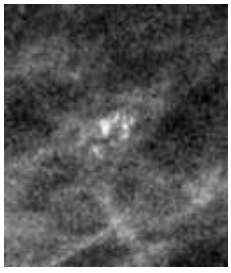

(54)

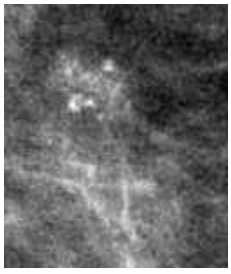

(55)

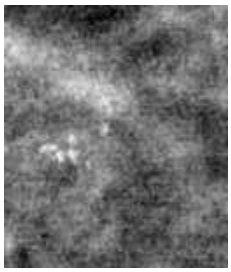

(56)


(57)

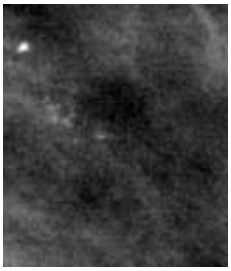

(58)


(59)

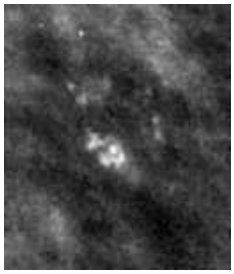

(60)



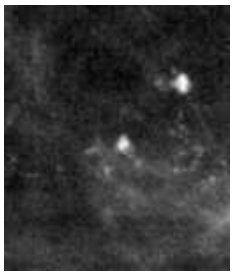
(61)



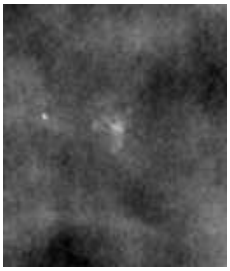
(62)



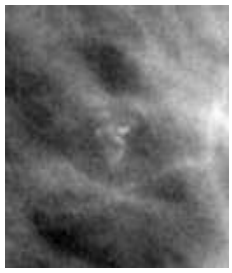
(63)



(64)



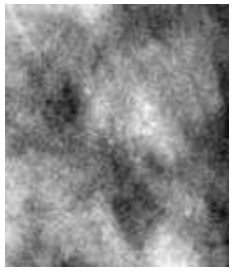
(65)



(66)

(67)

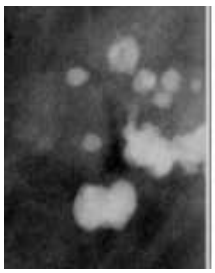
(68)



(69)

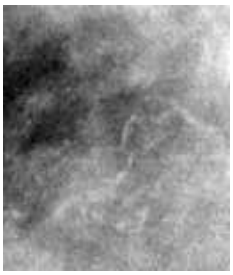
(70)

(71)



(72)

(73)

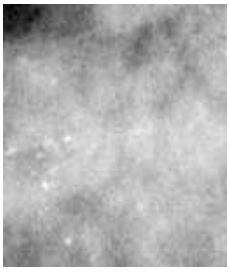


(74)

(75)

(76)

(77)



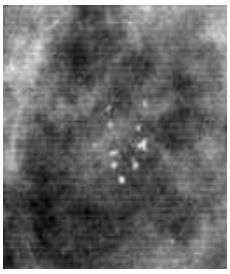
(78)

(79)

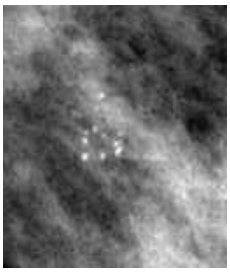
(80)

(81)

(82)

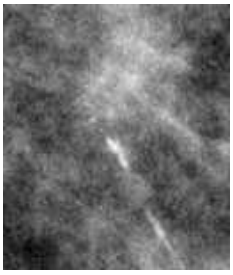


(83)

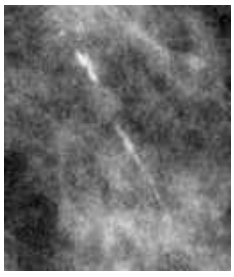


(84)

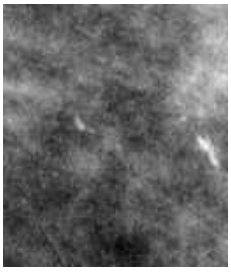
(85)



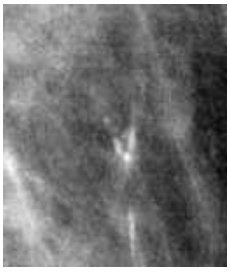
(86)



(87)

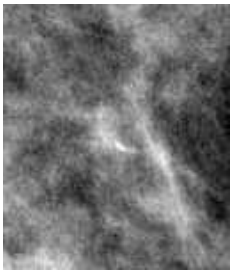


(88)

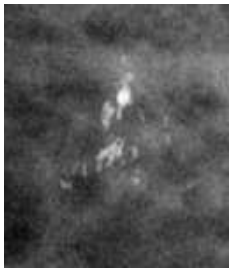


(89)

(90)



(91)



(92)

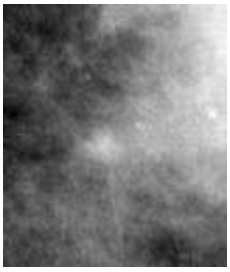
(93)

(94)

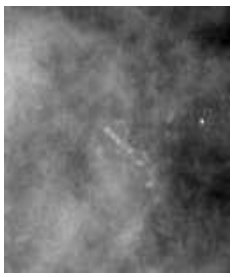
(95)

(96)

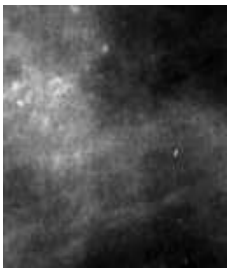
(97)



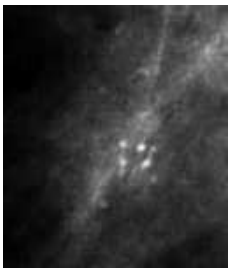
(98)



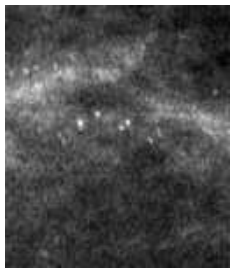
(99)



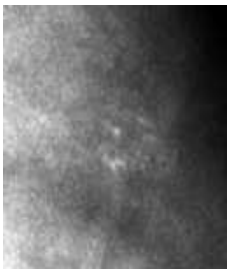
(100)



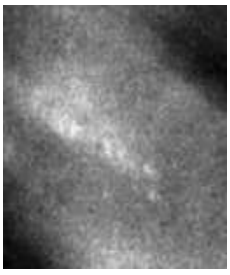
(101)



(102)

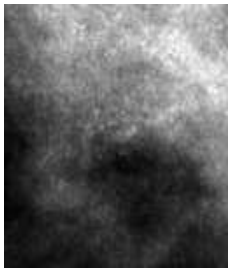


(103)

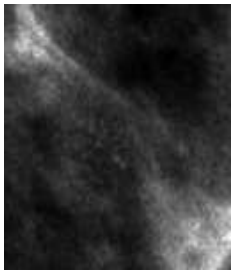


(104)

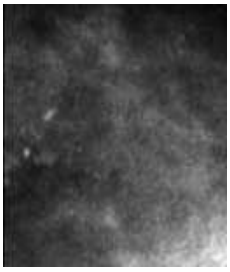
(105)



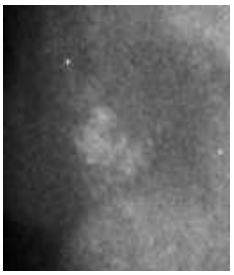
(106)



(107)

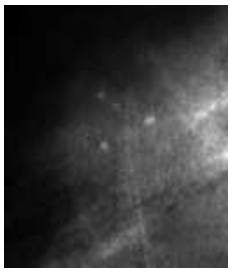


(108)

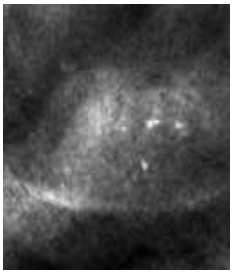


(109)

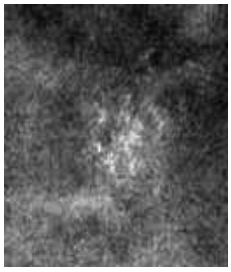
(110)



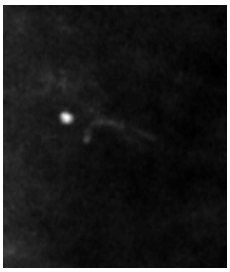
(111)



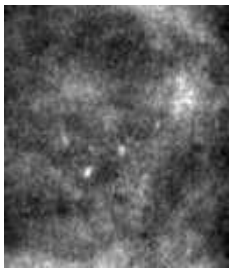
(112)



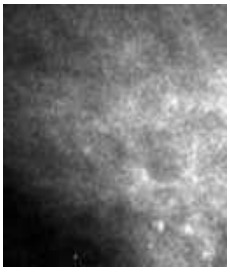
(113)



(114)

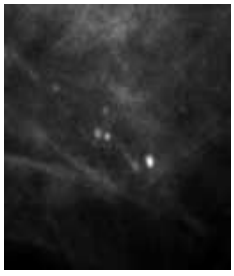


(115)



(116)

(117)



(118)

This page is blank.

Appendix B

Detailed DBN experiment results

This chapter provides the detailed results on pretraining error obtained from the experiments conducted in section 6.3.1.

B.1 Pre-Training Error in Layer One DBN

Hidden nodes	5000	2500	1500	200
Epoch	1.0e+ 06 *	1.0e+ 06 *	1.0e+ 06 *	1.0e+ 06 *
1	33.2945	16.4695	10.2234	2.4797
2	27.3891	14.9529	9.6610	2.4664
3	21.8551	13.1329	8.9351	2.4643
4	18.2450	11.5847	8.2341	2.4454
5	16.2534	10.3748	7.6277	2.4275
6	15.1743	9.4684	7.1222	2.4137
7	14.4243	8.5434	6.5342	2.3863
8	14.0324	7.7574	5.9174	2.3574
9	13.8739	7.1931	5.3621	2.3249
10	13.7639	6.9028	4.9115	2.2852
11	13.5149	6.8292	4.5803	2.2392
12	12.9828	6.8663	4.3719	2.1951
13	12.3205	6.9284	4.2846	2.1400
14	11.6111	6.9096	4.2618	2.0872
15	11.0566	6.7831	4.2703	2.0345
16	10.6908	6.5231	4.2883	1.9763
17	10.4975	6.1632	4.2739	1.9248
18	10.4195	5.7490	4.2283	1.8697
19	10.3238	5.3619	4.1149	1.8158
20	10.1166	5.0290	3.9672	1.7644
21	9.8478	4.8009	3.7802	1.7151
22	9.5434	4.6473	3.5765	1.6682
23	9.2448	4.5878	3.3738	1.6239
24	9.0079	4.5577	3.1876	1.5839
25	8.8350	4.5512	3.0274	1.5475
26	8.7433	4.5281	2.9076	1.5130
27	8.6736	4.4863	2.8137	1.4817
28	8.5730	4.4049	2.7490	1.4533
29	8.4234	4.3111	2.7125	1.4277
30	8.2715	4.1921	2.6889	1.4036
31	8.1119	4.0772	2.6815	1.3837
32	7.9418	3.9820	2.6658	1.3643
33	7.8283	3.9101	2.6431	1.3467
34	7.7341	3.8567	2.6246	1.3311
35	7.6439	3.8324	2.5912	1.3180
36	7.5579	3.8133	2.5502	1.3046
37	7.4861	3.7909	2.5076	1.2915
38	7.4173	3.7701	2.4670	1.2804
39	7.3018	3.7420	2.4251	1.2707
40	7.1991	3.7066	2.3868	1.2588
41	7.1285	3.6599	2.3514	1.2497
42	7.0295	3.6059	2.3218	1.2388
43	6.9840	3.5616	2.3031	1.2307
44	6.9430	3.5069	2.2866	1.2217
45	6.8503	3.4727	2.2724	1.2120
46	6.8286	3.4338	2.2627	1.2039
47	6.7371	3.3985	2.2457	1.1946
48	6.6944	3.3850	2.2360	1.1856
49	6.6210	3.3516	2.2252	1.1766
50	6.5677	3.3347	2.2075	1.1681
51	6.5182	3.3103	2.1940	1.1591
52	6.5004	3.2865	2.1762	1.1506
53	6.4187	3.2429	2.1587	1.1416
54	6.4008	3.2205	2.1368	1.1338
55	6.3546	3.1859	2.1193	1.1251
56	6.3068	3.1553	2.1028	1.1173
57	6.2643	3.1223	2.0828	1.1100
58	6.2167	3.0987	2.0720	1.1018
59	6.1950	3.0737	2.0572	1.0940
60	6.1546	3.0617	2.0406	1.0868
61	6.1246	3.0273	2.0293	1.0796

Hidden nodes	5000	2500	1500	200
Epoch	1.0e+ 06 *	1.0e+ 06 *	1.0e+ 06 *	1.0e+ 06 *
62	6.1042	3.0139	2.0171	1.0726
63	6.0667	2.9819	2.0047	1.0657
64	6.0304	2.9623	1.9948	1.0586
65	6.0081	2.9381	1.9788	1.0521
66	5.9889	2.9166	1.9696	1.0458
67	5.9418	2.8896	1.9574	1.0390
68	5.9290	2.8761	1.9443	1.0331
69	5.9125	2.8541	1.9271	1.0271
70	5.8816	2.8407	1.9140	1.0218
71	5.8846	2.8173	1.9004	1.0162
72	5.8458	2.7980	1.8886	1.0104
73	5.8157	2.7828	1.8726	1.0059
74	5.8126	2.7578	1.8637	1.0004
75	5.7782	2.7490	1.8535	0.9957
76	5.7531	2.7277	1.8420	0.9908
77	5.7576	2.7069	1.8262	0.9859
78	5.7513	2.6933	1.8145	0.9815
79	5.7334	2.6732	1.8079	0.9769
80	5.7182	2.6572	1.7939	0.9725
81	5.6973	2.6478	1.7830	0.9685
82	5.6842	2.6332	1.7715	0.9641
83	5.6635	2.6198	1.7612	0.9602
84	5.6761	2.6119	1.7488	0.9559
85	5.6740	2.5990	1.7419	0.9527
86	5.6511	2.5761	1.7283	0.9487
87	5.6350	2.5633	1.7216	0.9447
88	5.6311	2.5498	1.7079	0.9415
89	5.6175	2.5341	1.7047	0.9386
90	5.6099	2.5216	1.6900	0.9349
91	5.6073	2.5061	1.6830	0.9312
92	5.5989	2.5011	1.6772	0.9282
93	5.5880	2.4900	1.6641	0.9251
94	5.5980	2.4787	1.6560	0.9220
95	5.5886	2.4676	1.6464	0.9197
96	5.5701	2.4538	1.6411	0.9159
97	5.5736	2.4388	1.6281	0.9137
98	5.5718	2.4320	1.6228	0.9106
99	5.5541	2.4215	1.6141	0.9078
100	5.5594	2.4079	1.6046	0.9053
101	5.5402	2.4074	1.5950	0.9028
102	5.5414	2.3918	1.5899	0.9003
103	5.5477	2.3834	1.5785	0.8972
104	5.5475	2.3686	1.5727	0.8952
105	5.5361	2.3589	1.5609	0.8928
106	5.5410	2.3519	1.5566	0.8903
107	5.5420	2.3463	1.5466	0.8885
108	5.5237	2.3444	1.5395	0.8859
109	5.5346	2.3280	1.5330	0.8836
110	5.5319	2.3227	1.5204	0.8811
111	5.5266	2.3056	1.5180	0.8794
112	5.5095	2.3008	1.5111	0.8775
113	5.5202	2.2962	1.5017	0.8756
114	5.5152	2.2825	1.4990	0.8737
115	5.5136	2.2765	1.4887	0.8715
116	5.5213	2.2770	1.4767	0.8699
117	5.5158	2.2669	1.4741	0.8677
118	5.5159	2.2588	1.4675	0.8657
119	5.5263	2.2544	1.4592	0.8644
120	5.5008	2.2365	1.4527	0.8624
121	5.5208	2.2316	1.4443	0.8607
122	5.5216	2.2296	1.4414	0.8590

Hidden nodes	5000	2500	1500	200
Epoch	1.0e+ 06 *	1.0e+ 06 *	1.0e+ 06 *	1.0e+ 06 *
123	5.5076	2.2275	1.4304	0.8576
124	5.5182	2.2125	1.4272	0.8554
125	5.5164	2.2135	1.4236	0.8538
126	5.5088	2.1997	1.4139	0.8522
127	5.5061	2.2002	1.4108	0.8510
128	5.5063	2.1905	1.4021	0.8493
129	5.5029	2.1850	1.3939	0.8478
130	5.5079	2.1788	1.3867	0.8461
131	5.4980	2.1692	1.3816	0.8446
132	5.4938	2.1680	1.3770	0.8433
133	5.4958	2.1712	1.3693	0.8420
134	5.4962	2.1534	1.3675	0.8408
135	5.4852	2.1459	1.3572	0.8390
136	5.4815	2.1413	1.3529	0.8377
137	5.4957	2.1336	1.3477	0.8366
138	5.4719	2.1335	1.3414	0.8353
139	5.4733	2.1250	1.3357	0.8344
140	5.4836	2.1241	1.3335	0.8327
141	5.4616	2.1218	1.3261	0.8317
142	5.4651	2.1171	1.3215	0.8304
143	5.4893	2.1137	1.3127	0.8290
144	5.4640	2.1019	1.3084	0.8280
145	5.4635	2.1011	1.3037	0.8267
146	5.4757	2.0956	1.2996	0.8256
147	5.4596	2.0884	1.2914	0.8244
148	5.4692	2.0830	1.2840	0.8234
149	5.4767	2.0903	1.2809	0.8222
150	5.4626	2.0691	1.2810	0.8209
151	5.4601	2.0764	1.2718	0.8200
152	5.4523	2.0706	1.2688	0.8189
153	5.4557	2.0680	1.2616	0.8177
154	5.4584	2.0576	1.2541	0.8165
155	5.4419	2.0579	1.2492	0.8159
156	5.4521	2.0547	1.2490	0.8149
157	5.4401	2.0539	1.2426	0.8136
158	5.4530	2.0536	1.2400	0.8130
159	5.4359	2.0539	1.2299	0.8119
160	5.4495	2.0463	1.2277	0.8107
161	5.4374	2.0447	1.2216	0.8098

Hidden nodes	5000	2500	1500	200
Epoch	1.0e+ 06 *	1.0e+ 06 *	1.0e+ 06 *	1.0e+ 06 *
162	5.4380	2.0333	1.2195	0.8089
163	5.4498	2.0286	1.2131	0.8080
164	5.4302	2.0410	1.2116	0.8072
165	5.4256	2.0276	1.2054	0.8061
166	5.4300	2.0252	1.1975	0.8052
167	5.4243	2.0164	1.1953	0.8044
168	5.4371	2.0161	1.1895	0.8036
169	5.4243	2.0152	1.1833	0.8025
170	5.4236	2.0152	1.1814	0.8014
171	5.4225	2.0072	1.1781	0.8007
172	5.4148	2.0049	1.1692	0.7997
173	5.4143	2.0015	1.1695	0.7989
174	5.4237	1.9981	1.1658	0.7981
175	5.4189	1.9995	1.1589	0.7973
176	5.4129	2.0054	1.1597	0.7962
177	5.4013	1.9874	1.1524	0.7957
178	5.4103	1.9947	1.1472	0.7949
179	5.4075	1.9946	1.1431	0.7940
180	5.4160	1.9902	1.1395	0.7930
181	5.4040	1.9868	1.1347	0.7925
182	5.4094	1.9855	1.1300	0.7915
183	5.4020	1.9848	1.1284	0.7910
184	5.3968	1.9782	1.1214	0.7902
185	5.3880	1.9788	1.1170	0.7893
186	5.3945	1.9740	1.1151	0.7884
187	5.3910	1.9738	1.1104	0.7875
188	5.4041	1.9791	1.1075	0.7869
189	5.3787	1.9744	1.1037	0.7863
190	5.3680	1.9709	1.1040	0.7852
191	5.3713	1.9684	1.0961	0.7847
192	5.3678	1.9664	1.0939	0.7838
193	5.3840	1.9648	1.0902	0.7831
194	5.3749	1.9689	1.0840	0.7824
195	5.3867	1.9565	1.0851	0.7818
196	5.3752	1.9585	1.0787	0.7808
197	5.3722	1.9605	1.0704	0.7804
198	5.3665	1.9560	1.0720	0.7797
199	5.3599	1.9643	1.0658	0.7790
200	5.3673	1.9564	1.0610	0.7781

B.2 Pre-Training Error in Layer Two DBN

Hidden nodes	1500	1000	500	200
Epoch	1.0e+ 04 *	1.0e+ 04 *	1.0e+ 04 *	1.0e+ 04 *
1	3.3010	3.3032	3.3409	3.1520
2	3.2699	3.1483	2.8245	2.7274
3	2.8699	2.5581	2.4789	2.4846
4	2.2008	2.2051	2.2343	2.2800
5	1.9478	1.9324	2.0309	2.1093
6	1.7918	1.7804	1.8544	1.9684
7	1.6564	1.6767	1.7259	1.8187
8	1.5906	1.6287	1.6511	1.7055
9	1.5346	1.5839	1.6060	1.6349
10	1.4738	1.5392	1.5752	1.5877
11	1.3970	1.4742	1.5325	1.5493
12	1.3147	1.4089	1.4827	1.5134
13	1.2332	1.3409	1.4205	1.4720
14	1.1660	1.2703	1.3545	1.4261
15	1.0956	1.2085	1.2916	1.3717
16	1.0352	1.1411	1.2275	1.3164
17	0.9792	1.0830	1.1698	1.2577
18	0.9227	1.0276	1.1128	1.2075
19	0.8735	0.9740	1.0681	1.1618
20	0.8280	0.9266	1.0181	1.1179
21	0.7888	0.8791	0.9704	1.0793
22	0.7395	0.8398	0.9323	1.0435
23	0.7003	0.7980	0.8939	1.0039
24	0.6627	0.7599	0.8604	0.9707
25	0.6276	0.7248	0.8264	0.9396
26	0.5899	0.6921	0.7969	0.9084
27	0.5549	0.6548	0.7619	0.8816
28	0.5219	0.6235	0.7311	0.8575
29	0.4916	0.5947	0.7006	0.8325
30	0.4622	0.5616	0.6723	0.8041
31	0.4339	0.5323	0.6418	0.7793
32	0.4066	0.5045	0.6179	0.7558
33	0.3797	0.4766	0.5908	0.7332
34	0.3573	0.4518	0.5635	0.7111
35	0.3335	0.4268	0.5386	0.6866
36	0.3122	0.4008	0.5168	0.6656
37	0.2925	0.3774	0.4953	0.6430
38	0.2742	0.3571	0.4734	0.6229
39	0.2568	0.3383	0.4537	0.6028
40	0.2413	0.3196	0.4330	0.5841
41	0.2272	0.3010	0.4133	0.5637
42	0.2117	0.2843	0.3958	0.5446
43	0.1993	0.2681	0.3768	0.5303
44	0.1869	0.2535	0.3603	0.5145
45	0.1742	0.2403	0.3455	0.4973
46	0.1648	0.2267	0.3317	0.4805
47	0.1541	0.2139	0.3168	0.4657
48	0.1446	0.2033	0.3019	0.4530
49	0.1363	0.1940	0.2893	0.4387
50	0.1282	0.1827	0.2768	0.4248
51	0.1219	0.1739	0.2650	0.4126
52	0.1158	0.1638	0.2535	0.4002
53	0.1089	0.1566	0.2423	0.3874
54	0.1030	0.1492	0.2316	0.3747
55	0.0970	0.1419	0.2220	0.3648
56	0.0920	0.1349	0.2123	0.3543
57	0.0875	0.1290	0.2032	0.3435
58	0.0829	0.1221	0.1951	0.3345
59	0.0783	0.1162	0.1869	0.3252
60	0.0751	0.1115	0.1798	0.3150
61	0.0709	0.1073	0.1721	0.3056

Hidden nodes	1500	1000	500	200
Epoch	1.0e+ 04 *	1.0e+ 04 *	1.0e+ 04 *	1.0e+ 04 *
62	0.0679	0.1023	0.1667	0.2975
63	0.0648	0.0976	0.1594	0.2891
64	0.0616	0.0929	0.1536	0.2816
65	0.0593	0.0890	0.1472	0.2732
66	0.0566	0.0862	0.1414	0.2659
67	0.0541	0.0828	0.1364	0.2591
68	0.0522	0.0790	0.1323	0.2511
69	0.0501	0.0756	0.1271	0.2446
70	0.0477	0.0727	0.1229	0.2380
71	0.0461	0.0702	0.1178	0.2322
72	0.0444	0.0676	0.1144	0.2256
73	0.0428	0.0649	0.1104	0.2201
74	0.0410	0.0626	0.1057	0.2150
75	0.0398	0.0603	0.1035	0.2102
76	0.0387	0.0583	0.0996	0.2042
77	0.0365	0.0565	0.0952	0.1995
78	0.0352	0.0547	0.0927	0.1948
79	0.0341	0.0524	0.0893	0.1902
80	0.0332	0.0502	0.0864	0.1862
81	0.0320	0.0490	0.0840	0.1816
82	0.0307	0.0477	0.0815	0.1776
83	0.0298	0.0460	0.0787	0.1729
84	0.0286	0.0445	0.0764	0.1695
85	0.0281	0.0436	0.0736	0.1660
86	0.0269	0.0421	0.0715	0.1619
87	0.0261	0.0412	0.0699	0.1585
88	0.0254	0.0396	0.0674	0.1551
89	0.0246	0.0385	0.0658	0.1519
90	0.0242	0.0372	0.0636	0.1486
91	0.0236	0.0362	0.0618	0.1451
92	0.0229	0.0353	0.0602	0.1428
93	0.0224	0.0343	0.0587	0.1396
94	0.0216	0.0339	0.0570	0.1369
95	0.0209	0.0324	0.0556	0.1341
96	0.0202	0.0317	0.0542	0.1316
97	0.0197	0.0305	0.0522	0.1290
98	0.0195	0.0301	0.0515	0.1260
99	0.0187	0.0291	0.0496	0.1233
100	0.0183	0.0284	0.0484	0.1213
101	0.0177	0.0278	0.0475	0.1191
102	0.0173	0.0266	0.0460	0.1180
103	0.0171	0.0261	0.0449	0.1156
104	0.0165	0.0255	0.0438	0.1129
105	0.0162	0.0247	0.0433	0.1106
106	0.0155	0.0241	0.0420	0.1089
107	0.0154	0.0236	0.0416	0.1070
108	0.0151	0.0236	0.0405	0.1045
109	0.0146	0.0226	0.0401	0.1027
110	0.0145	0.0222	0.0391	0.1022
111	0.0140	0.0218	0.0380	0.0996
112	0.0138	0.0209	0.0369	0.0983
113	0.0134	0.0206	0.0363	0.0961
114	0.0130	0.0201	0.0353	0.0946
115	0.0130	0.0196	0.0348	0.0933
116	0.0125	0.0190	0.0340	0.0914
117	0.0122	0.0186	0.0330	0.0894
118	0.0124	0.0184	0.0325	0.0887
119	0.0118	0.0179	0.0318	0.0871
120	0.0115	0.0174	0.0315	0.0864
121	0.0114	0.0172	0.0307	0.0848
122	0.0113	0.0169	0.0299	0.0831

Hidden nodes	1500	1000	500	200
Epoch	1.0e+ 04 *	1.0e+ 04 *	1.0e+ 04 *	1.0e+ 04 *
123	0.0110	0.0167	0.0298	0.0818
124	0.0109	0.0161	0.0289	0.0810
125	0.0105	0.0158	0.0281	0.0799
126	0.0103	0.0154	0.0277	0.0784
127	0.0102	0.0154	0.0275	0.0775
128	0.0101	0.0147	0.0266	0.0754
129	0.0098	0.0146	0.0264	0.0749
130	0.0097	0.0142	0.0262	0.0736
131	0.0095	0.0140	0.0253	0.0720
132	0.0093	0.0136	0.0248	0.0708
133	0.0091	0.0135	0.0244	0.0696
134	0.0089	0.0134	0.0240	0.0700
135	0.0089	0.0132	0.0237	0.0673
136	0.0087	0.0129	0.0232	0.0669
137	0.0086	0.0129	0.0227	0.0657
138	0.0086	0.0122	0.0221	0.0648
139	0.0082	0.0122	0.0221	0.0642
140	0.0081	0.0120	0.0217	0.0627
141	0.0080	0.0117	0.0211	0.0623
142	0.0079	0.0115	0.0208	0.0617
143	0.0078	0.0115	0.0206	0.0605
144	0.0078	0.0113	0.0200	0.0613
145	0.0076	0.0111	0.0197	0.0590
146	0.0078	0.0108	0.0196	0.0585
147	0.0072	0.0107	0.0192	0.0573
148	0.0074	0.0105	0.0191	0.0563
149	0.0071	0.0102	0.0186	0.0562
150	0.0069	0.0100	0.0183	0.0553
151	0.0070	0.0098	0.0179	0.0547
152	0.0067	0.0096	0.0176	0.0540
153	0.0066	0.0096	0.0176	0.0536
154	0.0065	0.0095	0.0173	0.0529
155	0.0064	0.0092	0.0169	0.0518
156	0.0064	0.0093	0.0166	0.0520
157	0.0064	0.0091	0.0167	0.0513
158	0.0065	0.0089	0.0162	0.0507
159	0.0062	0.0088	0.0160	0.0505
160	0.0061	0.0088	0.0157	0.0493
161	0.0059	0.0088	0.0155	0.0483

Hidden nodes	1500	1000	500	200
Epoch	1.0e+ 04 *	1.0e+ 04 *	1.0e+ 04 *	1.0e+ 04 *
162	0.0059	0.0084	0.0153	0.0483
163	0.0057	0.0084	0.0149	0.0472
164	0.0057	0.0082	0.0150	0.0478
165	0.0057	0.0081	0.0147	0.0462
166	0.0056	0.0081	0.0150	0.0458
167	0.0055	0.0077	0.0146	0.0457
168	0.0054	0.0077	0.0143	0.0451
169	0.0055	0.0078	0.0140	0.0444
170	0.0052	0.0075	0.0140	0.0436
171	0.0052	0.0075	0.0134	0.0434
172	0.0053	0.0073	0.0133	0.0428
173	0.0051	0.0073	0.0132	0.0422
174	0.0050	0.0072	0.0131	0.0422
175	0.0049	0.0072	0.0128	0.0414
176	0.0049	0.0071	0.0126	0.0414
177	0.0049	0.0071	0.0125	0.0403
178	0.0047	0.0071	0.0123	0.0398
179	0.0048	0.0069	0.0121	0.0387
180	0.0046	0.0066	0.0122	0.0384
181	0.0046	0.0067	0.0120	0.0381
182	0.0045	0.0065	0.0117	0.0378
183	0.0045	0.0067	0.0116	0.0375
184	0.0043	0.0065	0.0116	0.0366
185	0.0045	0.0064	0.0116	0.0361
186	0.0043	0.0063	0.0113	0.0359
187	0.0042	0.0062	0.0116	0.0356
188	0.0044	0.0060	0.0110	0.0349
189	0.0044	0.0061	0.0111	0.0346
190	0.0041	0.0060	0.0107	0.0342
191	0.0040	0.0058	0.0109	0.0338
192	0.0041	0.0058	0.0106	0.0334
193	0.0040	0.0057	0.0105	0.0330
194	0.0040	0.0056	0.0103	0.0326
195	0.0039	0.0055	0.0103	0.0329
196	0.0039	0.0054	0.0100	0.0322
197	0.0038	0.0055	0.0100	0.0322
198	0.0038	0.0054	0.0102	0.0312
199	0.0039	0.0053	0.0099	0.0313
200	0.0037	0.0053	0.0098	0.0310

B.3 Pre-Training Error in Layer Three DBN

Hidden nodes	1000	500	200
Epoch	1.0e+ 04 *	1.0e+ 04 *	1.0e+ 04 *
1	2.6752	2.5138	2.3492
2	0.3870	0.3797	0.4488
3	0.3707	0.3199	0.3550
4	0.2886	0.2871	0.3027
5	0.2200	0.2364	0.2696
6	0.1875	0.2111	0.2372
7	0.1800	0.1934	0.2140
8	0.1558	0.1781	0.1958
9	0.1472	0.1630	0.1826
10	0.1303	0.1500	0.1687
11	0.1252	0.1437	0.1565
12	0.1162	0.1359	0.1508
13	0.1090	0.1272	0.1465
14	0.0983	0.1166	0.1403
15	0.0883	0.1049	0.1372
16	0.0840	0.1021	0.1286
17	0.0747	0.0944	0.1209
18	0.0708	0.0893	0.1138
19	0.0659	0.0821	0.1078
20	0.0642	0.0805	0.1039
21	0.0589	0.0717	0.0970
22	0.0532	0.0709	0.0898
23	0.0503	0.0655	0.0868
24	0.0457	0.0601	0.0836
25	0.0424	0.0573	0.0805
26	0.0406	0.0551	0.0755
27	0.0380	0.0510	0.0721
28	0.0337	0.0478	0.0695
29	0.0319	0.0438	0.0657
30	0.0310	0.0433	0.0635
31	0.0288	0.0405	0.0599
32	0.0258	0.0370	0.0570
33	0.0239	0.0357	0.0556
34	0.0228	0.0332	0.0537
35	0.0211	0.0289	0.0504
36	0.0195	0.0313	0.0485
37	0.0175	0.0270	0.0470
38	0.0168	0.0253	0.0456
39	0.0173	0.0248	0.0435
40	0.0147	0.0224	0.0417
41	0.0142	0.0221	0.0393
42	0.0133	0.0207	0.0371
43	0.0118	0.0181	0.0362
44	0.0115	0.0172	0.0346
45	0.0103	0.0170	0.0330
46	0.0104	0.0164	0.0313
47	0.0108	0.0146	0.0303
48	0.0099	0.0139	0.0284
49	0.0084	0.0122	0.0280
50	0.0082	0.0132	0.0271
51	0.0079	0.0116	0.0245
52	0.0068	0.0101	0.0241

Hidden nodes	1000	500	200
Epoch	1.0e+ 04 *	1.0e+ 04 *	1.0e+ 04 *
53	0.0073	0.0106	0.0227
54	0.0069	0.0110	0.0221
55	0.0059	0.0091	0.0203
56	0.0058	0.0079	0.0200
57	0.0063	0.0080	0.0183
58	0.0050	0.0079	0.0177
59	0.0051	0.0069	0.0181
60	0.0049	0.0069	0.0172
61	0.0046	0.0067	0.0159
62	0.0049	0.0060	0.0155
63	0.0040	0.0059	0.0150
64	0.0042	0.0049	0.0142
65	0.0036	0.0051	0.0127
66	0.0034	0.0044	0.0132
67	0.0033	0.0049	0.0130
68	0.0030	0.0047	0.0119
69	0.0034	0.0050	0.0115
70	0.0029	0.0042	0.0118
71	0.0032	0.0042	0.0104
72	0.0029	0.0043	0.0107
73	0.0036	0.0036	0.0105
74	0.0027	0.0031	0.0097
75	0.0025	0.0035	0.0092
76	0.0024	0.0031	0.0086
77	0.0027	0.0032	0.0087
78	0.0028	0.0028	0.0081
79	0.0027	0.0028	0.0081
80	0.0024	0.0029	0.0076
81	0.0016	0.0026	0.0074
82	0.0023	0.0027	0.0074
83	0.0024	0.0026	0.0072
84	0.0018	0.0029	0.0071
85	0.0023	0.0026	0.0066
86	0.0019	0.0019	0.0069
87	0.0017	0.0025	0.0057
88	0.0018	0.0021	0.0056
89	0.0019	0.0021	0.0051
90	0.0017	0.0021	0.0052
91	0.0019	0.0019	0.0051
92	0.0018	0.0020	0.0057
93	0.0016	0.0021	0.0047
94	0.0019	0.0018	0.0043
95	0.0014	0.0016	0.0049
96	0.0016	0.0016	0.0049
97	0.0016	0.0015	0.0045
98	0.0012	0.0018	0.0043
99	0.0014	0.0016	0.0044
100	0.0017	0.0017	0.0040
101	0.0012	0.0015	0.0038
102	0.0015	0.0014	0.0040
103	0.0015	0.0017	0.0038
104	0.0012	0.0013	0.0034

Hidden nodes	1000	500	200
Epoch	1.0e+ 04 *	1.0e+ 04 *	1.0e+ 04 *
105	0.0013	0.0012	0.0035
106	0.0015	0.0014	0.0034
107	0.0012	0.0011	0.0039
108	0.0013	0.0015	0.0035
109	0.0012	0.0010	0.0034
110	0.0011	0.0012	0.0035
111	0.0012	0.0012	0.0029
112	0.0014	0.0013	0.0032
113	0.0015	0.0013	0.0027
114	0.0009	0.0012	0.0031
115	0.0010	0.0012	0.0026
116	0.0012	0.0009	0.0024
117	0.0009	0.0012	0.0028
118	0.0011	0.0011	0.0027
119	0.0009	0.0010	0.0027
120	0.0011	0.0012	0.0028
121	0.0010	0.0009	0.0025
122	0.0011	0.0011	0.0026
123	0.0011	0.0010	0.0025
124	0.0009	0.0010	0.0022
125	0.0011	0.0011	0.0022
126	0.0010	0.0008	0.0021
127	0.0008	0.0009	0.0026
128	0.0010	0.0009	0.0021
129	0.0011	0.0011	0.0021
130	0.0009	0.0009	0.0022
131	0.0008	0.0009	0.0017
132	0.0008	0.0008	0.0020
133	0.0011	0.0009	0.0022
134	0.0009	0.0009	0.0021
135	0.0010	0.0008	0.0018
136	0.0011	0.0007	0.0020
137	0.0010	0.0009	0.0019
138	0.0008	0.0008	0.0019
139	0.0010	0.0007	0.0020
140	0.0010	0.0007	0.0015
141	0.0008	0.0008	0.0018
142	0.0007	0.0008	0.0015
143	0.0011	0.0009	0.0015
144	0.0007	0.0008	0.0016
145	0.0007	0.0007	0.0014
146	0.0007	0.0008	0.0016
147	0.0007	0.0008	0.0014
148	0.0007	0.0007	0.0013
149	0.0009	0.0008	0.0015
150	0.0008	0.0007	0.0014
151	0.0007	0.0007	0.0013
152	0.0007	0.0006	0.0015

Hidden nodes	1000	500	200
Epoch	1.0e+ 04 *	1.0e+ 04 *	1.0e+ 04 *
153	0.0008	0.0006	0.0012
154	0.0009	0.0007	0.0014
155	0.0008	0.0007	0.0014
156	0.0007	0.0008	0.0014
157	0.0008	0.0006	0.0012
158	0.0007	0.0006	0.0013
159	0.0008	0.0007	0.0011
160	0.0006	0.0006	0.0013
161	0.0006	0.0006	0.0011
162	0.0007	0.0006	0.0010
163	0.0006	0.0008	0.0014
164	0.0006	0.0007	0.0012
165	0.0007	0.0006	0.0013
166	0.0006	0.0007	0.0011
167	0.0007	0.0006	0.0012
168	0.0008	0.0006	0.0011
169	0.0007	0.0006	0.0009
170	0.0007	0.0005	0.0010
171	0.0006	0.0005	0.0011
172	0.0005	0.0006	0.0012
173	0.0006	0.0007	0.0009
174	0.0006	0.0006	0.0009
175	0.0007	0.0006	0.0010
176	0.0007	0.0005	0.0008
177	0.0005	0.0005	0.0011
178	0.0006	0.0006	0.0012
179	0.0006	0.0004	0.0010
180	0.0008	0.0007	0.0009
181	0.0006	0.0005	0.0010
182	0.0007	0.0005	0.0009
183	0.0005	0.0005	0.0011
184	0.0005	0.0007	0.0009
185	0.0005	0.0005	0.0008
186	0.0005	0.0005	0.0008
187	0.0005	0.0004	0.0009
188	0.0006	0.0005	0.0007
189	0.0006	0.0005	0.0009
190	0.0006	0.0004	0.0008
191	0.0005	0.0005	0.0008
192	0.0006	0.0004	0.0008
193	0.0007	0.0005	0.0007
194	0.0009	0.0004	0.0008
195	0.0006	0.0004	0.0008
196	0.0005	0.0006	0.0007
197	0.0005	0.0005	0.0007
198	0.0005	0.0005	0.0007
199	0.0005	0.0005	0.0008
200	0.0005	0.0005	0.0008

This page is blank.

Bibliography

ANDREADIS-I. I., SPYROU-G. M., AND NIKITA-K. S. (2015). A CADx Scheme for Mammography Empowered With Topological Information From Clustered Microcalcifications Atlases, *Journal of Biomedical and Health Informatics*, **19**(1), pp. 166–173.

AZEVEDOL-W. W., LIMAL-S. M. L., FEMANDES-I. M. M., ROCHA-A. D. D., CORDEIROL-F. R., DA SILVA-FILHOL-A. G., AND DOS SANTOS-W. P. (2015). Morphological Extreme Learning Machines Applied to Detect and Classify Masses in Mammograms, *IEEE Transactions on Biomedical Engineering*.

BALLAN-L., BAZZICA-A., BERTINI-M., DEL BIMBO-A., AND SERRA-G. (2009). Deep Networks for Audio Event Classification in Soccer Videos, *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 474–477.

BENGIO-Y. (2009). Learning Deep Architectures for AI, *Found. Trends Mach. Learn.*, **2**(1), pp. 1–127.

BENJELIL-M., KANOUN-S., MULLOT-R., AND ALIMI-A. (2009). Steerable Pyramid Based Complex Documents Images Segmentation, *10th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 833–837.

BEVERS-T. B., BONACCIO-E., ANDERSON-B. O., BUYS-S., DALY-M. B., DEMPSEY-P. J., FARRAR-W. B., FLEMING-I., GARBER-J. E., HARRIS-R. E., HEERDT-A. S., HELVIE-M., HUFF-J. G., KHAKPOUR-N., KHAN-S. A., KRONTIRAS-H., LYMAN-G., RAFFERTY-E., SHAW-S., SMITH-M. L., TSANGARIS-T. N., WILLIAMS-C., AND TANKEELOV-T. (2009). Breast Cancer Screening and Diagnosis, *National Comprehensive Cancer Network*, **7**(10), pp. 1060–1096.

BRACEWELL-R. (1999). *The Fourier Transform & Its Applications*, McGraw-Hill Series in Electrical and Computer Engineering, 3rd edn, McGraw-Hill Science/Engineering/Math.

Breast Screen Australia Monitoring Report 2009-2010 (2012).

BURRUS-C. S., GOPINATH-R. A., AND GUO-H. (1998). *Introduction to Wavelets and Wavelet Transforms: A Primer*, Prentice Hall.

CHAN-H. P., DOI-K., VYBORNY-C. J., SCHMIDT-R. A., METZ-C. E., LAM-K. L., OGURA-T., WU-Y. Z., AND MACMAHON-H. (1990). Improvement in Radiologists Detection of Clustered Microcalcifications on Mammograms – The Potential of Computer-Aided Diagnosis, *Investigative Radiology*, **25**(10), pp. 1102–1110.

CHAN-H., VYBORNY-C., MACMAHON-H., METZ-C., DOI-K., AND SICKLES-E. (1987). Digital Mammography: ROC Studies of the Effects of Pixel Size and Unsharp-Mask Filtering on the Detection of Subtle Microcalcifications, *Investigative Radiology*, **22**(7), pp. 581–589.

CHEN-C. H., AND LEE-G. G. (1997). On Digital Mammogram Segmentation and Microcalcification Detection using Multiresolution Wavelet Analysis, *Graphical Models and Image Processing*, **59**(5), pp. 349–364.

CHENG-H., CAI-X., CHEN-X., HU-L., AND LOU-X. (2003). Computer-Aided Detection and Classification of Microcalcifications in Mammograms: a Survey, *Pattern Recognition*, **36**(12), pp. 2967–2991.

COHEN-A., AND KOVACEVIC-J. (1996). Wavelets: The Mathematical Background, *Proceedings of IEEE*, **84**(4), pp. 514–522.

DEHGHAN-F., ABRISHAMI-MOGHADDAM-H., GITI-M., AND IEEE. (2008). Automatic Detection of Clustered Microcalcifications in Digital Mammograms: Study on Applying Adaboost with SVM-based Component Classifiers, *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4789–4792.

DE MELO-C. L. S., COSTA FILHO-C. F. F., COSTA-M. G. F., AND PEREIRA-W. C. A. (2010). Matching Input Variables Sets and Feedforward Neural Network Architectures in Automatic Classification of Microcalcifications and Microcalcification Clusters, *3rd International Conference on Biomedical Engineering and Informatics (BMEI)*, Vol. 1, pp. 358–362.

DEMIRHAN-A., TORU-M., AND GULER-I. (2015). Segmentation of Tumor and Edema Along With Healthy Tissues of Brain Using Wavelets and Neural Networks, *IEEE Biomedical and Health Informatics*, **19**(4), pp. 1451–1458.

DHAWAN-A. P., CHITRE-Y., AND KAISER-BONASSO-C. (1996). Analysis of Mammographic Microcalcifications using Gray-Level Image Structure Features, *IEEE Transactions on Medical Imaging*, **15**(3), pp. 246–259.

DHAWAN-A. P., CHITRE-Y. S., MOSKOWITZ-M., AND GRUENSTEIN-E. (1991). Classification of Mammographic Microcalcification and Structural Features using an Artificial Neural Network, *International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1105–1106.

EL AROUSSI-M., EL HASSOUNI-M., GHOUZALI-S., RZIZA-M., AND ABOUTAJDINE-D. (2009). Novel Face Recognition Approach Based on Steerable Pyramid Feature Extraction, *16th IEEE International Conference on Image Processing (ICIP)*, pp. 4165–4168.

FREEMAN-W. T., AND ADELSON-E. H. (1991). The Design and Use of Steerable Filters, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **13**(9), pp. 891–906.

FREER-T. W., AND ULISSEY-M. J. (2001). Screening Mammography with Computer-aided Detection: Prospective Study of 12,860 Patients in a Community Breast Center, *Radiology*, **220**(3), pp. 781–786.

GETTY-D., PICKETT-R., D'ORSI-C., AND SWETS-J. (1988). Enhanced Interpretation of Diagnositic Images, *Investigative Radiology*, **23**(4), pp. 240–252.

GHWANMEH-S. (2012). Applying Advanced NN-based Decision Support Scheme for Heart Diseases Diagnosis, *International Journal of Computer Applications*, **44**(2), pp. 37–41.

GORDON-R., AND RANGAYYAN-R. M. (1984). Feature Enhancement of Film Mammograms Using Fixed and Adaptive Neighborhoods, *Applied Optics*, **23**(4), pp. 560–564.

GORGEL-P., SERTBAS-A., AND UCAN-O. N. (2009). A Wavelet-Based Mammographic Image Denoising and Enhancement with Homomorphic Filtering, *Medical Systems*, **34**(6), pp. 993–1002.

GRAPS-A. (1995). An Introduction to Wavelets, *IEEE Computational Science and Engineering*.

GROMET-M. (2008). Comparison of Computer-Aided Detection to Double Reading of Screening Mammograms: Review of 231,221 Mammograms, *American Journal of Roentgenology*, **190**(4), pp. 854–859.

HEATH-M., BOWYER-K., KOPANS-D., KEGELMEYER-W. P., MOORE-R., CHANG-K., AND MUNISHKUMARAN-S. (1998). Current Status of the Digital Database for Screening Mammography, *Proceedings of the Fourth International Workshop on Digital Mammography*, Kluwer Academic Publishers, pp. 457 – 460.

HEATH-M., BOWYER-K., KOPANS-D., MOORE-R., AND KEGELMEYER-W. P. (2001). The Digital Database for Screening Mammography, *Proceedings of the Fifth International Workshop on Digital Mammography*, pp. 212–218.

HEINLEIN-P., DREXL-J., AND SCHNEIDER-W. (2003). Integrated Wavelets for Enhancement of Microcalcifications in Digital Mammography, *IEEE Transaction on Medical Imaging*, **22**(3), pp. 402–413.

HINTON-G. (2012). A Practical Guide to Training Restricted Boltzmann Machines, *Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science*, **7700**, pp. 599–619.

HINTON-G. E. (2007). Learning Multiple a Layers of Representation, *Trends in Cognitive Sciences*, **11**(10), pp. 428–434.

HINTON-G. E. (2010). Learning to Represent Visual Input, *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**(1537), pp. 177–184.

HINTON-G. E., AND SALAKHUTDINOV-R. (2006). Reducing the Dimensionality of Data with Neural Networks, *Science*, **313**(5786), pp. 504 – 507.

HINTON-G. E., OSINDERO-S., AND TEH-Y. W. (2006). A Fast Learning Algorithm for Deep Belief Nets, *Neural Computation*, **18**(7), pp. 1527–1554.

HUANG-W., SONG-G., HONG-H., AND XIE-K. (2014). Deep Architecture for Traffic Flow Prediction: Deep Belief Networks With Multitask Learning, *IEEE Transactions on Intelligent Transportation Systems*, **15**(5), pp. 2191–2201.

ISLAM-M. J., AHMADI-M., AND SID-AHMED-M. A. (2010). *Computer-Aided Detection and Classification of Masses in Digitized Mammograms Using Artificial Neural Network*, Vol. 6146, Springer Berlin Heidelberg, pp. 327–334.

KARABULUT-E. M., AND IBRIKCI-T. (2012). Effective Diagnosis of Coronary Artery Disease Using The Rotation Forest Ensemble Method, *Journal of Medical Systems*, **36**(5), pp. 3011–3018.

KNUTZEN-A., AND GISVOLD-J. (1993). Likelihood of Malignant Disease For Various Categories of Mammographically Detected, Nonpalpable Breast-Lesions, *Mayo Clinic Proceedings*, **68**(5), pp. 454–460.

LAINE-A., FAN-J., AND YANG-W. (1995a). Wavelets for Contrast Enhancement of Digital Mammography, *IEEE Engineering in Medicine and Biology*, **14**(5), pp. 536 – 550.

LAINE-A. F., SCHULER-S., FAN-J., AND HUDA-W. (1994). Mammographic Feature Enhancement by Multiscale Analysis, *IEEE Transaction on Medical Imaging*, **13**(4), pp. 725–740.

LAINE-A., HUDA-W., STEINBACH-B. G., AND HONEYMAN-J. C. (1995b). Mammographic Image-Processing Using Wavelet Processing Techniques, *European Radiology*, **5**(5), pp. 518–523.

LATTANZIO-V., AND SIMONETTI-G. (2010). *Mammography: Guide to Interpreting, Reporting and Auditing Mammographic Images-Re. Co. RM (From Italian Reporting and Codifying the Results of Mammography)*, Springer.

LEE-C. S., KIM-J. K., AND PARK-H. W. (1998). Computer-Aided Diagnostic System for Breast Cancer by Detecting Microcalcifications, *Image Display - Medical Imaging*, **3335**, pp. 615–626.

MCSWEENEY-M., AND SPRAWLS-P. (1983). Enhanced Image Mammography, *American Journal of Roentgenology*, **140**(1), pp. 9–14.

MORROW-W. M., PARANJAPE-R. B., RANGAYYAN-R. M., AND DESAUTELS-J. E. L. (1992). Region-based Contrast Enhancement of Mammograms, *IEEE Transactions on Medical Imaging*, **11**(3), pp. 392–406.

NEOFYTOU-M. S., TANOS-V., CONSTANTINOU-I., KYRIACOU-E. C., PATTICHIS-M. S., AND PATTICHIS-C. S. (2015). Computer-Aided Diagnosis in Hysteroscopic Imaging, *IEEE Biomedical and Health Informatics*, **19**(3), pp. 1129–1136.

NISHIKAWA-R. M. (2002). *Detection of Microcalcifications*, CRC, chapter 6.

NISHIKAWA-R. M. (2007). Current Status and Future Direct of Computer-Aided Diagnosis in Mammography, *Computerized Medical Imaging and Graphics*, **31**(4-5), pp. 224 – 235.

NISHIKAWA-R. M. (2010). *Digital Mammography*, Springer Berlin Heidelberg, chapter Computer-Aided Detection and Diagnosis, pp. 85–106.

O.DUDA-R., HART-P. E., AND STORK-D. G. (2000). *Pattern Classficationi*, 2nd edn, John Wiley & Sons, Inc.

PU-T., AND NI-G. (2000). Contrast-Based Image Fusion using the Discrete Wavelet Transform, *Optical Engineering*, **39**(8), pp. 2075–2082.

RANGAYYAN-R. M. (2005). *Biomedical Image Analysis*, CRC Press.

RANGAYYAN-R. M., AYRES-F. J., AND LEO DESAUTELS-J. E. (2007). A Review of Computer-Aided Diagnosis of Breast Cancer: Toward the Detection of Subtle Signs, *Journal of the Franklin Institute*, **344**(3-4), pp. 312–348.

RAYTCHEV-B., YODA-I., AND SAKAUE-K. (2004). Head Pose Estimation by Nonlinear Manifold Learning, *In Proceedings of the 17th ICPR*, **4**, pp. 462 – 466.

REN-J. (2012). ANN vs. SVM: Which One Performs Better in Classification of MCCs in Mammogram Imaging, *Knowledge-Based Systems*, **26**, pp. 144–153.

REN-J., WAN-D., AND JIANG-J. (2011). Effective Recognition of MCCs in Mammograms Using An Improved Neural Classifier, *Engineering Applications of Artificial Intelligence*, **24**, pp. 638–645.

RIOUL-O., AND VETTERLI-M. (1991). Wavelets and Signal Processing, *IEEE Signal Processing Magazine*, pp. 14–38.

RITA E. SOHLICH, EDWARD A. SICKLES-E. S. B., AND DEE-K. E. (2002). Interpreting Data from Audits when Screening and Diagnostic Mammography Outcomes are Combined, *American Journal of Roentgenology*, **178**(3), pp. 681–686.

ROBSON-K. J. (2010). Advances in Mammographic Imaging, *The British Journal of Radiology*, **83**, pp. 273–275.

SAKELLAROPOULOS-P., COSTARIDOU-L., AND PANAYIOTAKIS-G. (2003). A Wavelet-Based Spatially Adaptive Method for Mammographic Contrast Enhancement, *Physics in Medicine and Biology*, **48**(6), pp. 787 – 803.

SALVADO-J., AND ROQUE-B. (2005). Detection of Calcifications in Digital Mammograms using Wavelet Analysis and Contrast Enhancement, *IEEE International Workshop on Intelligent Signal Processing*, pp. 200–205.

SCHARCANSKI-J., AND JUNG-C. R. (2006). Denoising and Enhancing Digital Mammographic Images for Visual Screening, *Computerized Medical Imaging and Graphics*, **30**(4), pp. 243–254.

SHEN-S., SANDHAM-W., GRANAT-M., AND STERR-A. (2005). MRI Fuzzy Segmentation of Brain Tissue Using Neighborhood Attraction With Neural-Network Optimization, *IEEE Information Technology in Biomedicine*, **9**(3), pp. 459–467.

SHUSEN-Z., QINGCAI-C., AND XIAOLONG-W. (2010). Discriminative Deep Belief Networks for Image Classification, *17th IEEE International Conference on Image Processing (ICIP)*, pp. 1561–1564.

SIMONCELLI-E. (1994). Design of Multi-Dimensional Derivative Filters, *IEEE International Conference on Image Processing (ICIP)*, Vol. 1, pp. 790–794.

SIMONCELLI-E. P., AND FREEMAN-W. T. (1995). The Steerable Pyramid: A Flexible Architecture for Multi-Scale Derivative Computation, *IEEE International Conference on Image Processing*, pp. 444–447.

SMITHUIS-R., AND PIJNAPPEL-R. (2008). Breast – Calcifications Differential Diagnosis, Radiology Assistant, Rijnland Hospital, Leiderdorp and Martini Ziekenhuis, Groningen, the Netherlands, http://www.radiologyassistant.nl/en/p4793bfde0ed53 (publication date May 11, 2008).

SOLTANIAN-ZADEH-H., POURABDOLLAH-NEZHAD-S., AND RAFIEE-RAD-F. (2001). Shape-Based and Texture-Based Feature Extraction for Classification of Microcalcifications in Mammograms, *Medical Imaging: Image Processing*, **2**, pp. 301–310.

SOLTANIAN-ZADEH-H., RAFIEE-RAD-F., AND POURABDOLLAH-NEJAD D-S. (2004). Comparison of Multiwavelet, Wavelet, Haralick, and Shape Features for Microcalcification Classification in Mammograms, *Pattern Recognition*, **37**(10), pp. 1973–1986.

SONGYANG-Y., AND LING-G. (2000). A CAD System for the Automatic Detection of Clustered Microcalcifications in Digitized Mammogram Films, *IEEE Transactions on Medical Imaging*, **19**(2), pp. 115–126.

SPIESBERGER-W. (1979). Mammogram Inspection by Computer, *IEEE Transactions on Biomedical Engineering*, **26**(4), pp. 213–219.

STRICKLAND-R. N., AND HEE IL-H. (1996). Wavelet Transforms for Detecting Microcalcifications in Mammograms, *IEEE Transactions on Medical Imaging*, **15**(2), pp. 218–229.

TAMILSELVAN-P., WANG-Y., AND WANG-P. (2012). Deep Belief Network Based State Classification for Structural Health Diagnosis, *2012 IEEE Aerospace Conference*, p. doi:10.1109/AERO.2012.6187366.

TANG-J., LIU-X., AND SUN-Q. (2009a). A Direct Image Contrast Enhancement Algorithm in the Wavelet Domain for Screening Mammograms, *IEEE Journal of Selected Topics in Signal Processing*, **3**(1), pp. 74 – 80.

TANG-J., RANGAYYAN-R., XU-J., EL NAQA-I., AND YANG-Y. (2009b). Computer-Aided Detection and Diagnosis of Breast Cancer With Mammography: Recent Advances, *IEEE Transactions on Information Technology in Biomedicine*, **13**(2), pp. 236 –251.

The Radiology Information Resource For Patients: Mammography (2011).

TIRTAJAYA-A., AND SANTIKA-D. D. (2010). Classification of Microcalcification Using Dual-Tree Complex Wavelet Transform and Support Vector Machine, *2nd International Conference on Advances in Computing, Control, and Telecommunication Technologies*, pp. 164 – 166.

TSUJII-O., FREEDMAN-M. T., AND MUN-S. K. (1999). Classification of Microcalcifications in Digital Mammograms using Trend-Oriented Radial Basis Function Neural Network, *Pattern Recognition*, **32**(5), pp. 891–903.

UGUZ-H. (2012). A biomedical system based on artificial neural network and principal component analysis for diagnosis of the heart valve diseases, *Journal of Medical Systems*, **36**(1), pp. 61–72.

VAN DER MAATEN-L., POSTMA-E. O., AND VAN DEN HERIK-H. J. (2008). Dimensionality Reduction: A Comparative Review, *Tilburg University Technical Report, TiCC-TR 2009-005*.

VAPNIK-V. (1995). *The Nature of Statistical Learning Theory*, Springer.

VELDKAMP-W. J. H., AND KARSSEMEIJER-N. (1999). An Improved Method for Detection of Microcalcification Clusters in Digital Mammograms, *Proceedings of SPIE*, **3661**, pp. 512–522.

VENKATARAJAN-M., AND BRAUN-W. (2004). New Quantitative Descriptors of Amino Acids based on Multidimensional Scaling of a Large Number of Physicalchemical Properties, *Journal of Molecular Modeling*, **7(12)**, pp. 445 – 453.

VERMA-B. (1998). A Neural Network Based Technique to Locate and Classify Microcalcifications in Digital Mammograms, *IEEE World Congress on Computational Intelligence Neural Networks*, Vol. 3, pp. 1790–1793.

VERMA-B., AND ZAKOS-J. (2001). A Computer-Aided Diagnosis System for Digital Mammograms Based on Fuzzy-Neural and Feature Extraction Techniques, *IEEE Transactions on Information Technology in Biomedicine*, **5**(1), pp. 46–54.

VETTERLI-M., AND KOVACEVIC-J. (1995). *Wavelets and Subband Coding*, Prentice Hall Signal Processing Series.

WARREN-R. M. L., AND DUFFY-S. W. (1995). Comparison of Single Reading With Double Reading Of Mammograms and Change In Effectiveness with Experience, *British Journal of Radiology*, **68**(813), pp. 958–962.

WEI-L. Y., YANG-Y. Y., NISHIKAWA-R. M., AND JIANG-Y. L. (2005). A Study on Several Machine-Learning Methods for Classification of Malignant and Benign Clustered Microcalcifications, *IEEE Transactions on Medical Imaging*, **24**(3), pp. 371–380.

WHO Cancer Fact Sheet (2014).

WINSBERG-F., ELKIN-M., MACY-J., BORDAZ-V., AND WEYMOUTH-W. (1967). Detection Of Radiographic Abnormalities In Mammograms By Means Of Optical Scanning And Computer Analysis, *Radiology*, **89**(2), pp. 211–215.

WONG-M. T., HE-X., NGUYEN-H., AND YEH-W.-C. (2012). *Mass Classification in Digitized Mammograms Using Texture Features and Artificial Neural Network*, Vol. 7667, Springer Berlin Heidelberg, pp. 151–158.

YU-D., AND DENG-L. (2011). Deep Learning and Its Applications to Signal and Information Processing, *IEEE Signal Processing Magazine*, **28**(1), pp. 145 – 154.

List of Acronyms

AM-DBN	associated memory DBN
ANN	artificial neural network
BP-DBN	back-propagation DBN
BP-NN	back-propagation neural network
CAD	computer-aided diagnostic
CADe	computer-aided diagnostic detection
CADx	computer-aided diagnostic diagnosis
CC	cranio-caudal
CD	contrastive divergence
DBN	deep belief network
DDSM	Digital Database for Screening Mammography
FDR	fishers discriminant ratio
FFNN	feed-forward neural network
FN	false negative
FP	false positive
GA	genetic algorithm
kNN	k-nearest neighbouring
MLO	mediolateral-oblique

MLP	multi-layer perceptron
MRA	multiple resolution analysis
MRF	Markov random field
RBF	radial-basis function
RBM	restricted Boltzmann machine
RCV	region contrast value
ROC	receiving operating characteristic
ROI	region of interest
SNR	signal to noise ratio
SP	steerable pyramid
SP-DBN	steerable pyramid deep belief network
SRDM	surrounding region-dependence method
STFT	short time Fourier transform
SVM	support vector machine
TDLU	terminal ductal lobular unit
TN	true negative
TP	true positive
WHO	World Health Organization
WT	wavelet transform

Index

abnormalities
    architectural distortions, 4
    asymmetric, 4
    mass, 4, 5
    microcalcification, 4, 6, 16, 61, 100, 103
activation unit
    Bernoulli, 84, 138
    Gaussian, 84, 138
autoencoder, 92, 149

basis filters, 52, 54, 59, 108, 109, 111, 113, 114
boxplot analysis, 103, 112, 118, 125

chain code, 100, 102
classification, 66, 107
classifiers
    supervised, 66, 77, 155
    unsupervised, 66, 149
code layer, 149
computer-aided diagnostic, 9
    detection, 11, 15
    diagnosis, 9, 11, 15, 26, 66
confusion matrix, 106
contrast stretching, 17
contrastive divergence, 87, 140
cross validation, 106

data acquisition, 98
data normalization, 121
data scaling, 104
deep belief network, 97, 138, 147, 150, 156
    associcated memory, 89
    back-propagation, 89
Digital Database for Screening Mammography, 98
dynamic range, 104

early screening, 3
eigenvalues, 122, 123
eigenvectors, 122
energy, 117, 123, 131
enhancement
    conventional, 17
    features-based, 19
    region-based, 19
entropy, 117, 123, 131
equilibrium, 86
Euclidian distance, 71

feature extraction, 66, 97, 108, 138, 149
Fourier analysis, 38
Fourier series, 38
Fourier transform, 40, 42, 50

Gibbs phenomenon, 39
Gibbs sampling, 86
gradient descent, 77
greedy training, 83
grid-search, 106, 118
ground truth, 99, 102
growing seed, 19

hidden layer, 75, 82, 139, 147
histogram equalisation, 18
hyperplane, 68

kernel
Gaussian radial-basis function, 74, 118
linear, 74
polynomial, 74
Kuhn-Tucker constraints, 72

Lagrangian formulation, 72
learning rate, 78

mammogram
bengin, 155
benign, 5, 7, 99, 100, 102, 106
cranio-caudal, 3, 99
malignant, 5, 7, 99, 100, 102, 106, 155
mediolateral-oblique, 3, 99
mammography, 2, 98
multi-layer perceptron, 82
multiple orientation, 107, 113, 138
multiple resolution, 107
multiple resolution analysis, 43
multiple resolution analysis, MRA, 109

neural network, 30, 67, 139
back-propagation, 77
feed-forward, 30, 31, 75
neurons, 74

pre-processing
denoising, 22
principal component analysis, 108, 121, 124, 125

receiving operating characteristic, 107
receiving operator characteristic, 137, 150
reconstruction error, 113
region contrast value, 21
region of interest, 23, 66, 98, 100, 102, 110, 149
restricted Boltzmann machine, 84, 138, 141, 146

segmentation
individual, 24
region-based
multi-scale, 25
statistical, 25
sensitivity, 106
short time Fourier transform, 41
signal to noise ratio, 109, 110
specificity, 107
statistical features
energy, 164
entropy, 164
steerable filters, 59
steerable pyramid, 50, 51, 57, 107, 111, 115, 123, 131, 136, 156
sum squared error, 77, 141, 142
support vector machine, 30, 65, 97, 105, 106, 116, 137, 155, 157
system evaluation
false negative, 106, 150
false positive, 106, 150
true negative, 106, 150
true positive, 106, 150

terminal ductal lobular unit, 6
thresholding, 23, 68
time-frequency localisation, 50
time-frequency localization, 42

unsharp mask, 18

wavelet shrinkage, 22
wavelet transform, 20, 29, 47, 109, 116, 123

Biography

Aqilah Baseri Huddin is a tutor of Electrical, Electronic and Systems Engineering at the National University of Malaysia. She received her First Class Bachelor of Engineering from the University of Adelaide in 2007. She gained industrial experience; working as an Industrial Trainer at the Perusahaan Otomobil Kedua Limited Company (Perodua Motors Ltd. Co); an automobile manufacturer in Malaysia. She also gained research experience through her attachment at the Sultan Idris Education University in Malaysia working as a research assistant.

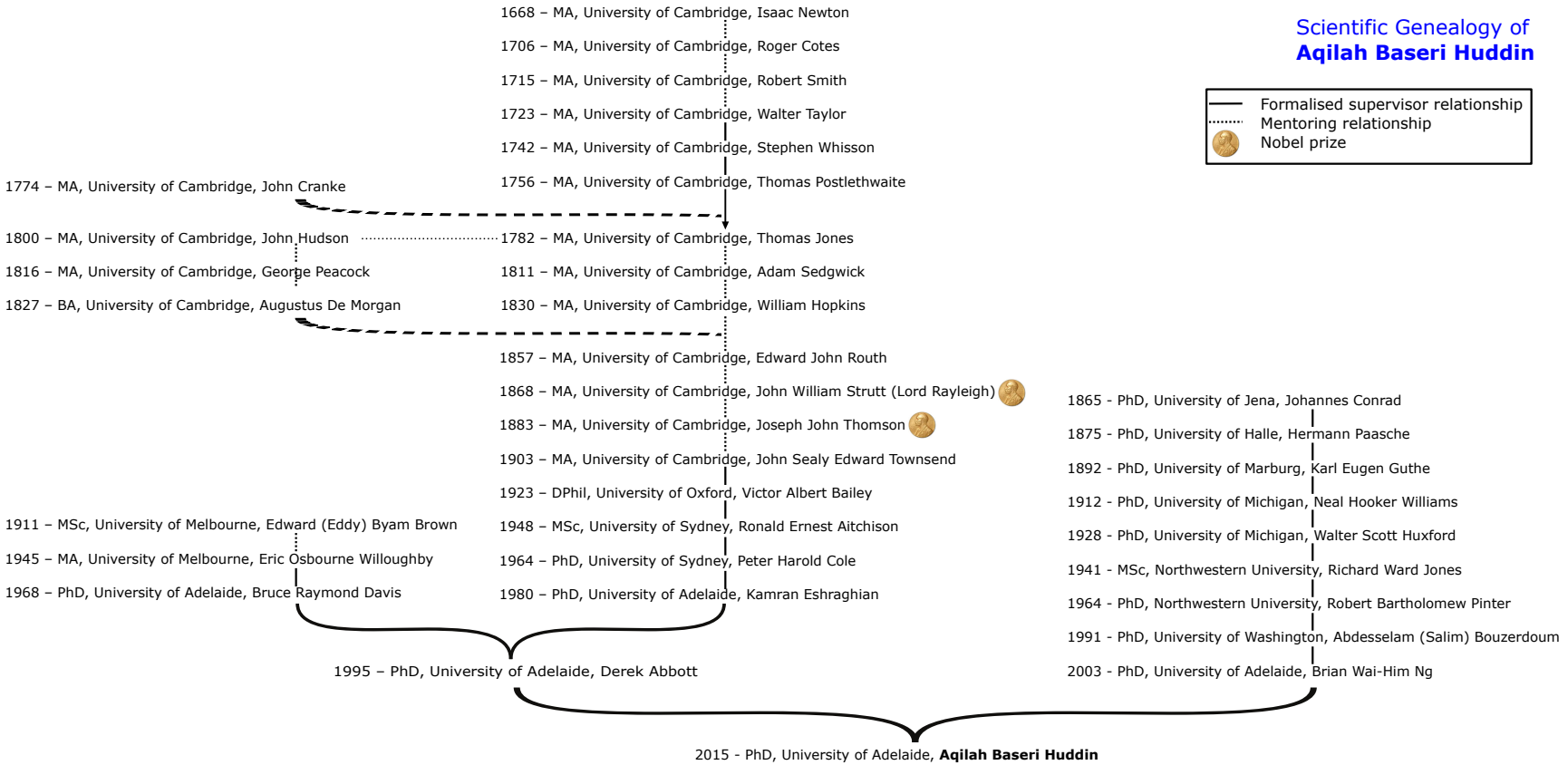
Following her First Class Honours Degree from the University of Adelaide, she was accepted by the National University of Malaysia to the tutorship appointment. She served the Electrical, Electronic and Systems Engineering Department of the university; assisting several Professors in the teaching and learning activities including several laboratory teaching supervision and evaluation of students' project reports. Following her experience received during her span as an Industrial Trainer during the undergraduate studies at the University of Adelaide, she gained competency as a visiting supervisor to several students from her department, who were undergoing industrial practice in several locations in Kuala Lumpur, Malaysia.

After serving for two years at the National University of Malaysia, she won the Malaysian Government Scholarship that entitled her to embark on a research study towards a PhD. She decided to return to Adelaide, one of *the happiest places to live in the world*. This time she went over to Adelaide accompanied by her newly wed husband, who also won a scholarship for a PhD study at the School of Computer Science of the same university.

For her PhD thesis, she worked on the application of signal processing in medical modality, ie. improving breast cancer diagnostic on the mammogram under the supervision of Dr Brian Ng and Professor Derek Abbott. The study involved the in-depth investigation and implementation of new techniques to improve the efficiency of the breast screening, specifically in diagnosing tiny and non-palpable tumor microcalcification in mammograms.

Now Aqilah Baseri Huddin is continuing her career as a member of academic staff at the National University of Malaysia. She is now actively involved in various academic activities in the Department of Electrical, Electronic and Systems Engineering, intending to develop herself to be a professional university educator.

Aqilah Baseri Huddin
aqilah@ukm.edu.my


Scientific Genealogy of
Aqilah Baseri Huddin
Formalised supervisor relationship
Mentoring relationship
Nobel prize
1668 – MA, University of Cambridge, Isaac Newton
1706 – MA, University of Cambridge, Roger Cotes
1715 – MA, University of Cambridge, Robert Smith
1723 – MA, University of Cambridge, Walter Taylor
1742 – MA, University of Cambridge, Stephen Whisson
1756 – MA, University of Cambridge, Thomas Postlethwaite
1774 – MA, University of Cambridge, John Cranke
1782 – MA, University of Cambridge, Thomas Jones
1800 – MA, University of Cambridge, John Hudson
1811 – MA, University of Cambridge, Adam Sedgwick
1816 – MA, University of Cambridge, George Peacock
1830 – MA, University of Cambridge, William Hopkins
1827 – BA, University of Cambridge, Augustus De Morgan
1857 – MA, University of Cambridge, Edward John Routh
1868 – MA, University of Cambridge, John William Strutt (Lord Rayleigh)
1883 – MA, University of Cambridge, Joseph John Thomson
1903 – MA, University of Cambridge, John Sealy Edward Townsend
1923 – DPhil, University of Oxford, Victor Albert Bailey
1911 – MSc, University of Melbourne, Edward (Eddy) Byam Brown
1948 – MSc, University of Sydney, Ronald Ernest Aitchison
1945 – MA, University of Melbourne, Eric Osbourne Willoughby
1964 – PhD, University of Sydney, Peter Harold Cole
1968 – PhD, University of Adelaide, Bruce Raymond Davis
1980 – PhD, University of Adelaide, Kamran Eshraghian
1865 - PhD, University of Jena, Johannes Conrad
1875 - PhD, University of Halle, Hermann Paasche
1892 - PhD, University of Marburg, Karl Eugen Guthe
1912 - PhD, University of Michigan, Neal Hooker Williams
1928 - PhD, University of Michigan, Walter Scott Huxford
1941 - MSc, Northwestern University, Richard Ward Jones
1964 - PhD, Northwestern University, Robert Bartholomew Pinter
1991 - PhD, University of Washington, Abdesselam (Salim) Bouzerdoum
2003 - PhD, University of Adelaide, Brian Wai-Him Ng
1995 – PhD, University of Adelaide, Derek Abbott
2015 - PhD, University of Adelaide, Aqilah Baseri Huddin