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