# Supervised Pattern Recognition

## "Case Study on Using Pattern Recognition System for Providing Second Opinion in the Diagnosis of Breast Cancer"

A thesis submitted in partial fulfillment of the requirements for the Master Degree in Applied Statistics

By

**Muhamed Wael Farouq Abdel-Azeim**
Demonstrator Department of Statistics, Mathematics & Insurance
Faculty of Commerce Ain Shams University

Under supervision of

**Dr. Medhat Mohamed Abdel Aal**
Assistant Professor – Mathematics & Statistics Department
Faculty of Commerce Ain Shams University

**Prof. Dr. Abdel Badeeh Mohamed Salem**
Professor - Faculty of Computer and Information Sciences
Ain Shams University

**Prof. Dr. Hala Abou Sena**
Professor - Faculty of Medicine
Ain Shams University

**2009**

# *Approval Sheet*

**Title of thesis : Supervised Pattern Recognition "Case Study on Using Pattern Recognition System for Providing Second Opinion in the Diagnosis of Breast Cancer"**

**Academic Degree :** M.Sc. in Applied Statistics
**Name of student :** Muhamed Wael Farouq Abdel-Azeim.

This thesis submitted in partial fulfillment of the requirements for the Master Degree in Applied Statistics has been approved by:

## Examination committee

**1- Professor Dr. Mustafa Galal Mustafa**
Professor – Mathematics & Statistics Department
Faculty of Commerce
Ain Shams niversity ......................................................

**2- Professor Dr. Abdel Badeeh Mohamed Salem**
Professor - Faculty of Computer and Information Sciences
Ain Shams University ......................................................

**3- Professor Dr. Hala Abou Sena**
Professor - Faculty of Medicine
Ain Shams University ......................................................

**4-Professor Dr. Abdulla Ahmed Abd El-Ghaly**
Professor - Faculty of Economics and Political Sciences
Cairo University ......................................................

**5-Dr. Medhat Mohamed Abdel Aal**
Assistant Professor – Mathematics & Statistics Department
Faculty of Commerce
Ain Shams University ......................................................

**Date of Dissertation Defense / / 2009 Approval date / / 2009**

Acknowledgements

First and foremost, I would like to thank ALLAH for giving me the opportunity and strength to accomplish this research work.

I sincerely acknowledge and express my deep appreciation to **Professor Dr. Mustafa Galal Mustafa** for his support, encouragement, and patience during my years of study.

I would like to express my gratitude to my supervisor **Professor Dr.Medhat Mohamed Abdelaal** for allowing me to work on this project. I have learned a lot from him, not only about the technical aspect of this thesis, but also about the methods of performing a small-scale research project. His constructive appraisal has kept me motivated and interested in this work the whole time.

My gratitude also goes to **Professor Dr. Abdel-Badeeh Mohamed Salem** for valuable discussions and guidance in all aspects of the research.

My deepest appreciations go to **Professor Dr. Hala Abou Sena** for her generous support in the medical aspects of this work, also for her friendly supervision, patience, technical guidance and support.

I strongly acknowledge the contribution of **Professor Dr. Abdulla Ahmed Abd El-Ghaly** for his valuable time, patience and professionalism.

I would like to express my sincere thanks to all who have stood by me during the course of this degree. I could not have completed this work without their guidance, encouragement and support throughout the difficult times.

My special thanks go to my friends and many others for their help and goodwill. Last but not least, I am indebted to my parents, brothers and sisters for all their support and comfort.

*To the soul of my father….*

*Muhamed Wael Farouq*

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيمِ

﴿وَمَا تَوْفِيقِي إِلاَّ
بِاللهِ عَلَيْهِ تَوَكَّلْتُ
وَإِلَيْهِ أُنِيبُ﴾

بسم الله الرحمن الرحيم

إهـــــداء

إننى أتشرف بإهداء هذا العمل المتواضع

الى معلمي الأول

إلى ينبوع الحنان الذي لا ينضب

إلى مثال الإرادة والصبر والتحمل

إلى من كنت أتمنى أن يشاركني هذه اللحظات

وأرى الابتسامة تملأ وجهه الجميل

وتملأ معها حياتى بالسعادة والأمل

إلى أبى

"اللهم إغفر له وارحمه واسكنه فسيح جناتك واجعل قبره

روضة من رياض الجنة... اللهم ارزقه الفردوس الأعلى....اللهم إنى

دعوتك كما أمرتنى فاستجب لى كما وعدتنى"

اللهم آمين.. آمين

وائل

كلية التجارة
قسم الإحصاء والرياضة والتأمين

## رسالة ماجستير

اسم الباحث: محمد وائل فاروق عبد العظيم

عنوان الرسالة: نظام تمييز الأنماط "دراسة تحليلية لاستخدام نظام تمييز الأنماط لإبداء رأي مساعد لتشخيص أورام الثدي"

الدرجة العلمية: ماجستير إحصاء تطبيقي

تاريخ الرسالة: ٢٠٠٩م

## لجنة المناقشة والحكم

١- الأستاذ الدكتور/ مصطفى جلال مصطفى

أستاذ الإحصاء قسم الإحصاء والرياضة والتأمين كلية التجارة – جامعة عين شمس

٢- الأستاذ الدكتور /عبد البديع محمد محمود سالم

الأستاذ بكلية الحاسبات والمعلومات – جامعة عين شمس

٣- الأستاذة الدكتورة/ هالة أبو سنة

الأستاذة بكلية الطب – جامعة عين شمس

٤- الأستاذة الدكتورة/ عبد الله أحمد عبد الغالى

أستاذ الإحصاء بكلية الاقتصاد والعلوم السياسية – جامعة القاهرة

٥- الأستاذ الدكتور/ مدحت محمد أحمد عبد العال

أستاذ الإحصاء والرياضة والتأمين المساعد بكلية التجارة – جامعة عين شمس

***********************************

## الدراسات العليا

**ختم الإجازة** | **أجيزت الدراسة**

**بتاريخ / / ٢٠٠٩**

**موافقة مجلس الكلية**

**بتاريخ / / ٢٠٠٩**

**موافقة مجلس الجامعة**

**بتاريخ / / ٢٠٠٩**

## Table of Contents

**List of Tables** ........ **III**

**List of Figures** ........ **V**

**List of Abbreviation** ........ **VII**

**Chapter 1: Introduction**

**1.1 Introduction** ........ **1**

**1.2 Overview of Breast Cancer** ........ **3**

**1.2.1 Stages of Breast Cancer** ........ **7**

**1.2.2 Diagnosis of Breast Cancer** ........ **7**

**1.2.3 Treatment of breast cancer** ........ **12**

**1.3 Thesis Importance** ........ **13**

**1.4 Thesis Objective** ........ **15**

**1.5 Thesis Outline** ........ **15**

**Chapter 2: Classifying Techniques**

**2.1 Discriminant Analysis** ........ **18**

**2.1.1 The Decision Process for Disciminant Analysis** ........ **21**

**2.2 Logistic Regression** ........ **32**

**2.2.1 Assessing the Goodness of Fit of the Estimated Model** ........ **33**

**2.3 Artificial Neural Networks** ........ **34**

**2.3.1 Introduction** ........ **34**

**2.3.2 Biological Neurons** ........ **36**

**2.3.3 Artificial Neural Networks (ANN)** ........ **41**

**2.3.3.1 Architecture** ........ **41**

**2.3.3.2 Learning Algorithm** ........ **47**

**2.3.3.3 Activation Function** ........ **50**

**2.3.4 Multilayer Perceptrons (MLPs)** ........ **54**

**2.3.4.1 Architecture** ........ **54**

**2.3.4.2 Nomenclature** ........ **55**

**2.3.4.3 Algorithm** ........ **57**

**2.3.4.4 Training Algorithm** ........ **58**

**2.3.4.5 Problems and Features of MLP** ........ **63**

**Chapter 3: Digital Image Processing**

**3.1 Introduction ........ 68**

**3.2 Image Sampling and Quantization ........ 73**

**3.3 Image Segmentation ........ 77**

**Chapter 4: Data Analysis Approaches for the Wisconsin Dataset**

**4.1 Introduction........ 86**

**4.2 Nature of the Variables ........ 86**

**4.3 Extraction of Digital Nuclear Features ........ 91**

**4.4 Discriminant Analysis ........ 97**

**4.5 Logistic Regression ........ 100**

**4.6 Neural Networks ........ 103**

**Chapter 5: Data Analysis Approaches for the DDSM Dataset**

**5.1 Introduction........ 109**

**5.2 Nature of the Variables ........ 110**

**5.3 Extraction of Digital Nuclear Features ........ 113**

**5.4 Discriminant Analysis ........ 118**

**5.5 Logistic Regression ........ 121**

**5.6 Neural Networks ........ 124**

**Chapter 6: Results and Conclusions ........ 129**

**References........ 134**

**Appendix A: Sensitivity about the mean-Wisconsin dataset ........ 139**

**Appendix B: Sensitivity about the mean-DDSM dataset ........ 144**

# List of Tables

Table (1.1): Age-Specific Probabilities of Developing Invasive breast Cancer ............................ 3
Table (1.2): Site Distribution of Cancer at NCI, Cairo University (1990-1997) .......................... 6

Table (2.1): Some correspondences between the discipline concerns of machine learning and statistics ........................................................................................................ 34

Table (2.2): ANN types and applications ....................................................................... 48

Table (4.1): Variables entered the equation, Wisconsin dataset ................................................... 97
Table (4.2): Variables entered/removed, Wisconsin dataset........................................................ 97
Table (4.3): Significance of the discriminant function, Wisconsin dataset ................................. 98
Table (4.4): Variables in the analysis, Wisconsin dataset............................................................ 98
Table (4.5): Eigen Values, Wisconsin dataset .............................................................................. 98
Table (4.6): Fisher's linear discriminant function coefficients, Wisconsin dataset. 98
Table (4.7): Classification Results, Wisconsin dataset................................................................ 99
Table (4.8): Variables in the equation=Step 0, Wisconsin dataset ............................................... 100
Table (4.9): Variables not in the equation-Step 0, Wisconsin dataset .......................................... 100
Table (4.10): Variables entered the equation, Wisconsin dataset ................................................. 100
Table (4.11): Model summary, Wisconsin dataset ....................................................................... 101
Table (4.12): Hosmer & Lemeshow test, Wisconsin dataset........................................................ 101
Table (4.13): Variables in the equation, Wisconsin dataset......................................................... 102
Table (4.14): Classification table, Wisconsin dataset .................................................................. 102
Table (4.15): Results of Training and cross validation of Wisconsin dataset.............................. 104
Table (4.16-a): Classification matrix of Wisconsin dataset ......................................................... 104
Table (4.16-b): Performance of Wisconsin dataset....................................................................... 104
Table (4.17): Sensitivity of Wisconsin dataset ............................................................................ 106

Table (5.1): Variables entered the equation, DDSM Dataset....................................................... 118
Table (5.2): Variables entered/removed, DDSM Dataset ............................................................. 118
Table (5.3): Significance of the discriminant function, DDSM Dataset....................................... 119
Table (5.4): Variables in the analysis, DDSM Dataset ................................................................. 119
Table (5.5): Eigen Values, DDSM Dataset................................................................................... 119
Table (5.6): Fisher's linear discriminant function coefficients, DDSM dataset 119
Table (5.7): Classification Results, DDSM Dataset ..................................................................... 120
Table (5.8): Variables in the equation=Step 0, DDSM Dataset.................................................... 121
Table (5.9): Variables not in the equation-Step 0, DDSM Dataset............................................... 121

Table (5.10): Variables entered the equation, DDSM Dataset........................................................ 121
Table (5.11): Model summary, DDSM Dataset................................................................................ 122
Table (5.12): Hosmer & Lemeshow test, DDSM Dataset .............................................................. 122
Table (5.13): Variables in the equation, DDSM Dataset ................................................................ 123
Table (5.14): Classification table, DDSM Dataset ......................................................................... 123
Table (5.15): Results of Training of DDSM dataset....................................................................... 125
Table (5.16-a): Classification matrix of DDSM dataset ................................................................. 125
Table (5.16-b): Performance of DDSM dataset.............................................................................. 125
Table (5.17): Sensitivity of DDSM dataset..................................................................................... 127

Table (6.1): Comparison table for the data analysis approaches……………………………….. 131

# List of Figures

Figure (1.1): The incidence of breast cancer in England women in 2001 per 100,000 population as a function of age ........ 4
Figure (1.2): Craniocaudal vie of left breast in mammogram image ........ 10
Figure (1.3): Mediolateral view of left breast in mammogram image ........ 10

Figure (2.1): Univariate Representation of Discriminant Z Scores ........ 20
Figure (2.2): Stages 1-3 in the Discriminant Analysis Decision Diagram ........ 21
Figure (2.3): Stages 4-6 in the Discriminat Analysis Decision Diagram ........ 22
Figure (2.4): Optimal Cutting Score with Equal Sample Sizes ........ 29
Figure (2.5): Optimal Cutting Score with Unequal Sample Sizes ........ 29
Figure (2.6): Examples of Fitting the Logistic Curve to Sample Data ........ 33
Figure (2.7): A structure of a typical biological neuron. ........ 37
Figure (2.8): The generalization principle illustrated as a mapping of the domain to the solution space for a new input x'. ........ 39
Figure (2.9): Neurons in the hidden and output layers ........ 43
Figure (2.10): A single-layer neural net ........ 44
Figure (2.11): A multilayer neural net ........ 45
Figure (2.12): Autoassociative and hetero associative types of neural networks ........ 46
Figure (2.13): Identity Function ........ 50
Figure (2.14): Binary step function ........ 51
Figure (2.15): Binary sigmoid, steepness parameters $\sigma$ =1 and $\sigma$ =3 ........ 52
Figure (2.16): Bipolar Sigmoid ........ 53
Figure (2.17): Backpropagation neural network with one hidden layer ........ 55
Figure (2.18): Schematic representation of the forward propagation process of the function signal and backpropagation of the error signal ........ 62
Figure (2.19): The "overtraining" phenomenon ........ 64
Figure (2.20): Global minimum and local minima of the error function: a one-dimensional example ........ 66

Figure (3.1): Representation of digital images by arrays of discrete points on a rectangular grid: a 2-D image, b 3-D image. ........ 69

Figure (3.2): Digital images consist of pixels. ............................................................................. 70
Figure (3.3): The Electromagnetic spectrum arranged according to energy per photon................ 72
Figure (3.4): Generating a digital image. (a) Continuous image. (b) A scan line from A to B in the continuous image, used to illustrate the concepts of sampling and quantization. (c) Sampling and quantization. (d) Digital scan line. ...................... 74
Figure (3.5): (a) Continuos image projected onto a sensor array.(b) Result of image sampling and quantization. ........................................................................................ 75
Figure (3.6): An example of a bimodal histogram with selected threshold T................................ 79
Figure (3.7): An example of global thresholding. (a) Original image, (b) histogram of image (a) , (c) result of thresholding with T=127, (d) outlines of the white cells after applying a 3 X 3 to the image shown in (c) ........................................................ 80
Figure (3.8): Example image with simple region ......................................................................... 83
Figure (3.9): Edges outlining regions in Figure (3.8) .................................................................. 83

Figure (4.1): Initial approximate boundaries of the cell nuclei,outlined using a computer mouse 88
Figure (4.2): Snakes after convergence to cell nucleus boundaries ............................................... 90
Figure (4.3): Line segments used to compute radius .................................................................... 91
Figure (4.4): Line segments used to compute smoothness ........................................................... 93
Figure (4.5): Line segments used to compute concavity and concave points. .............................. 93
Figure (4.6): Line segments used to compute symmetry. ............................................................. 94
Figure (4.7): Sequence of perimeter measurements for computing fractal dimension. ................ 95
Figure (4.8): Training MSE and cross validation MSE of Wisconsin dataset ............................... 103
Figure (4.9): Sensitivity about the mean - Wisconsin dataset....................................................... 105
Figure (4.10): Sensitivity of Radius............................................................................................... 106

Figure (5.1): Mass descriptors for shape....................................................................................... 111
Figure (5.2): Mass descriptors for margin ..................................................................................... 112
Figure (5.3): Mammogram image representing side view of left breast........................................ 114
Figure (5.4): Histogram of image (3.22)........................................................................................ 114
Figure (5.5): Training MSE of DDSM dataset ............................................................................... 124
Figure (5.6): Sensitivity about the mean – DDSM dataset ............................................................ 126
Figure (5.7): Sensitivity of Age ..................................................................................................... 127

# List of Abbreviation

| | |
|---|---|
| **ACS** | **: American cancer society.** |
| **AI** | **: Artificial intelligence.** |
| **ANN** | **: Artificial neural networks.** |
| **BPTT** | **: Back – propagation through time.** |
| **CAD** | **: Computer aided detection / diagnosis.** |
| **CBE** | **: Clinical breast examination.** |
| **CO_POINT** | **: Concave points.** |
| **COMPACT** | **: Compactness.** |
| **DDSM** | **: Digital database for screening mammography.** |
| **EM** | **: Electromagnetic.** |
| **FDA** | **: Food and drug administration (U.S.A. dept of health and human sciences).** |
| **FN** | **: False negative.** |
| **FNA** | **: Fine needle aspirate.** |
| **FP** | **: False positive.** |
| **MAE** | **: Mean absolute error.** |
| **MANOVA** | **: Multivariate analysis of variance.** |
| **MFNN** | **: Multi – layer feedforward neural network.** |
| **ML** | **: Machine learning.** |
| **MLPs** | **: Multilayer perceptrons.** |
| **MRI** | **: Magnetic resonance imaging.** |
| **MSE** | **: Mean square error.** |
| **NCI** | **: National cancer institute** |
| **WDBC** | **: Wisconsin diagnostic breast cancer.** |
| **SMOOTH** | **: Smoothness.** |

# Chapter One

# Introduction