

**Computer Science Department
Faculty of Computer and Information Sciences
Ain Shams University**



Diagnosis of Skin Cancer Using Machine Learning Techniques

**Dissertation submitted to Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams University
In partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science**

By

Munya Abdulmajid Arasi

**Lecturer Assistance of Information Technology Department
Faculty of Engineering, University of Aden, Yemen.**

Supervised By

Prof. Dr. Abdel-Badeeh M. Salem

Professor of Computer Science
Faculty of Computer and
Information Sciences
Ain Shams University

Prof. Dr. El-Sayed M. El- Horbaty

Professor of Computer Science
Faculty of Computer and
Information Sciences
Ain Shams University

Prof. Dr. El-Sayed A. El-Dahshan

Professor at Faculty of Computer and Information Technology
The Egyptian University for E-learning

Cairo, 2018

Acknowledgement

First of all, I humbly give my deepest thanks to God for giving me the opportunity and the strength to accomplish this work.

I would like to thank all the people who helped me to make this work possible, especially my supervision committee. I would like to give my deep appreciation and thanks my supervisor, **Prof. Dr. Abdel-Badeeh M. Salem** for the confidence given to me and his wisdom intelligence, and creative view. His supervision inspired and helped me to complete the researches and writing this thesis, he always gave me the precise advice at the precise moment throughout my PhD study.

Also, I would like to express my sincere gratitude to my co-supervisor **Prof. Dr. El-Sayed M. El- Horbaty**, for always motivating me for going further and further. His enthusiasm, dedication and impressive depth of knowledge have been a great inspiration to me. I learned from him much about the field beyond my thesis topic and I know that this will help me throughout my career. He patiently helped me through my explorations and spared time to discuss ideas with me. I credit most of my intellectual growth during this period to him.

I am most grateful to my co-supervisor **Prof. Dr. El-Sayed A. El-Dahshan**, for his support and guidance of this study. He provided invaluable suggestion and encouragement from the beginning of my research. I benefited numerous discussions through his shared ideas, views from the perspective of their fields of expertise, and helped me when I needed some advice in different matters.

I also thank to the dean of the faculty of computer and information sciences, staff members, and registration committee for their advice and supporting.

Last but not least, I owe special thanks to my parents, my children Shahd and Abdullah and my sisters and brother for their patience with me, and their unlimited support and encouragement to do well.

Abstract

Skin cancer can be considered as one of the most growing types and deadly in the world of cancer; the importance of these tumors is the malignant melanoma. It is one the growing threats in the developed world, which could lead to increasing mortality. The main danger of melanoma is its ability to grow in any part in the body by lymphatic and blood vessels, making it difficult to diagnose in its early stages. As such, mortality rates increase by the time the melanoma is diagnosed in its late stages. Therefore, early diagnosis is crucial for reducing mortality and morbidity. Early stage diagnostic of benign and malignant melanoma is a crucial problem of skin cancer.

Nowadays, medical decision support employs computer technology, which helps the dermatologist to recognize a benign from malignant melanoma. Such technology takes less time, and gives accurate diagnosis with reduced related errors.

The thesis is divided into two parts; the first part summarizes the survey on machine learning techniques for diagnostic sciences in medical imaging, and focuses on a literature review of the Computer Aided Diagnosis (CAD) system that implemented in identifying and diagnosing malignant melanoma of dermoscopy images with emphasis on computer-aided image analysis starting with image acquisition and pre-processing, and finishing with the malignant melanoma classification of dermoscopy images. The techniques employed in each step are explained, and their benefits and limitations are identified. Apart from that, performance of a number of classifiers specifically developed for the diagnosis of malignant melanoma is compared.

The second part is based on various methodologies for CAD system, which are proposed on the problem of malignant melanoma detection using machine learning and deep learning techniques from macroscopic and dermoscopic images. These methodologies are composed of four stages: image acquisition, pre-processing, feature extraction, classification, and evaluation.

Different data sets are used like: Med-Node, Dermatology Information System (DermIS), and DermQuest. Various techniques are implemented in pre-processing such as: morphological operators, adaptive median filter and contrast adjustment. The three proposed techniques for feature extraction include Discrete Wavelet Transform (DWT), Principle Component Analysis (PCA), and the texture feature. These features serve as the input to the various classification techniques such as: *K*-Nearest Neighbor (KNN), Artificial Neural Networks (ANN), and Support Vector Machines (SVM), Neuro-Fuzzy, Decision Trees (DT), Naïve Bayes (NB), deep learning using Stacked Auto Encoders (SAEs). Different statistical tests are performed for analysis the macroscopic and dermoscopic images, and to verify and validate that the accuracy is true for comparing machine learning techniques. Also it is provided guarantee the reliability of study.

The experimental results show that the proposed methodologies are efficient, and achieving high classification accuracies in the range [75 % - 100 %] on macroscopic images, and achieving high classification accuracies in the range [90.0 % - 98.8 %] on dermoscopic images. We compared the obtained results with other techniques. The comparative results showed that the proposed CAD methodologies are more accurate than other techniques in this field of melanoma diagnosis.

Table of Contents

Acknowledgement	i
Abstract.....	ii
List of Tables	vii
List of Figures	ix
List of Abbreviations	xii
List of Symbols	xiv
List of Algorithms	xv
List of Publications	xvi

CHAPTER 1: Introduction

1.1 Introduction.....	1
1.2 An Overview of Cancer Medical Informatics.....	2
1.3 Applied Medical Informatics	3
1.4 Motivation	8
1.5 Problem Definition	11
1.6 Objectives	11
1.7 Contribution	12
1.8 Thesis Organization	13

CHAPTER 2: Machine Learning Techniques for Medical Imaging

2.1 Introduction	14
2.2 An Overview of Medical Imaging	15
2.3 Survey of Machine Learning Techniques in Medical Imaging	19
2.4 Summary	26

CHAPTER 3: Background of Melanoma Skin Cancer

3.1 The Human Skin	28
3.2 Malignant Melanoma	30
3.3 Review of CAD Systems for Malignant Melanoma Detection.....	33
3.4 Summary	45

CHAPTER 4: Analysis of Different Techniques Used in CAD System for Melanoma Diagnostic

4.1 Introduction	47
4.2 Image acquisition and Pre-processing	50
4.3 Features Extraction	56
4.3.1 Texture Features	56
4.3.2 Color Features	57
4.3.3 Geometric features	58
4.3.4 Discrete Wavelet Transform (DWT)	59
4.4 Classification	62
4.4.1 K-Nearest Neighbour Algorithm (k_NN)	63
4.4.2 Decision Trees (DT)	64
4.4.3 Support Vector Machine (SVM)	65
4.4.4 Artificial Neural Network (ANN)	67
4.4.5 Self-Organizing Map (SOM)	68
4.4.6 Neuro-Fuzzy	69
4.4.7 Fuzzy C-Mean (FCM)	70
4.4.8 Naïve Bayes (NB)	71
4.4.9 Clustering	72
4.5 Summary	76

CHAPTER 5: Machine Learning Techniques For Macroscopic Images

5.1 Experiment of Hybrid DWT+PCA features with Machine Learning Techniques	77
5.2 Results and Discussion of Hybrid DWT+PCA features with Machine Learning Techniques.....	86
5.3 Experiments of Resampling by k-Fold Cross-Validation	91
5.4 Results and Discussion of Resampling by k-Fold Cross Validation	93
5.5 Experiment of Statistical Tests for Analysis the Macroscopic Images.	95
5.6 Results and Discussion of Statistical Tests for Analysis the Macroscopic Images	97
5.7 Experiment of Statistical Tests for Comparing Machine Learning Techniques	102
5.8 Results and Discussion of Statistical Tests for Comparing Machine Learning Techniques.....	106
5.9 Summary	110

CHAPTER 6: Machine Learning Techniques For Dermoscopy Images Based on Hybrid DWT and PCA Features

6.1	Experiment of Hybrid DWT+PCA with SVM, k_NN, ANN on DermIS Data Set	113
6.2	Results and Discussion of Hybrid DWT+PCA with SVM, k_NN, ANN on DermIS Data Set	118
6.3	Experiment of Hybrid DWT+PCA with ANN, ANFIS and SVM on composite Data Sets.....	122
6.4	Results and Discussion of Hybrid DWT+PCA with ANN, ANFIS and SVM on composite Data Sets	130
6.5	Summary.....	137

CHAPTER 7: Machine Learning Techniques For Dermoscopy Images Based on DWT and Texture Features

7.1	Experiment Based on Deep Learning Using SAEs for Melanoma Diagnostic.....	139
7.2	Results and Discussion Based on Deep Learning Using SAEs for Melanoma Diagnostic.....	146
7.3	Experiment of Hybrid DWT+PCA and Texture features with Naive Bayes and Decision Tree for Melanoma Diagnostic.....	152
7.4	Results and Discussion of Hybrid DWT+PCA and Texture features wi	158
7.5	Experiment of Statistical Tests for Analysis the Dermoscopy Images ..	164
7.6	Results and Discussion of Statistical Tests Analysis on Dermoscopy Images	164
7.7	Experiment of Statistical Tests for Comparing Machine Learning Techniques on Dermoscopy Images.....	170
7.8	Results and Discussion of Statistical Tests for Comparing Machine Learning Techniques on Dermoscopy Images	171
7.9	Summary	173

CHAPTER 8: Conclusions and Future Work

8.1	Conclusions	177
8.2	Future Work	178
	References	180
	APPENDIX	
	Glossary of Medical Terms	196

List of Tables

Table 2.1	Imaging devices for the diagnosis of skin cancer	18
Table 2.2	A survey on Decision Tree Technique.....	20
Table 2.3	A Survey on SVM Technique.....	22
Table 2.4	A survey on ANN Technique	24
Table 2.5	A survey on NB Technique.....	25
Table 3.1	Layers of the skin and its function.....	29
Table 3.2	Review of CAD System for Malignant Melanoma Detection	34
Table 3.3	Overview of the relationships commonly data set of dermoscopy images and the features for each color space	45
Table 4.1	Benefits and limitations of features extraction techniques	61
Table 4.2	Benefits and limitations of machine learning techniques	73
Table 5.1	Performance of Hybrid DWT+PCA features with machine learning techniques	88
Table 5.2	Performance Comparison for Classification Hybrid DWT and PCA features with ANN, SVM, k-NN	88
Table 5.3	Performance Comparison for Classification Hybrid DWT and PCA with NB, DT.....	89
Table 5.4	Comparative Results of Different Studies	91
Table 5.5	Comparison of Resampling by 10-Fold Cross-Validation and Hold Out	94
Table 5.6	Definition's Columns ANOVA Table	96
Table 5.7	Definition's Rows ANOVA Table.....	96
Table 5.8	The Output Arguments of Hypothesis t-Test Results	97
Table 5.9	The output results of the hypothesis t-test statistical tests	102
Table 5.10	a P-value in terms of a fixed level α test.....	103
Table 5.11	The results of Classifiers by using McNamara's test.....	104
Table 5.12	Comparison Accuracies of NB and k_NN Models using McNemar test variations	108
Table 5.13	Comparison Accuracies of k_NN and DT Models using repeated cross-validation t tests	110
Table 6.1	Performance of Hybrid DWT+PCA with SVM, k-NN and ANN	119

Table 6.2	Comparison of Classification Results for Hybrid DWT+PCA with SVM, k-NN and ANN	120
Table 6.3	Comparative Results of Different Studies	121
Table 6.4	Results of Wavelet Accuracy with Different Number of Hidden Neurons	132
Table 6.5	Results of Wavelet Accuracy with Different Number of Epochs	132
Table 6.6	Performance of Hybrid DWT+PCA with ANN, ANFIS and SVM	133
Table 6.7	Comparison between ANN, ANFIS and SVM Results	133
Table 6.8	Comparative Results of Different Studies	135
Table 7.1	Results of Wavelet Accuracy with a Logistic Sigmoid Function	147
Table 7.2	Results of Wavelet Accuracy with a Linear Transfer Function....	147
Table 7.3	Results of Texture Accuracy with a Linear Transfer Function.....	148
Table 7.4	Results of Texture Accuracy with a Logistic Sigmoid Function..	148
Table 7.5	Performance of DWT and Texture with SAEs	149
Table 7.6	Comparison Between DWT and Texture with SAEs Results.....	150
Table 7.7	Comparative Results of Different Studies	151
Table 7.8	The performance for Hybrid DWT+PCA and Texture features with NB and DT Techniques	159
Table 7.9	Performance comparisons for Classification Texture Features with NB, DT.....	160
Table 7.10	Performance Comparison for Classification Hybrid DWT and PCA with NB, DT	161
Table 7.11	Comparative Results of Different Studies	163
Table 7.12	The output results of the hypothesis t-test statistical tests	169
Table 7.13	Comparison Accuracies of machine learning using McNemar test variations	173

List of Figures

Figure 1.1	Diagram for Medical cancer informatics	3
Figure1.2	A standard general scheme of a CAD system for skin lesions	10
Figure 2.1	Medical Imaging Modalities	16
Figure 2.2	Image Classification Process	17
Figure 2.3	An example of Dermoscope equipment	17
Figure 2.4	Hyper plane of SVM	21
Figure 2.5	A structure of the ANN	24
Figure 3.1	Anatomy the layers of normal and abnormal melanocytes skin in different stages	29
Figure 3.2	The ABCD Rule for early detection of melanoma	31
Figure 4.1	General scheme of a CAD system for malignant melanoma diagnosis	47
Figure 4.2	Various techniques in malignant melanoma diagnosis.....	48
Figure 4.3	Images 1–8, and 9–15 show benign and malignant lesions, respectively	51
Figure 4.4	Overview most commonly data sets used malignant melanoma diagnosis System	52
Figure 4.5	Issues in border detection	53
Figure 4.6	Overview of the most commonly used preprocessing techniques in CAD Systems	54
Figure 4.7	Image Decomposition (3 Level) using DWT	60
Figure 4.8	Principle of SVM	66
Figure 4.9	Feed forward neural network	68
Figure 5.1	The Proposed Methodology of Hybrid DWT+PCA features with Machine Learning Techniques.....	78
Figure 5.2	Macroscopic image examples: (a) benign image; (b) melanoma image	79
Figure 5.3	a) Original Image b) Gray Image	79
Figure 5.4	Filtered image	80
Figure 5.5	The decomposed wavelet at level 1	81
Figure 5.6	Confusion Matrix of ANN on macroscopic images	86

Figure 5.7	ROC for each output class on macroscopic images.....	87
Figure 5.8	The bar graph diagram displays a comparison between hybrid DWT and PCA with Machine Learning Techniques.....	89
Figure 5.9	ROC Curves for Hybrid DWT+PCA with k_NN, SVM, NB, and DT Techniques.....	90
Figure 5.10	The bar graph diagram displays Comparison of Resampling by using 10-Fold CrossValidation and Hold Out	94
Figure 5.11	Screenshot of ANOVA Table for Test (1).....	99
Figure 5.12	Screenshot of ANOVA Table for Test (2).....	99
Figure 5.13	Screenshot of ANOVA Table for Test (3).....	100
Figure 6.1	The Proposed Methodology of Hybrid DWT+PCA with SVM, k_NN, ANN.....	113
Figure 6.2	Images 1–3, and 4–6 show benign and malignant lesions, respectively	114
Figure 6.3	The processes of pre-processing on dermoscopy images a) original image, b) Gray image, C) Filtred image, d) Image after hair removal, f) Image after contrast using gamma value.	115
Figure 6.4	Feature extraction using 2D DWT a) Original image b) decomposed by using wavelet at level 1.....	117
Figure 6.5	Confusion Matrix of ANN.....	119
Figure 6.6	Performance Results for Hybrid DWT+PCA with SVM, k-NN and ANN.....	120
Figure 6.7	The Proposed methodology of Hybrid DWT+PCA with ANN, ANFIS and SVM.....	123
Figure 6.8	Dermoscopy image examples: (a) melanoma image; (b) benign image.....	123
Figure 6.9	The processes of pre-processing a) original image, b) gray image, C) image after hair removal	124
Figure 6.10	The wavelet decomposition of image	125
Figure 6.11	Feature extraction of image using DWT and PCA	126
Figure 6.12	ANFIS System for melanoma diagnostic	129
Figure 6.13	The bar graph diagram displays a comparison between ANN, ANFIS and SVM Results.....	134

Figure 6.14	ROC for each output class	134
Figure 7.1	The Proposed methodology Based on Deep Learning Using SAEs for Melanoma Diagnostic	140
Figure 7.2	Dermoscopy Image	141
Figure 7.3	The processes of pre-processing a) original image, b) gray image, C) image after hair removal, d) filtered image, e) after contrast adjustment.....	142
Figure 7.4	The wavelet decomposition of image	143
Figure 7.5	Stacked Auto Encoder.....	145
Figure 7.6	Confusion matrix for ANN	149
Figure 7.7	The bar graph diagram displays a comparison between DWT and Texture with SAEs results	150
Figure 7.8	The Proposed Methodology of Hybrid DWT+PCA and Texture features with Naive Bayes and Decision Tree for Melanoma Diagnostic.	153
Figure 7.9	An example of deromscopy image	153
Figure 7.10	Resized and gray image	154
Figure 7.11	After removal the hair	154
Figure 7.12	Filtered image	154
Figure 7.13	The bar graph diagram displays Texture features with NB and DT Results.....	160
Figure 7.14	The bar graph diagram displays a comparison between hybrid DWT and PCA with NB and DT results.....	161
Figure 7.15	ROC Curves for Hybrid DWT+PCA and Texture features with NB and DT Classification.....	162
Figure 7.16	Screenshot of ANOVA Table for Test (1).....	166
Figure 7.17	Screenshot of ANOVA Table for Test (2).....	166
Figure 7.18	Screenshot of ANOVA Table for Test (3).....	167

List of Abbreviations

ABCD	rule (A: Asymmetry, B: Boundary, C: Colour and D: Diameter of the lesion)
AC	Accuracy
AHE	Adaptive Histogram Equalization
ANN	Artificial Neural Network
ANOVA	Analysis Of Variance
AR	Association rules
BNs	Bayesian Networks
BPNs	Back propagation network
CAD	Computer Aided Diagnosis
CLAHE	Contrast Limited Adaptive Histogram Equalization technique
CNNs	Convolutional Neural Networks
CPN	Counter Propagation Network
CSLM	Confocal Scanning Laser Microscopy
CT	Computed tomography
DBNs	Deep Belief Networks
Deg-malig	Degree of malignancy
DLNN	Deep learning based Neural Networks
DR	Digital radiography
DSS	Decision support system
DT	Decision Tree
DWT	Discrete Wavelet Transform
EBPN	Error Back Propagation Network
FCM	Fuzzy C-means
FP-Tree	Frequent pattern tree
FTA	Fuzzy-T ARM
GLCM	Gray Level Co-occurrence Matrix
HNB	Hidden Naive Bayes
HSV	Hue-Saturation-Value
IHP	Independent Histogram Pursuit

kNN	k-nearest neighbor
LMT	Logistic Model Tree
LVQ	The Learning vector quantization
MLP	Multi-Layer Perceptron
MRI	Magnetic resonance images
MSIM	Multistage Illumination Modeling
NB	Naïve Bayes
NCI	National Cancer Institute
NFARM	Novel Fuzzy Association Rule Mining
NOS	Normalized Otsu's Segmentation
OCT	Optical Coherence Tomography
PCA	Principle Component Analysis
PNN	Probabilistic neural network
RBF	Radial Basis Function
RF	Random Forest
RNNs	Recurrent Neural Networks
ROC	Receiver Operating Characteristic
SAEs	Stacked Auto Encoders
SE	Sensitivity
SLP	Single Layer Perceptron
SMO	Sequential Minimal Optimization
SOM	Self-Organizing Map
SP	Specificity
SVD	Singular Value Decomposition
SVM	Support Vector Machine
UMCG	University Medical Center Groningen
UV	Ultraviolet
WBCHIR	Wavelet Based Color Histogram Image Retrieval

List of Symbols

Symbol	Meaning
$P(C)$	prior probability
$p(\mathbf{x} C)$	class likelihood
$P(C \mathbf{x})$	posterior probability
$\mathbf{x}$	observation
$\mathbf{x} = (x_1, \dots, x_n)$	a finite set of features
C	Confidence factor
x_i	sample
y_i	associated class label
$(C_1, C_2, \dots, C_K)$	The data is partitioned into K clusters
TPR	true positive Rate
TNR	true negative Rate
FPR	false positive Rate
FNR	false negative Rate
df	Degrees of freedom of the test
sd	standard deviation
ci	Confidence interval vector
H_0	Null Hypothesis
H_a	Alternate Hypothesis
μ	mean value
P-value	a probability –value
α	significance level
