



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكروفيلم

بسم الله الرحمن الرحيم



MONA MAGHRABY



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكروفيلم



شبكة المعلومات الجامعية التوثيق الإلكتروني والميكروفيلم



MONA MAGHRABY



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكروفيلم

جامعة عين شمس

التوثيق الإلكتروني والميكروفيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
علي هذه الأقراص المدمجة قد أعدت دون أية تغيرات



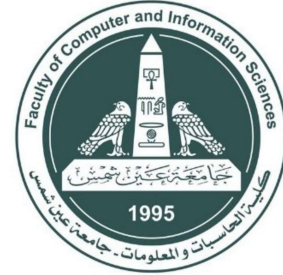
يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار



MONA MAGHRABY

AIN SHAMS UNIVERSITY
FACULTY OF COMPUTER &
INFORMATION SCIENCES
COMPUTER SCIENCE DEPARTMENT



Medical Image Classification based on Ensemble Machine learning Methods

A thesis submitted to the Department of Computer Science,
Faculty of Computer and Information Science, Ain Shams
University, in partial fulfillment of the requirements for the degree
of Master of Science in Computer and Information Science

By:

Nada Sherif Abdel Galil El Askary

Bachelor in Computer Science,
Faculty of Computer and Information Science,
Ain Shams University, Cairo, Egypt

Supervised By:

Prof. Dr. Mohamed Ismail Roushdy
Professor, Department of Computer Science
Faculty of Computers & Information Science
Ain Shams University

Dr. Mohammed A.-M. Salem
Associate Professor, Department of Scientific Computing
Faculty of Computer and Information Sciences
Ain Shams University
Faculty of Media Engineering and Technology
German University in Cairo

September 2020

*Dedicated to my beloved father Sherif El Askary
(Rahimahu Allah) who suffered alot from the cancer
disease and passed away before seeing his daughter
trying to find a solution to help him. . .*

Declaration of Authorship

I, Nada Sherif ABDEL GALIL EL-ASKARY, declare that this thesis titled, 'Medical Images Classification Based on Ensemble Machine Learning Methods' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Nada Sherif Abdel Galil El-Askary

Date: May 2020

“Who doesn’t taste the bitter of Science ‘education’ an hour; he will sip the Humiliation of Ignorance all his life..”

Imam Shafi’i (Rahimahu Allah)

“Our greatest weakness lies in giving up. The most certain way to succeed is always to try just one more time.”

Thomas A. Edison

“Successful and unsuccessful people do not vary greatly in their abilities. They vary in their desires to reach their potential.”

John Maxwell

Abstract

Faculty of Computer and Information Science
Computer Science Department

Master Degree of Science Thesis

Medical Images Classification Based on Ensemble Machine Learning Methods

by Nada Sherif ABDEL GALIL EL-ASKary

Machine learning became a leading technology in all fields of life. In bioinformatics, different techniques are used to learn a machine how to detect and classify pathologies from digital medical images. These images could be of various types such as Computed Tomography (CT) and X-ray. As for lung CT images, machine learning can help radiologist to automatically detect lung nodules which can lead to lung cancer in their early stages and with high accuracy. Lung nodule is an abnormal growth in the lung. These nodules can be either benign (non-cancerous) or malignant (cancer) and in this later case the patient is suffering from pathological lung. Early detection of lung nodule decreases the risk of advanced stages in lung cancer disease and raises the possibility of saving precious human lives.

Random forest (RF), an ensemble machine learning algorithm, is used to detect the lung nodules and classify soft-tissues into nodules and non-nodules. A lung nodule classification approach is proposed to improve early detection for nodules in addition to optimizing RF and reached lung nodule localization. A five stages model has been built and tested using 214 cases from the LIDC database. Stage 1 is image acquisition and preprocessing. Stage 2 is extracting 119 features from each pixel in the lung CT image. Stage 3 is refining feature vectors by removing all duplicate instances and undersampling the non-nodule class. Stage 4 is tuning the RF parameters. Stage 5 is examining different collections from the extracted feature sets to select those scores best for classification. The accuracy achieved by RF is the highest compared to other machine learning classifiers such as KNN, SVM, and DT. The proposed method aimed to analyze and select features that maximize classification results. Pixel based feature set and wavelet based set scored best for higher accuracy. RF was tuned with 80 trees and 0.04 for in-bag-fraction. Best results were achieved by the proposed model are 94.57%, 98% and 96.28% for sensitivity, specificity, and accuracy respectively.

Acknowledgements

Thanks to God for leading me all the way though out my thesis journey.

Thanks to my supervisors; Prof. Dr. Mohamed Roushdy for guiding my to the right way and Dr. Mohammed Abd El Megeed for his endless knowleadge giving and advices for me to be more acheivable student. I would love also to thank Dr.

Shereen Mousa for helping me in the publications processes.

Great thanks to my mother, my husband, my little children and all my family for supporting and being so kind to me while I was so depressed and frustrating.

Their encouragement gave me the push to continue.

Finally, I should also acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study. . .

Contents

1	Introduction	1
1.1	Motivation.....	1
1.2	Problem statement	2
1.3	Objectives.....	3
1.4	Thesis Outline.....	3
2	Machine Learning for Medical Image Classification	5
2.1	Background.....	5
2.1.1	Artificial Intelligence	6
2.1.2	Machine Learning.....	7
2.1.3	Deep Learning	8
2.1.4	Data Science	9
2.1.5	Types of Learning	9
	2.1.5.1 Supervised learning	9
	2.1.5.2 Unsupervised learning	10
	2.1.5.3 Semi-supervised learning.....	10
	2.1.5.4 Reinforcement Learning	10
2.1.6	Machine Learning in Medical Applications.....	11
2.1.7	Ensemble Methods for Classifying Medical Images	11
2.2	Literature Review.....	12
2.2.1	Previous work.....	12
2.2.2	Datasets	13
	2.2.2.1 Computed Tomography (CT) Images	14
	2.2.2.2 Other Types of Lung Datasets	17
2.2.3	Preprocessing and Segmentation	17
2.2.4	Feature Extraction and Reduction	20
	2.2.4.1 Intensity Based Features	20
	2.2.4.2 Texture based features derived from Gray level Co-occurrence Matrix	22
	2.2.4.3 Texture based features derived from Gray Level Run Length matrix.....	24

2.2.4.4	Texture based features derived from 2D and 3D gradient	25
2.2.4.5	List of Common Shape Based Features (Region Properties)	26
2.2.4.6	List of Other Used Features	28
2.2.4.7	Feature Reduction	28
2.2.5	Lung Nodule Detection and Classification Methods	29
2.2.5.1	Using K-Nearest Neighbor Models	29
2.2.5.2	Using Bayesian Classifier	30
2.2.5.3	Using Support Vector Machine	31
2.2.5.4	Using Random Forest	33
2.2.5.5	Using Deep Learning Methods	35
2.2.5.6	Using Other Models	37
2.3	Discussion	39
3	Random Forest	43
3.1	Decision trees	43
3.2	Bagging	44
3.3	Random selection of features	44
3.4	Random forest	45
3.5	RF algorithm steps	46
3.6	Parameters for RF algorithm	47
3.7	RF real life applications	48
3.8	Advantages of using RF	48
4	Lung Nodule Classification using Random Forest	50
4.1	Introduction	50
4.2	Stage 1: Image acquisition & Pre-processing	52
4.3	Stage 2: Features Extraction	54
4.4	Stage 3: Feature Matrix Cleaning and Balancing	61
4.5	Stage 4: Tuning RF parameters	63
4.6	Stage 5: Classification using RF	63
4.6.1	Feature reduction	63
4.6.1.1	PCA	64
4.6.1.2	Dividing Features to Common Sets	65
4.6.2	Classification	65
5	Results and Discussions	66
5.1	Experimental environment	66
5.2	LIDC Dataset	67
5.3	Results and Discussions	67
5.3.1	Phase 1	68
5.3.2	Phase 2	71

6	Conclusion and Future work	75
6.1	Summary	75
6.2	Conclusion	76
6.3	Future work	77

List of Figures

2.1	Traditional programming vs. ML	7
2.2	Abstract block diagram for lung nodule detection models	14
2.3	Some samples from different used datasets	14
2.4	Sample output from some methods	19
2.5	Flow chart for method reduces the false positive rate using SVM	32
2.6	Flow chart for method used DCT filters	35
2.7	Flow chart for Two level NN for lung cells classification	36
2.8	Flow Chart for a lung nodule detection module	38
3.1	Bagging	44
3.2	Simplified Random Forest	46
4.1	Block diagram for the propsed 5 stages model	51
4.2	A sample tiff image from the LIDC dataset, with nodule pointed with the arrow.....	52
4.3	The left image is a sample for a lung CT image that has nodule very attached to the lung wall, marked with red circle. Then the right image is after disconnecting the nodule from the lung wall.....	53
5.1	Chart explains rate of changing the accuracy by changing the size percentage of testing. Experiment environment was 64 trees, 0.01 in-bag-fraction (Phase 1).....	68
5.2	Chart explains rate of changing accuracy by changing the value of in-bag-fraction in the RF. Experiment environment was 64 trees, 20% of the data set for testing (Phase 1).....	69
5.3	Chart explains the rate of changing the accuracy by changing the number of trees in the RF. 0.007 used for the in-bag-fraction, 20% of data for testing (Phase 1).	69
5.4	Rate of changing accuracy by changing the value of the in-bag-fraction. Number of trees used are 50 trees (Phase 2).	72
5.5	Rate of changing accuracy by changing number of trees in the forest. In-bag-fraction is set to 0.04 (Phase 2).	72
5.6	Comparison between the accuracy achieved in phase 1 and accuracy achieved in phase 2 for each feature set	73

List of Tables

2.1	List of researches used LIDC-IDRI dataset in their studies.....	15
2.2	List of Intensity based features.....	20
2.3	List of Texture based features derived from GLCM.....	22
2.4	List of Texture based features derived from Gray Level Run Length matrix.....	24
2.5	List of 2D and 3D gradient based features.....	25
2.6	List of common Shape based features (region properties).....	26
2.7	List of other used features.....	28
2.8	Summary on results of presented researches and methods.....	39
4.1	List of used features	54
4.2	Dividing features list into Sets according to the type or family they descend from.....	65
5.1	Sets of features divided based on their types and families.....	70
5.2	Results in phase 1 using different feature sets	70
5.3	Comparing RF model phase 1 with other machine learner classifiers	71
5.4	Results in phase 2 from different feature sets	73
5.5	Comparison between results from previous work and proposed work using same environment for comparison fairness. <i>Note that number of trees and in-bag-fraction changed from the those used in table 5.4 and so results changed.</i>	74
5.6	Comparing proposed model with other models that uses RF for classification and detection	74

List of Publications and Submissions

List of published work

1. Nada S. El-Askary, Mohammed A.-M. Salem, and Mohamed I. Roushdy, “Lung Nodule Detection and Classification using Random Forest: A Review”, In Proceedings of the 2019 9th International Conference on Intelligent Computing and Information Systems (ICICIS '19), Cairo, Egypt, 2019, pp. 105-111, doi: 10.1109/ICICIS46948.2019.9014706.
2. Nada S. El-Askary, Mohammed A.-M. Salem, and Mohamed I. Roushdy. “Feature Extraction and Analysis for Lung Nodule Classification using Random Forest”, In Proceedings of the 2019, 8th International Conference on Software and Information Engineering (ICSIE '19), Cairo, Egypt, 2019. ACM pp. 248–252. DOI:<https://doi.org/10.1145/3328833.3328872>

List of submitted work to international journal

1. Nada S. El-Askary, Mohammed A.-M. Salem, and Mohamed I. Roushdy, “Features processing for Random Forest optimization in lung nodule localization” 2019